KYIV SCHOOL OF ECONOMICS

MASTER'S PROGRAMME «SOCIAL PSYCHOLOGY»

MASTER'S THESIS
«PERCEIVED PERSUASIVENESS OF AI-GENERATED VS. HUMAN-WRITTEN TEXTS AMONG UKRAINIAN YOUNG ADULTS»

Student: Iryna Tuzko
Supervisor: Andrii Solonskyi

For obtaining a degree: Master's
Speciality: C4 Psychology
Educational programme «Social psychology»

Kyiv 2026

CONTENTS

Abstract

This research examined how Ukrainian young adults perceive the persuasiveness and source credibility of human- and AI-written texts, and what influences their authorship judgements. In a mixed within- and between-subject design, 123 participants rated four unlabelled human-authored and AI-generated texts (randomised within-participant) with a total of 492 evaluations. The resulting data was analysed with linear and generalised mixed-effects models and content analysis of readers' justifications. The texts of actual human and AI authorship did not differ in persuasiveness and credibility ratings. Yet, texts that readers believed to be written by artificial intelligence were associated with significantly lower ratings, regardless of whether the judgement was correct, of the participant's baseline agreement with the topic statement, or of their attitudes towards AI. The cues participants used for identification of AI authorship were largely unreliable (overall 52% accuracy, excluding unsure responses), however, contextual detachment from the realities of war proved the most effective signal.

Keywords: persuasiveness, credibility, artificial intelligence, large language models, perceived authorship, AI identification cues, wartime content.

Word count: 18,693.

INTRODUCTION

The issue of persuasion has traditionally been considered within social psychology as processes of social influence and attitude change in interpersonal communication. With large language models (LLMs) and generative artificial intelligence (AI) becoming increasingly accessible, the question of what makes a message persuasive is taking on a new dimension in human-machine communication. People now interact not only with content produced by other people but increasingly with text generated by LLMs capable of producing logically structured, stylistically coherent, and semantically rich argumentation. Crucially, this content most often reaches readers without any label of its source, so that judgments about authorship become inferences the reader must make rather than facts the reader is given. Accordingly, **the relevance of this study** follows from the need to search for the psychological patterns that underpin the perception of information produced by human authors and by artificial intelligence.

Research problem. A central premise supported by the literature, is that readers are often unable to recognise artificial authorship reliably (Jakesch et al., 2023), while AI-generated texts can be equally or even more persuasive than human-written texts (Spitale et al., 2023; Bai et al., 2025). Research into this issue is particularly relevant among young people, who are the most integrated into the digital environment and actively interact with various types of information sources. It is even more pressing for Ukrainian youth, as misjudged trust in information and its sources can carry very real stakes during wartime. However, the examination of human vs AI persuasiveness and the factors that make readers attribute a text to AI within the Ukrainian context of intensive digital socialisation and heightened information threats remains underrepresented in existing studies.

The following **research question** therefore arises: does the perceived persuasiveness of an unlabelled argumentative text and the credibility of its source differ, depending on the true source (human or artificial), or on the authorship the readers attribute to it, and which factors (source-recognition accuracy, prior topic stance, or attitudes toward AI) may shape those judgments among Ukrainian young adults?

The study explores this question with three groups of **hypotheses**, set out in full in the Research Framework. The first concerns the effect of the true source on persuasiveness and credibility; the second concerns the effect of the source the reader attributes to the text, and the third concerns the conditions under which the perceived source effect strengthens or weakens. A separate exploratory hypothesis asks which affective evaluations may predict whether a reader attributes a text to AI.

The study employs the following **research methods:**

- *theoretical methods*: analysis, generalisation, synthesis of scientific sources;
- *empirical methods*: experimental survey with evaluation of stimulus text material;
- *quantitative methods (mathematical statistics)*: descriptive analysis, linear mixed modelling, generalised linear mixed-effects modelling;
- *qualitative methods of data analysis*: content analysis.

Research aim: to identify the social-psychological factors that influence the perceived persuasiveness of AI-generated and human-written texts among Ukrainian young adults and the cues underlying authorship judgements.

Research objectives:

1. Analyse theoretical approaches to understanding persuasiveness in social and cognitive psychology.
2. Identify the role of AI as a source of information in building trust.
3. Analyse existing empirical studies on the persuasiveness of AI-generated texts.
4. Characterise Ukrainian youth as an audience.
5. Conduct an empirical study of the characteristics of assessing the persuasiveness of texts by different authors.
6. Interpret the results obtained from the perspective of contemporary social-psychological theories and models.
7. Establish a basis for developing practical recommendations aimed at fostering critical thinking and strengthening the resilience of Ukrainian youth, particularly against the influence of manipulative text content generated by AI.

The study uses a mixed-methods **design** that follows from the research question, since persuasiveness and credibility differences by source are best measured with quantitative tests, and the cues readers use when they decide who wrote a text is a question of meaning, which calls for analysing the reasons readers give in their own words. The main strength of this approach is complementarity, however, it adds complexity to the analysis.

Structure of the paper. The paper consists of an introduction, two chapters, a discussion and conclusion, a list of references and appendices. The first chapter examines the theoretical foundations of research into persuasive communication in the context of human-AI interaction. The second chapter is devoted to the organisation and conduct of the empirical study, the analysis of the results obtained, their interpretation, and practical recommendations. The discussion and conclusion part summarises the main findings of the study and identifies prospects for further research.

CHAPTER 1.
THEORETICAL FOUNDATIONS OF PERSUASION IN THE CONTEXT OF HUMAN AND ARTIFICIAL INTELLIGENCE INTERACTION

1.1. Persuasive Communication in Social and Cognitive Psychology

Persuasion is central to many forms of human communication, from political campaigns to product advertisements. Traditionally, the communication process is conceptualised in terms of source (the communicator), message (information being delivered), channel (medium through which the message is communicated) and receiver (recipient or an audience) (Shannon, 1948; Berlo, 1960). Persuasive communication can be defined as "a symbolic process in which communicators try to convince other people to change their attitudes or behaviour regarding an issue through the transmission of a message, in an atmosphere of free choice" (Perloff, 2003, p. 8), and it has long been a major topic of interest for social psychologists. Scholarly work across communication, social and cognitive psychology demonstrates that persuasive communication is a deeply social phenomenon, and our understanding of its nature and effectiveness have largely evolved over the past decades.

One of the earliest systematic accounts in the field of persuasion started within the Yale Communication and Attitude Change Programme. Hovland and Weiss (1951) found that the persuasive effect of a message depends on the perceived credibility of its source, that is, the expertise and trustworthiness of the communicator. Their work introduced the idea that recipients do not evaluate information independently of the communicator. Building upon this idea, Hovland, Janis, and Kelley (1953) developed a more comprehensive model, treating persuasion as a learning process, and argued that its effectiveness depends on the recipient's attention, comprehension and acceptance of the message. Following the Yale group and other scholars, Pornpitakpan (2004) defined source credibility along three dimensions: competence (perceived expertise), trustworthiness (perceived objectivity and reliability), and goodwill (perceived concern for the audience). Their findings concluded that sources featuring the above-mentioned characteristics were significantly more persuasive and influenced the attitudes and behavioural intentions of the audience in a positive way (Pornpitakpan, 2004). The approach that the Yale group, thus, establishes solid foundations for studying persuasive communication from a psychological perspective.

The works mentioned above, however, do not consider how individuals determine the source of information, which is particularly relevant in the digital environment where

people consume multiple messages per day, while source cues are often fragmented or obscured. Addressing this issue requires turning to the source monitoring theory developed by Johnson, Hashtroudi, and Lindsay's (1993). The authors describe source monitoring as a cognitive process "based on characteristics of memories in combination with judgment processes" (Johnson et al., 1993, p. 4) by which individuals attribute information to a particular origin and differentiate facts from fantasy. The two categories in which source-monitoring errors may occur include: disruptions in the original encoding of an event, limiting the retention of perceptual, contextual, emotional, semantic, or cognitive details that normally help individuals identify where information came from, and judgment stage disruptions, when individuals fail to access relevant cues, apply inappropriate criteria, or do not adequately coordinate heuristic and systematic processes when evaluating a source (Johnson et al., 1993). The researchers state that their framework may serve to explain why individuals confuse familiarity with real prior knowledge, reproduce ideas even when they cannot recall their sources, or take fictional material as their own true beliefs and memories (Johnson et al., 1993).

Another line of persuasion research by Sherif and Hovland (1961) added a perspective according to which people evaluate a persuasive message in relation to their prior attitudes (referred to as the anchor point), and map the message into one of the three zones of judgment: the latitude of acceptance (reasonable, acceptable ideas), the latitude of rejection (statements the person finds completely objectionable or unreasonable), and the latitude of noncommitment (neutral or non-defined statements). The authors argued that individuals perceive a message within the latitude of acceptance as much closer to their position than it is (the assimilation effect), while they may be more likely to dismiss the messages falling into the latitude of rejection since they view them as much further from their anchor point (contrast effect) (Sherif & Hovland, 1961). Thus, people are more likely to be persuaded by the messages that do not contradict their existing beliefs, highlighting the important role attitudes play in persuasion.

To better understand why individuals favour belief-consistent messages, we should examine attitudes and their influence on perception and judgment. As Eagly and Chaiken (1993) suggest, attitudes should be defined as "a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor" (p. 1). The authors argued that while an attitude is a singular judgment, it forms as a summary of three distinct types of psychological responses: cognitive, affective, and behavioural (Eagly & Chaiken, 1993). Scholarly work on attitudes hence shows that they organise how individuals interpret new information: a person's likelihood of finding a message persuasive often depends on whether it aligns with their pre-existing views regarding the

topic, rather than the inherent strength of the message itself. The theory of motivated reasoning (Kunda, 1990) adds that motivation affects judgments by directing which cognitive strategies an individual will rely on when constructing and evaluating beliefs. Motivation in Kunda's research (1990) refers to "any wish, desire, or preference that concerns the outcome of a given reasoning task" (p. 480). If accuracy is the motivation, individuals tend to use the strategies they see as best suited for judging correctly, but when motivated by a particular outcome, they choose strategies that the desired outcome is most likely to be achieved (Kunda, 1990). These findings are also mirrored in information processing theories. In 1980, Chaiken began studying how people process messages with varying levels of cognitive effort. Earlier work (Tversky & Kahneman, 1974) demonstrated that individuals rely on cognitive heuristics (or cognitive shortcuts) in the face of uncertainty to quickly and accurately make decisions and minimise cognitive resources spent. Chaiken (1980) applied the concept of cognitive heuristics in her Heuristic-Systematic Model (HSM) to distinguish between two types of information processing: systematic processing involves careful attention to the content, evidence, and logic, whereas heuristic processing involves reliance on less systematic decision rules based on the relative level of expertise of the source of the message or the apparent clarity and confidence of the message. Unlike other dual-processing models, HSM posits that both forms may occur simultaneously, for example, an individual may engage in the analysis of the argument while, at the same time, being influenced by the credibility or presentation style of the communicator. Accordingly, HSM suggests that individuals are neither fully rational nor fully superficial, and they may engage in either or both forms of processing, depending on the level of motivation, ability and/or context.

In the same period as Chaiken's HSM was developed, Petty and Cacioppo (1984) found that personal involvement dictates how people evaluate persuasive messages: when that involvement is high, people scrutinise message quality, but when the involvement is low, people rely on cognitive shortcuts, such as the quantity of the arguments, rather than their substance. They further developed this idea into the Elaboration Likelihood Model (ELM), which acknowledges the central role of the message recipient in the persuasion process and posits that persuasion occurs either by the central or peripheral route, which corresponds to high or low thought, i.e. elaboration (Petty & Cacioppo, 1986). The central route is associated with high-elaboration processing and careful scrutiny of the logic in a message, and the peripheral route involves low-elaboration processing that depends on cues such as source attractiveness, authority, emotional tone, and a general impression of credibility (Petty & Cacioppo, 1986). As such, Petty and Cacioppo (1986) affirm that attitude change resulting from

central processing tends to be stronger, more closely related to behaviour, and more resistant to counter-arguments, whereas the peripheral route is more likely to lead to less consistent attitude changes. Thus, ELM shows that the depth of processing is one of the main conditions determining whether persuasion will be lasting or temporary, and explains why strong arguments do not always persuade and why weak arguments may sometimes be effective.

Ajzen and Fishbein (1980) and their Theory of Reasoned Action (TRA) extend the discussion of persuasion by connecting attitudes and subjective norms to the behaviour. TRA suggests that when a particular action is the desired outcome, persuasive appeals should target the beliefs that support behavioural intentions, clarifying that a persuasive message may change a general attitude without changing a specific intention that translates into action. Following this, Ajzen (1991) developed the theory of planned behaviour (TPB), which considers attitudes, subjective norms and perceived behavioural control (a person's belief in their own ability to successfully perform a specific behaviour, based on their assessment of the personal resources, opportunities, and obstacles). TPB posits that individuals are more likely to form an intention or to engage in a behaviour when their evaluation of the behaviour is positive, they believe important others would support it, and feel capable of performing the behaviour (Ajzen, 1991). Therefore, persuasion aimed at behaviour must address not only what the person believes about the message itself, but how realistic they think the behaviour is and whether it would be socially appreciated.

Research on behavioural perspectives and the role of subjective norms in persuasion demonstrates that individuals rarely make decisions in isolation from social expectations and group influences. To better understand the mechanisms underlying these influences, it is important to turn to Cialdini's study of social influence. His work consolidated six principles of persuasion that function as shortcuts for decision-making and compliance (Cialdini, 2009). Specifically, the principles include: reciprocity, the impulse to return a favour; consistency, the wish to align new actions with past commitments; social proof, the tendency to take cues from the behaviours of others; authority, the tendency to defer to experts; liking, our susceptibility to those we find appealing or similar; and scarcity, the sense that limited opportunities carry greater value (Cialdini, 2009). Cialdini (2016) later expanded upon this list with the inclusion of the seventh principle of unity, which indicates that the more we identify with others, the more we are influenced by them. All of the principles identified above through Cialdini's work demonstrate that persuasion is a deeply social phenomenon that works through identifiable social and emotional cues.

In summary, research shows that persuasive communication is a dynamic process aimed at changing the recipient's attitude or behaviour, and the effectiveness of this process depends on the interaction between (and the characteristics of) the source, message, audience and channel (Perloff, 2003; Shannon, 1948; Berlo, 1960). The Yale tradition addressed the first three of these and established that persuasion occurs through attention, comprehension, and acceptance of the communicated information (Hovland & Weiss, 1951; Hovland et al., 1953). However, because people attribute information to its original source through fallible memory and judgment, as Johnson and colleagues (1993) argued, they are left vulnerable to absorb ideas from forgotten sources or take fictional information as personally known. Where a person already stands on the topic, and their prior attitudes also play an important role in message acceptance (Sherif & Hovland, 1961; Eagly & Chaiken, 1993). The dual-processing models, including Chaiken's HSM (1980) and Petty and Cacioppo's ELM (1986), specify that acceptance is reached through either careful information processing or heuristic cues. Cialdini's principles (2009; 2016) demonstrate that persuasion, as a social influence, operates through psychological cues of reciprocity, consistency, social proof, authority, liking, scarcity, and unity. Finally, Ajzen and Fishbein's TRA (1980) and Ajzen's TPB (1991) show how attitudes, subjective norms, and perceived behavioural control lead to actual behavioural outcomes of the persuasion process. Thus, in the present thesis, persuasiveness should be understood not as an inherent property of a message or a source, but as their perceived capacity to influence attitudes, judgments, intentions, and actions of an audience. The degree of persuasiveness may be conditioned by the perceived source credibility and ability to recall the source, individual differences in cognitive processing of the message content, prior attitudes, and context.

1.2. Persuasion, Trust, and Attitudes in Human–AI Communication

The theories reviewed in the previous part were developed to explain persuasive communication in a human-to-human(s) context, yet a growing share of the messages that audiences now encounter, especially in the digital environment, are produced by machines. The term artificial intelligence covers both the attempt to understand human intelligence by recreating a mind within a machine and the development of technologies that perform tasks once associated with human intelligence (Broussard, 2018; Frankish & Ramsey, 2014). This study adopts the latter, pragmatic sense, concerned with technologies such as conversational agents and automated writing systems, designed to carry out specific tasks within the communication process that were formerly the

province of people. Such communicative technologies have grown largely out of two intertwined subfields, Natural Language Processing (NLP) and Natural Language Generation (NLG), whose joint aim is to let machines interpret messages expressed in human rather than machine language and to produce new messages in human language (Allen, 2003). The text-generating system examined here, ChatGPT (OpenAI, 2022), is a large language model of exactly this kind: a tool that receives a prompt in ordinary language and returns fluent, human-readable prose, and one whose output is now routine enough in everyday communication to make the perception of its authorship a question of practical consequence.

Thus, it is imperative to explore whether the same social psychological frameworks apply when the communicator is not human. Scientific literature suggests that people respond to machines as social actors and thus process their messages through the same psychological mechanisms, while also engaging specific heuristics and dispositions tied to the perceived agency, trustworthiness, and human-likeness of the technological source.

The foundational paradigm in this regard is the Computers Are Social Actors (CASA). It holds that individuals do not distinguish between real social actors and simulated ones, and "mindlessly" respond to computers and other media using the same social scripts and expectations they apply to other people, rather than investing additional cognitive effort to decide how to respond (Reeves & Nass, 1996). As Nass and Moon (2000) concluded, people may act politely toward computers, view information more favourably when it comes from "their own" device, and attribute personality-like traits such as friendliness, competence, or reliability to technological systems. Thus, even purely technical systems may acquire social characteristics and evoke social response.

Certain studies further demonstrate that a technological system is more than a channel in the communication process. While computers typically function as a medium for human-to-human communication, Sundar and Nass (2000) showed that they can be perceived as a source, rather than only a transmitter of content, indicating that a technological entity may possess a degree of agency. This idea is further developed within the MAIN model, which stands for Modality, Agency, Interactivity, and Navigability, proposed by Sundar (2008). According to this model, the mode of information presentation (design features), the perceived agency of the source, the possibility of interaction, and the structure of navigation activate heuristics that influence users' perception, trust and engagement with the presented content (Sundar, 2008). With faster processing power and richer graphics, media agents have become

increasingly capable of human-like behaviour and appearance, becoming distinctly anthropomorphic and further influencing the way individuals understand and relate to technology. Anthropomorphism refers to the perception of human traits or qualities in a technological entity, and signifies that entity's potential for social interaction (Waytz et al., 2010). Waytz and colleagues' research (2010) further demonstrates that individual differences in the tendency to anthropomorphise predict the moral care for, as well as responsibility and trust extended to an agent, and determine how readily the person treats it as a source of social influence on the self. These findings indicate that the form, perceived agency, and human-likeness of the technology all position a technological system as a social actor whose messages may be processed through the same psychological mechanisms that govern human-to-human persuasion.

This reconceptualisation of technology as a communicative partner rather than a passive medium has grown into a distinct field of Human-Machine Communication (HMC). It is most closely associated with Andrea L. Guzman (2018), who formalised HMC as a distinct academic area of study that integrates communication works from Human-Computer Interaction, Human-Robot Interaction, and Human-Agent Interaction. Guzman (2018) defines HMC as the creation of meaning between humans and machines, theorised as a technological agent and a functional "source" – a subject with which people communicate – rather than a channel through which people communicate with one another. To account for new technological advancements, scholars have also revisited the CASA framework. Gambino and colleagues (2020) extend this paradigm and reconceptualise the technologies relevant to the CASA model as not merely computers nor all technologies in general, proposing the term "media agents". The authors define a media agent as "any technological artifact that demonstrates sufficient social cues to indicate the potential to be a source of social interaction" (Gambino et al., 2020, p. 73). As they note, this category is broad, and conversational agents, chatbots, smart devices with social interfaces, wearables, and social robots all qualify as media agents (Gambino et al., 2020). The authors also argue that humans have developed distinct human-media social scripts for interacting with media agents, stating that subsequent research should aim at clarifying what specific types of those scripts exist (Gambino et al., 2020). Because research on persuasive communication has always treated the source as one of the central variables, these works indicate that the questions of credibility, trustworthiness, and authority developed for human communicators become applicable to media agents.

Since this work investigates the persuasiveness of AI-generated texts, it is imperative to consider the perception of this technology in particular. Sundar's (2020)

later research emphasises that AI systems are treated as quasi-agents, capable of initiating actions, making decisions, and communicating in ways that produce social reactions. Additionally, research into task subjectivity suggests that people are more willing to trust algorithms for objective, data-related tasks than for subjective or moral ones (Castelo et al., 2019). Castelo, Bos, and Lehmann (2019) demonstrate that trust in algorithms is lower for tasks perceived as subjective than for those perceived as objective, but they also identify two important qualifications. First, framing a subjective task as more objective increases algorithmic trust for that task (Castelo et al., 2019). Second, the deficit on subjective tasks is grounded in the belief that algorithms lack the affective capacities the task requires, and increasing the perceived affective human-likeness of an algorithm through anthropomorphic cues increases its use for subjective tasks (Castelo et al., 2019). This lays out an important groundwork for understanding the cognitive biases that develop in human-algorithm interactions.

The concept of machine heuristic plays a central role here. It refers to a cognitive shortcut where information created by algorithmic systems may be automatically perceived as more objective, accurate, and unbiased (Sundar & Kim, 2019). A similar phenomenon is known as algorithmic appreciation (Logg et al., 2019), which describes that people tend to trust algorithmic decisions more than human ones, especially in situations characterised by uncertainty or task complexity. At the same time, researchers found the opposite phenomenon, called algorithm aversion, in which people prefer human judgment after noticing even a small error made by an algorithm (Dietvorst et al., 2015; Burton et al., 2020). From a social-psychological perspective, algorithmic appreciation and algorithm aversion reflect a tension between rational and heuristic mechanisms of processing information. Therefore, both trust in algorithms and avoidance of algorithms can be understood as different forms of source evaluation, in which prior experience and general trust in technology play an important role.

According to the study by Lee and See (2004), trust in technology can be characterised as a dynamic process that relies on the user's expectations, prior experience, and assessment of possible risks. Glikson and Woolley's (2020) framework identifies two main antecedents of human trust in AI as the system's representation (full physical presence, virtual, or embedded presence) and its level of machine intelligence (the AI's perceived capabilities). According to the authors, robotic systems are initially granted low trust, which increases through interaction, but virtual or embedded systems initially gain high trust, which declines with use (Glikson & Woolley, 2020). Certain studies further explain that trust can vary systematically from person to person. In particular, the analysis conducted by Russo et al. (2025) demonstrates that gender takes

part in moderating the relationship between AI anxiety and AI attitudes: “higher levels of AI anxiety mainly affect men’s attitudes toward AI compared to women, although gender differences are less evident at high levels of anxiety. Still, at low levels of AI anxiety, women report lower levels of positive attitudes toward AI than men” (p. 6). These patterns indicate that the perception of AI cannot be described as uniform.

Given their role as moderating factors, individual and group differences are central to the present study. Schepman and Rodway (2020) developed the General Attitudes towards Artificial Intelligence Scale (GAAIS), which is a validated instrument for measuring individual differences in AI attitudes empirically. The authors explain these attitudes as relatively stable psychological orientations that impact AI evaluation along both positive dimensions, such as usefulness, efficiency, and the potential to improve life, and negative ones, such as risk, distrust, and fear of losing control (Schepman & Rodway, 2020). A subsequent study by the same authors (Schepman & Rodway, 2023) confirmed the two-factor structure of the scale and found that AI attitudes are predicted by individual psychological characteristics. They further specify that “introverts displayed more positive attitudes than extraverts on both subscales of the GAAIS”, while conscientious individuals “showed more positive attitudes towards the negative aspects of AI, as did people who measured higher on agreeableness” (Schepman and Rodway, 2023, p. 2738). These findings support the treatment of AI attitudes as a stable, measurable variable, with the GAAIS providing the operational instrument, which can be used to test them empirically.

In conclusion, the works reviewed in this section establish two interconnected claims that ground the present study. First, AI can occupy the position of a source in the persuasive communication process, since audiences treat algorithmic systems as quasi-agents capable of communicating in ways that produce social responses and with which people can exchange meaning (Reeves & Nass, 1996; Nass & Moon, 2000; Sundar, 2008; Sundar, 2020; Guzman, 2018). Second, general trust in technology (Lee & See, 2004; Glikson & Woolley, 2020), attitudes toward AI, and broader individual characteristics (Russo et al., 2025; Schepman & Rodway, 2020, 2023) affect how people perceive AI systems and their outputs. Additionally, the machine heuristic (Sundar & Kim, 2019) together with algorithmic appreciation (Logg et al., 2019) may help explain why AI-generated content may be perceived more favourably, while algorithm aversion (Dietvorst et al., 2015; Burton et al., 2020) may contribute to the understanding of the opposite tendency. This creates a theoretical basis for further research on differences in the perception of texts depending on their human or artificial origin.

1.3. Research on Persuasiveness and Identification of AI-Generated Information

Existing empirical research on the persuasiveness and detection of generative artificial intelligence as a source shows that social perceptions of synthetic information are complex and, to a certain extent, ambivalent. While AI-generated messages can be as persuasive as or even more persuasive than human-written ones, people cannot reliably detect AI writing, and AI labelling does not always decrease the credibility of machine-produced pieces.

To begin with, texts produced by large language models can be highly effective in their communicative influence. Comparing the perception of GPT-3-generated tweets and tweets written by Twitter (X platform) users, Spitale, Biller-Andorno, and Germani (2023) found that GPT-3 produced accurate information that was more easily understood than human versions, but it also produced more compelling disinformation, and participants were unable to distinguish AI-generated tweets from human-written ones. Thus, AI can be more effective than humans at conveying accurate information, more effective at conveying inaccurate information, and is unrecognisable as an artificial source in either case. This line of research is also strengthened by Goldstein et al. (2024), who argue that AI-generated propaganda is highly persuasive, and with minor human editing, its effect becomes equal to, and even stronger than, that of real propaganda messages. Thus, this research illustrates that the persuasiveness of AI extends into socially significant domains with a high potential for manipulation. The study of Bai et al. (2025) also supports AI persuasiveness and suggests that the perception of human and synthetic texts is not uniform. The authors demonstrated that AI-generated texts produce real attitude change on public policy issues and were generally comparable to human-written texts in their level of persuasiveness, however, the mechanisms of influence differed: human messages were more often perceived as original and personal, while AI-generated messages were perceived as more logical, factual, and neutral (Bai et al., 2025). These findings imply that AI persuasiveness should be investigated along with the specific cues that may increase its impact, such as perceived logic, neutrality, factuality, and originality.

AI-generated content that resonates with specific backgrounds and demographics produces a more pronounced effect. As Matz and colleagues (2024) demonstrated, ChatGPT-generated messages consistently exert more persuasive influence across domains that included marketing of consumer products and climate action appeals when the messages are adapted to the psychological profile of the recipient. Salvi et al. (2025) tested the same idea in a controlled live-debate setting. They paired participants with

either a human or a GPT-4 opponent and tested how persuasive they were rated with and without access to a participant's sociodemographic data. In pairs where AI and human opponents were not equally persuasive, personalised messages from GPT-4 were perceived as more persuasive 64.4% of the time, with 81.2% relative increase in the odds of higher post-debate agreement (Salvi et al., 2025). The finding reinforces Matz and colleagues' conclusion that the persuasive advantage of an AI source over a human is significant when personalisation is involved, and that this effect applies across both static messages and live interaction.

As much as AI persuasion depends on the perceived qualities of the message and its alignment with the needs and motivations of an individual, researchers note, it may be influenced by additional factors. Hölbling, Maier, and Feuerriegel (2025) offer a systematic perspective supporting this claim through their meta-analysis of seven studies covering 17,422 participants. Across the studies included, LLMs and humans did not differ significantly in overall persuasive effectiveness, but the studies themselves showed a substantial heterogeneity (Hölbling et al., 2025). In their exploratory analysis, Hölbling and colleagues (2025) found that "differences in LLM model, conversation design, and domain are important contextual factors in shaping persuasive performance, and that single-factor tests may understate their influence" (p. 1). These findings, even though based on a small sample of studies, show that investigations of AI versus human persuasion should account for the specific conditions that may make AI more or less persuasive, such as model, interaction format, and topic. The authors also note that future researchers should consider "samples beyond WEIRD populations, systematic testing of personalisation and interactivity, and longitudinal, real-world research designs that capture how persuasion unfolds over time" (Hölbling et al., 2025).

However, when exploring AI persuasion, it is equally important to consider the recipient's ability to identify the AI source in the first place. Research on source recognition shows that people mostly ascribe AI authorship using unreliable cues. As Jakesch et al. (2023) demonstrated, people cannot reliably determine whether a message was written by a human or an AI model in professional, hospitality, and dating contexts. They also point out that the cues their study participants relied on to detect human authorship were misleading and included first-person narration, references to family, personal life experience, and contractions, yet these features can be successfully imitated by language models (Jakesch et al., 2023). Chein and colleagues (2024) provide a similar finding, demonstrating that, on average, participants' differentiation accuracy was only slightly above chance, while individuals higher in fluid intelligence identified AI texts more accurately, those who used social media more often misclassified AI texts

as human-written, and those who determined AI authorship better were less willing to share artificial content. Thus, the perception of AI authorship depends on cognitive resources and digital habits. Whether this recognition gap actually disadvantages AI-generated content, however, is a separate question. Research on perceived source and message credibility of AI and humans, by Huschens and colleagues (2023), which used unlabelled testing in different interface conditions, found no significant variation in these variables, but noted that artificial texts were more often rated as clearer and more interesting. Thus, even when audiences are uncertain about authorship, the AI-generated text can attract more favourable evaluations. Collectively, these findings show that AI authorship identification is largely heuristic and potentially mistaken and that AI-generated texts are not penalised and may even be advantaged on certain dimensions of perceived quality, all of which makes audiences psychologically vulnerable to manipulation.

A parallel question concerns what happens when authorship is openly disclosed. The labelling studies show that even a simple indication that a text was produced by a machine, or a suspicion that it may have been, can change how people perceive the credibility of the message, the “humanness” of the author, and the overall quality of the text. Specifically, Altay and Gilardi (2024) found that headlines labelled as AI-generated were rated lower on perceived accuracy and were less likely to be shared, regardless of whether the headlines were true or false, or actually human- or AI-produced, and the impact of an AI label was three times smaller than that of a "false" label. The authors connect this aversion of AI-generated content to the participants’ expectation of no human supervision and argue that effective labelling therefore requires transparency about the degree of human involvement in content creation. Jia and colleagues (2024) identified the mechanism behind these effects testing news bylines that were specifically labelled as “written by staff writer,” “written by staff writer with artificial intelligence (AI) tool,” “written by staff writer with artificial intelligence (AI) assistance,” “written by staff writer with artificial intelligence (AI) collaboration,” and “written by artificial intelligence (AI)”. The researchers report that what predicted readers’ perceptions of source and message credibility was not the AI labelling but readers’ perceived AI contribution — the extent to which they believed AI had been involved in producing the article (Jia et al., 2024). When participants attributed greater AI involvement to a piece, both message credibility and source credibility declined (Jia et al., 2024). Lermann Henestrosa and Kimmerle (2024) tested the same mechanism through a labelling manipulation on a GPT-written article that was labelled either as AI-written or as human-written, and discovered that the AI label reduced perceived message credibility

along with perceived source credibility, perceived anthropomorphism, and perceived intelligence of the author (Lermann Henestrosa & Kimmerle, 2024). The labelling literature, therefore, identifies a real but conditional effect that depends on how readers perceive the label, and this conditionality has consequences beyond credibility judgments.

It is worth noting that according to McPhedran, Silk, and Ecker (2025), effective interventions that reduce susceptibility to AI-generated misinformation include pre-emptive source discreditation (warning audiences in advance that a particular source is unreliable) and debunking (correcting the false claim after exposure), the former depends on the recipient first recognising the AI source or correctly understanding the meaning of the label. If neither is satisfied, this suggests the corrective intervention cannot do the work it is intended to do. Therefore, it is imperative to investigate how readers perceive AI authorship and to understand which textual, contextual, and individual factors influence their perception of the AI contribution.

Taken together, empirical evidence indicates that AI messages, in particular when personalised, already match and exceed the persuasiveness of human-written ones in many contexts (Hölbling et al., 2025; Spitale et al., 2023; Bai et al., 2025; Goldstein et al., 2024; Matz et al., 2024; Salvi et al., 2025), however, people cannot reliably identify artificial authorship and rely on superficial, inaccurate heuristics (Jakesch et al., 2023; Chein et al., 2024; Huschens et al., 2023). Labelling effects reduce source and message credibility, but depend on perceived level of AI assistance, and the same recognition gap that conditions label effects also limits the possibilities to implement corrective interventions for AI misinformation countering (Altay & Gilardi, 2024; Lermann Henestrosa & Kimmerle, 2024; Jia et al., 2024; McPhedran et al., 2025). Future research should hence examine how individuals infer AI involvement when authorship is not disclosed and extend to samples and contexts that remain underrepresented in the existing literature.

1.4. Ukrainian Young Adults as an Audience

The analysis of persuasion is impossible without considering the specific characteristics of the audience as the information recipients. Ukrainian youth have a variety of different qualities as a social and demographic group that may affect their information processing, trust information, and finally determine the evaluation of digital content. These characteristics include age, digital socialisation, emotional state, war

context, and patterns of interaction with technology. As such, these factors are central to understanding how this group forms judgments about information online.

Firstly, prolonged war, information overload, and constant stress serve a distinct and critical context in which Ukrainians live and interact with information. Polyvianaia et al. (2025) provide evidence of elevated rates of PTSD symptoms among Ukrainian university students, experienced by 48.1% of respondents, while moderate-to-severe anxiety, depression, and insomnia affected 34.1%, 33.6%, and 19.3%, respectively. The authors also noted greater social media use among the predictors of higher depression scores (Polyvianaia et al., 2025). According to Gradus Research (2024), adults aged 25 to 34 are among the most critical of their own psychological state, reporting elevated emotional fatigue (46%) and psychological tension (44%). Such conditions may impede the ability to perform deep information analysis, precisely when the stakes are highest.

The 2022 full-scale invasion also reconfigured how Ukrainian audiences relate to shared identity, and thus changed the lens through which they evaluate information. Marukhovska-Kartunova et al. (2025) document a pronounced shift in Ukrainian cultural self-identification: 84% of Ukrainians stopped engaging with Russian visual media, 71% with Russian audio media, and identification with Ukrainian cultural identity rose from 56.3% in 2006 to 80.8% in 2023. These developments indicate that contemporary Ukrainian young adults evaluate information from a position of strengthened national and cultural self-identification, in which the values information carries are read against a clearer sense of in-group affiliation. The way this in-group affiliation operates in digital information environments is documented by Kyrychenko, Brik, van der Linden, and Roozenbeek (2024). Their analysis of 1.6 million posts from Ukrainian news sources on Facebook and Twitter indicates that after the full-scale invasion, posts expressing in-group solidarity were 92% more likely to gain engagement on Facebook and 68% – on Twitter, while posts expressing out-group hostility had only a 1% effect. For the present study, the implication is that emotionally-charged content reaches Ukrainian audiences through a psychological architecture that is very dependent on ingroup solidarity, shared values, and collective identity.

It is also important to acknowledge that the age-related characteristics and digital habits of young people influence the way they process information. Prensky (2001, 2010) contended that later generations raised in a digitalised environment (the so-called “digital natives”) think differently from earlier ones (“digital immigrants”): they are inclined to process visual and fragmented messages more rapidly, prefer short and dynamic forms of presentation, and are oriented towards interactivity and immediate feedback. Benini and Murray (2014) challenge the dichotomy of digital natives and

digital immigrants proposed by Prensky (2001), arguing that while generations differ in how they engage with technology, they also share similarities that are "mainly based on how much experience people have with the tools" (p. 80). Thus, whether by generation or by accumulated use, young people's digital habits form in a dense, fast-moving environment and condition how they evaluate information they encounter.

Digital socialisation occupies a particular place among the factors that shape information perception in this group. Ukrainians aged 18 to 35 are characterised by a high level of digital competence and active use of online platforms as their main sources of information. Internews' (2023) digital media usage statistic shows that 97% of surveyed Ukrainians aged 18–35 are tracking daily news and reading texts exclusively through mobile devices. Surveys on media literacy show that a large part of young people consume socio-political content largely through social networks and messengers, with Telegram (54%) and YouTube (39%) emerging as leaders (MediaSapiens, 2026). Gradus Research (2023) further emphasises that wartime stressors have turned the vast majority of young Ukrainians into "news-addicted" actors who rely primarily on horizontal peer-to-peer applications and messengers (such as Telegram). It should be noted that the information presented in such platforms often comes unattributed or from anonymous sources, raising the potential for the spread of manipulative narratives.

As the present study concerns manipulative information generated with artificial intelligence, it is necessary to mention that AI use among Ukrainian young adults is substantial and growing. According to the Ministry of Digital Transformation of Ukraine (2026), 42% of adults and 70% of adolescents in Ukraine use AI tools, with younger cohorts driving the adoption. As noted by Yavorskyi (2025), trust in AI tends to be relatively high among young people because digital tools are integrated into their daily lives. These digital habits may translate into a more favourable disposition toward AI and the information it produces. However, MediaSapiens (2026) reports that 43% of respondents do not trust Google descriptions and reviews if they were created by AI, 46% believe that AI can create and spread distorted information, and 70% insist that media producers should label content created with the help of artificial agents. Thus, there is a widespread awareness of AI's role in content creation among Ukrainian youth, alongside demand for transparency about AI involvement.

A critical attitude toward AI and digital sources deserves particular attention, especially in the context of widespread disinformation and manipulative content. Halpern (2014) defines critical thinking as the ability to consciously analyse information, identify logical errors, evaluate the credibility of sources, and make well-grounded decisions. Ukrainian young adults demonstrate high levels of critical

thinking. Internews' (2023) reports that 72% of surveyed Ukrainians aged 18–35 manifest high internal confidence in their ability to distinguish truthful text from unreliable or manipulated narratives. According to the Detector Media's (2026) media literacy index, a majority of Ukrainian audiences regard disinformation as a significant problem, are aware of and employ specific strategies to detect it, which include checking source references and looking for corroborating evidence. This critical involvement may partly explain the scepticism and precision with which this audience evaluates AI-generated and human-written texts. Additionally, Kolodiichuk (2025) notes that disinformation, cognitive influence, and the integration of artificial intelligence in media production as the central pressures on how Ukrainians perceive information and that these pressures play a particular role in the formation of trust in media under hybrid-warfare conditions. For the present study, the implication is that Ukrainians as an audience are already highly attuned to cognitive influence as a security concern, and the strategies they use to evaluate information independently of this adjustment.

This combination of high digital literacy, paired with increasing engagement with AI tools, and a wartime reality creates the specific social and informational context of Ukrainian young adults as an audience. They consume news rapidly, primarily on mobile devices and through peer-to-peer messengers where source attribution is often absent (Internews, 2023; MediaSapiens, 2026; Gradus Research, 2023). Engagement with AI tools is widespread yet coexists with explicit scepticism (MediaSapiens, 2026). Strengthened national and cultural self-identification makes this audience judge information based on a strong sense of in-group affiliation (Marukhovska-Kartunova et al., 2025; Kyrychenko et al., 2024). The self-reported misinformation awareness and ability to recognise manipulative content are high among Ukrainian youth (Internews, 2023). However, studies done to-date confirm increases in PTSD, anxiety and emotional fatigue, which may diminish the amount of cognitive resources available to engage in deep analysis of information (Polyvianaia et al., 2025; Gradus Research, 2024). Based on these characteristics of Ukrainian young adults, it is appropriate to view this audience as individuals who are experiencing a unique type of digital socialisation, but also a high level of cognitive load due to the ongoing war, and express strong in-group solidarity and high misinformation awareness, all of which affects their patterns of information processing.

Conclusion to Chapter 1

The theoretical analysis conducted in this chapter allows us to conclude that persuasiveness is a complex and multidimensional psychological phenomenon. It cannot be reduced only to the formal logic of a message or to the objective "strength" of argumentation. Rather, persuasiveness is formed through the interaction between the content of a message, the way it is presented, the characteristics of the information source, and the individual psychological features of the recipient (Hovland & Weiss, 1951; Petty & Cacioppo, 1986; Eagly & Chaiken, 1993). The analysis concluded that persuasion may occur through two main routes. The first and the central route, which involves processing content analytically, and the second peripheral route, which relies on heuristics and surface-level cues, as described in the Elaboration Likelihood Model and the Heuristic-Systematic Model (Petty & Cacioppo, 1986; Eagly & Chaiken, 1993).

The perception of the information source is one of the key factors in the formation of persuasiveness, and Human-Machine Communication scholarship extends this vocabulary to artificial sources. The Computers as Social Actors paradigm and the MAIN model establish that audiences treat AI systems as social communicators and apply attributional reasoning to them comparable to that applied to human sources, though shaped by specific dispositions toward technology, including the machine heuristic, algorithmic appreciation, and algorithm aversion (Reeves & Nass, 1996; Sundar, 2008; Sundar & Kim, 2019; Logg, Minson, & Moore, 2019; Lee & See, 2004). As a result, recipients' evaluation of AI texts depends on the perceived qualities of the source as much as on the qualities of the message.

Empirical studies on the persuasiveness of AI-generated texts show that large language models are capable of creating messages whose persuasiveness is comparable to, and in some cases even higher than, that of human-written texts, especially under conditions of personalisation (Hölbling et al., 2025; Spitale et al., 2023; Matz et al., 2024; Salvi et al., 2025). At the same time, people are often unable to reliably identify artificial authorship and tend to rely on inaccurate cues that vary across individuals (Jakesch et al., 2023; Chein et al., 2024). The mere labelling of a text as AI-generated does not necessarily reduce its persuasive effect, but operates through the perceived AI contribution it implies — a recognition gap that also constrains the reach of corrective interventions against AI-generated misinformation (Altay & Gilardi, 2024; Jia et al., 2024; Lermann Henestrosa & Kimmerle, 2024; McPhedran, Silk, & Ecker, 2025).

Ukrainian youth, as the target audience of the present study, is characterised by a combination of high digital involvement, developed critical thinking, and high wartime

stress (Polyvianaia et al., 2025; Gradus Research, 2024; MediaSapiens, 2026; Kolodiichuk, 2025). At the same time, strong digital competence, a strengthened sense of in-group affiliation, widespread AI use coexisting with scepticism, and high disinformation awareness create conditions for a more differentiated information-evaluation practice (Kyrychenko et al., 2024; Marukhovska-Kartunova et al., 2025; Gradus Research, 2023; Internews, 2023). There are no studies that directly examine how Ukrainian young adults perceive information created by AI, and specifically how young Ukrainian readers detect the involvement of AI in texts that are not labelled as AI-generated.

Research Framework

The study is built around a set of constructs, which must be explained before they can be measured. The central concepts are persuasiveness, source credibility, and perceived authorship. There is also a group of supporting constructs that specify the conditions under which the central constructs operate: attitudes toward artificial intelligence, prior topic stance, and the affective evaluations of the texts. Each construct is defined below, based on the reviewed literature.

The central concept of this study is *persuasiveness*, interpreted as a psychological effect that emerges through the interaction between the message content, the qualities of its source, and the individual characteristics of the recipient (Hovland et al., 1953). It is also understood as the perceived capacity of a message to influence the reader's attitudes and judgments, rather than as an objective property of the text. This follows the tradition that treats persuasion as a process in which a communicator seeks to change the reader's attitude or behaviour through a message received under conditions of free choice (Perloff, 2003), and in which the effect depends on the reader's attention to, comprehension, and acceptance of that message (Hovland et al., 1953). Defining persuasiveness as perceived rather than actual influence is the decision that organises the whole study: it locates the phenomenon in the reader's evaluation, which allows the same text to be more or less persuasive depending on how its source is construed.

Source credibility is regarded as the reader's perception of the communicator's competence and trustworthiness, the two classical dimensions established in the Yale tradition (Hovland & Weiss, 1951) and confirmed in later syntheses (Pornpitakpan, 2004). The construct is deliberately kept perceptual: it concerns how qualified and well-intentioned the source of information appears to the reader, not reflecting any verifiable expertise of the actual author. This definition links credibility to

persuasiveness, because a source that is judged as competent and honest strengthens its message independently of the message's content.

Perceived authorship is the reader's belief about who produced a text, whether it is human or artificial, without any explicit labelling. This construct is treated as conceptually distinct from the text's true source, due to the reviewed literature showing that readers cannot reliably tell machine-written from human-written text (Jakesch et al., 2023). As a result, the believed source and the actual source can diverge, which is one of this study's core conceptual approaches. Whereas source credibility theory describes how a known source is perceived, the present work focuses on how an unknown source is inferred. It is the inferred, attributed authorship, not the true one, that is expected to shape evaluation. The approach is based on two frameworks. The first one is the source-monitoring framework (Johnson et al., 1993), which describes that people attribute information through judgment processes that are open to error. The second one comprises the CASA framework (Reeves & Nass, 1996) by which an algorithm is considered a social actor or quasi-agent capable of triggering social responses through communication (Gambino et al., 2020). Empirical findings also suggest that AI-generated texts can demonstrate a level of persuasiveness comparable to, or even higher than, human-written texts (Spitale et al., 2023; Bai et al., 2025).

Attitudes toward artificial intelligence are conceptualised as a relatively stable evaluative orientation toward AI as a class of technology (Schepman & Rodway, 2020), here separated into positive and negative dimensions. Treating the two dimensions as distinct is a deliberate choice because an individual may hold both favourable and wary views at once, and the two may relate differently to how messages attributed to AI are perceived. This idea is anchored in two phenomena that may influence how the message is ultimately interpreted by the reader: the machine heuristic, or algorithm appreciation (Sundar & Kim, 2019; Logg et al., 2019), by which machine is seen as more objective, and, conversely, algorithm aversion, by which people are less forgiving toward algorithms and prefer human judgement (Dietvorst et al., 2015; Burton et al., 2020).

The prior attitude toward a topic is defined as the reader's baseline agreement with the topic statement measured before exposure to the text on that topic. This follows the premise that a message is evaluated relative to the reader's existing position (Sherif & Hovland, 1961; Eagly & Chaiken, 1993), and is included to separate the source effects from the effect of existing disposition toward the topic.

Finally, *the affective evaluations,* the valence and arousal a text evokes, its perceived controversy, and the internal resistance it provokes, are conceptualised as the reader's immediate responses to the text. These affective appraisals are based on the

view that persuasion combines careful, content-focused processing with faster, cue-driven processing (Petty & Cacioppo, 1986; Chaiken, 1980), and they capture the peripheral, experiential side of that process. Resistance, in particular, is defined as the friction or push-back that a reader may experience while processing a message. Resistance, as one of the affective appraisals of reading, is of special interest to this study as a cue to how a reader can potentially determine the authorship of the text.

The true source of the text, as human or artificial, functions as the main independent variable (X) in this study and is examined alongside the perceived source, a predictor rather than an independent variable, since it is not manipulated in the present study. Reported persuasiveness and credibility evaluations measured following exposure to the texts are treated as the dependent variables (Y). In addition to the main variables, the study includes two moderators that specify the conditions under which the relationship between X and Y may change. The first moderator is the participant's prior stance toward the given topic, which allows for taking into account the possible influence of the topic itself on the evaluation of the text. The second moderator is the participant's attitude toward artificial intelligence, understood as a relatively stable psychological orientation toward AI technologies, measured using the positive and negative subscales of the GAAIS scale (Schepman & Rodway, 2020).

The theoretical foundation of the study is formed by two interconnected reasoning models. The first concerns the main effect of the source on persuasiveness (Hovland & Weiss, 1951). The second line of reasoning describes the conditions under which the main effect may become weaker or stronger. According to the Elaboration Likelihood Model (Petty & Cacioppo, 1986) and the Heuristic-Systematic Model (Chaiken, 1980), the depth of cognitive processing determines the extent to which a recipient relies on heuristics rather than on the objective logical evaluations of text content. Participants who may suspect the text was AI-generated may be more critical about the message and rate it as less persuasive. At the same time, general attitudes toward AI may function as an implicit prior orientation: a positive attitude may strengthen the machine heuristic and increase receptiveness to AI-generated arguments, while a negative attitude may reduce their persuasiveness regardless of the substantive quality of the text.

This framework thus combines confirmatory and exploratory research design and allows for the formulation of the following hypotheses.

Hypothesis 1a.

H0: There is no statistically significant difference between the persuasiveness ratings of texts created by AI and those written by human authors.

HA: AI-generated texts will be rated as more persuasive than human-written texts.

Hypothesis 1b.

H0: There is no statistically significant difference in the credibility ratings between the texts created by AI and those written by human authors.

HA: AI-authored texts will get higher credibility ratings than human-written ones.

Hypothesis 2a.

H0: There is no statistically significant difference in the persuasiveness ratings between the texts believed to be AI-authored and those believed to be human-authored.

HA: Texts believed to be AI-authored will be rated as less persuasive than those believed to be human-authored.

Hypothesis 2b.

H0: There is no statistically significant difference in the credibility ratings between the texts believed to be AI-authored and those believed to be human-authored.

HA: Texts believed to be AI-authored will get lower credibility ratings than those believed to be human-authored.

Hypothesis 3.

H0: The authorship assumption of the respondent does not change the relationship between the actual source of the text and its persuasiveness rating.

HA: The effect of the actual source on persuasiveness is moderated by the perceived (attributed) source (expecting lowest persuasiveness when the actual AI is guessed AI).

Hypothesis 4.

H0: Baseline topic stance of the respondent does not change the relationship between the perceived source of the text and its persuasiveness rating.

HA: The effect of attributed source on persuasiveness is moderated by the respondent's baseline topic stance.

Hypothesis 5a.

H0: Negative attitudes toward AI, measured by the GAAIS subscale, will have no statistically significant moderating effect on the relationship between the perceived source of the text and its persuasiveness.

HA: The effect of perceived (attributed) authorship on persuasiveness is moderated by negative AI attitudes.

Hypothesis 5b.

H0: Positive attitudes toward AI, measured by the GAAIS subscale, will have no statistically significant moderating effect on the relationship between the perceived source of the text and its persuasiveness.

HA: The effect of perceived (attributed) authorship on persuasiveness is moderated by positive AI attitudes.

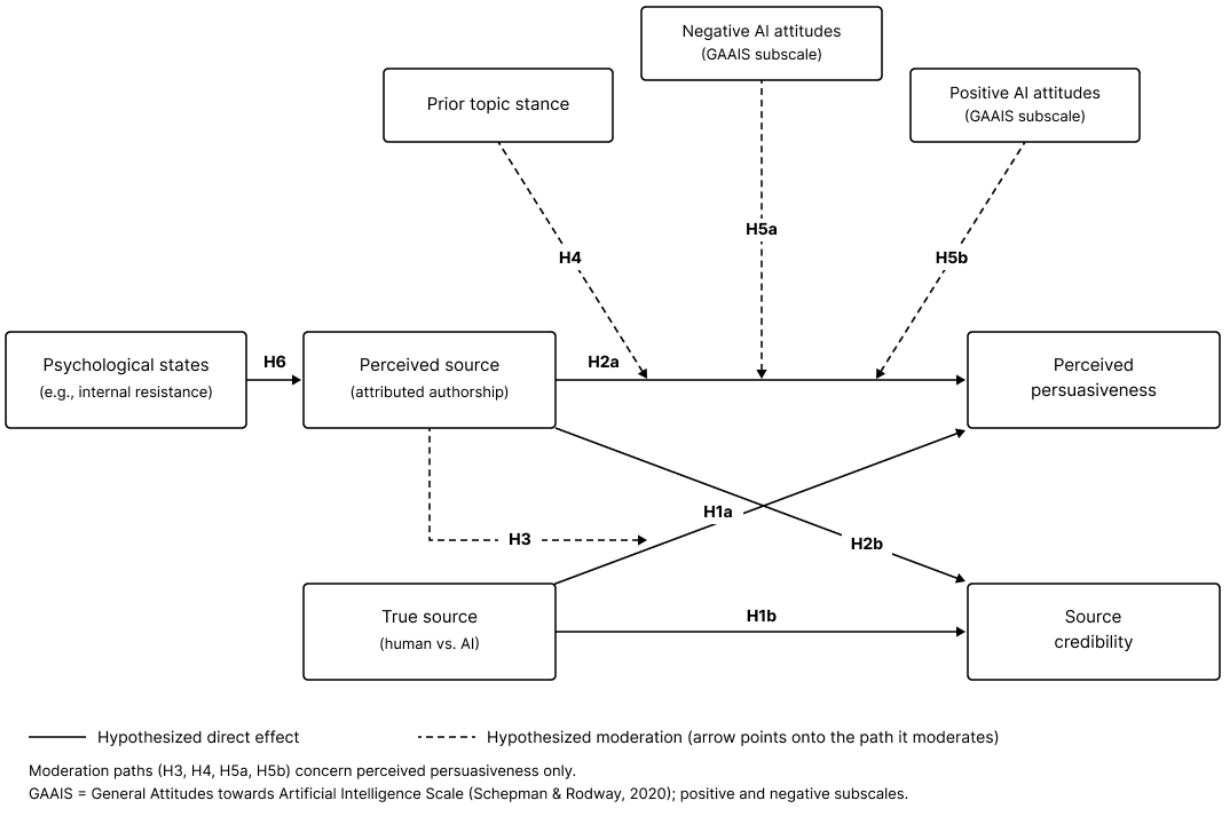
Hypothesis 6 (exploratory).

H0: There is no statistically significant relationship between the participants' affective evaluations and the likelihood of attributing authorship to AI.

HA: Specific affective evaluations (e.g., high internal resistance) predict the likelihood of attributing authorship to AI.

The hypotheses thus operationalise the research question along three complementary lines, illustrated in Figure 1 below: the relationship between the actual source and perceived persuasiveness (Hypotheses 1a, 1b), the effect of perceived source (Hypotheses 2a, 2b), and the conditions under which the effects may strengthen or weaken (Hypotheses 3, 4, 5a, 5b). Hypothesis 6 adds an exploratory dimension by asking whether participants' affective appraisals predict how they attribute authorship, a question motivated by the recognition gap identified in the literature.

Figure 1. Expected Relationship within the Present Research Framework



Moderation paths (H3, H4, H5a, H5b) concern perceived persuasiveness only.
GAAIS = General Attitudes towards Artificial Intelligence Scale (Schepman & Rodway, 2020); positive and negative subscales.

CHAPTER 2.
EMPIRICAL STUDY OF THE PERCEIVED PERSUASIVENESS OF AI-GENERATED AND HUMAN-WRITTEN TEXTS

2.1. Organisation and Methodology of the Study

The empirical study was organised to identify the factors that influence how people evaluate the persuasiveness of non-labelled texts of different authorship, specifically texts generated by artificial intelligence and texts written by human authors. It is important to note that the study is focused not on recording participants' opinions, but rather aimed at examining the psychological markers in their judgements about information and its source.

This research relied on a combination of quantitative and qualitative methods, as the use of both gives a more complete picture of the different factors that might influence how individuals evaluate texts of different authorship. This choice follows from the research question: whether persuasiveness and credibility ratings differ by source is a matter of comparison and degree, best answered quantitatively; what cues individuals rely on when they decide who wrote a text is rather about meaning, better answered by analysing the explanations readers give in their own words.

The experimental survey employed a within- and between-participant design structure. Each participant read and rated four argumentative texts, one for each of four topics. Since every participant read and rated four argumentative texts, one for each of four topics, the topics varied within participants and since texts per topic existed in two versions, one human-written and one AI-generated, of which a participant saw only one, the two versions of any single topic were compared between participants. True source, therefore, could vary within each participant across topics, as well as between participants for a given topic. The arrangement holds the topic constant while sources are compared, and it also lowers cognitive fatigue and removes the cue that seeing two versions of the same text side by side would have created, which was not intended in the present study. The strength of this design lies in the experimental control over the source variable, along with the opportunity to study individual differences in how texts are read. Its main limitation is that perceived authorship, unlike true source, was measured rather than manipulated, so any effect that runs through the reader's belief about the author is correlational and cannot establish direction on its own. The small stimulus set, four texts per participant, is a further constraint, since it leaves the statistical models with little text-level variation.

Participants and sampling

The target population of the study consisted of Ukrainian young adults who have permanently resided in Ukraine for at least the past five years. This group is most relevant for the present research as it is characterised by high digital involvement, frequent interaction with AI tools, active participation in socially significant online discussions and prolonged exposure to both the direct and informational effects of the war. Inclusion criteria were age between 18 and 35 years, granted informed consent, and permanent residence in Ukraine. Participants of non-target age, permanently residing outside of Ukraine, those who did not provide informed consent to take part or who incorrectly filled out the attention check item, were excluded from the final sample. Recruitment procedures were based on convenience sampling and the snowball method. The survey was distributed on social networks, including Instagram, LinkedIn, and Facebook, and administered in one offline session with the first-year law students at the Kyiv School of Economics. After the exclusions, the final sample comprised 123 participants, who together produced 492 text-level evaluations, four per person.

Stimulus materials

The stimuli were eight short argumentative texts arranged as four thematic pairs, with one human-written and one AI-generated text in each pair. The human texts (Appendix A) were prepared by members of the Ukrainian youth NGO "Ukrainian Debate Federation" upon the author's request to produce short texts (around 150 words) in support of the provided statements, with the requirement of no AI usage for text creation. The AI-generated texts (Appendix B) were produced with the large language model GPT-5.3 Instant using zero-shot prompting, with a fresh incognito session opened for each text so that no generation could influence another. The standardized prompt formulation was as follows: "Напиши аргументативний текст українською мовою на підтримку тези, але не повторюй її: {Теза}. Обсяг 120–160 слів" ["Write an argumentative text in Ukrainian in support of the thesis, but do not repeat it: {Thesis statement}. Length: 120–160 words"; author's translation]. The human texts were of comparable argumentative length, and all eight stimuli were then matched at roughly 150–200 words; human texts were lightly edited to remove only the most obvious errors. Source was set at the level of the individual text and was never shown to participants, so authorship always had to be inferred rather than read from a label. The four topics were

chosen to be controversial and socially significant for Ukrainian youth, each representing a wartime-relevant environmental dilemma:

1. After shelling, it is necessary to first test and decontaminate the soil and water to remove hazardous substances, and only then rebuild infrastructure – even if this significantly delays restoring electricity and heating to households.

2. Ukraine should demand environmental reparations as a distinct category of compensation, even if doing so complicates or prolongs peace negotiations.

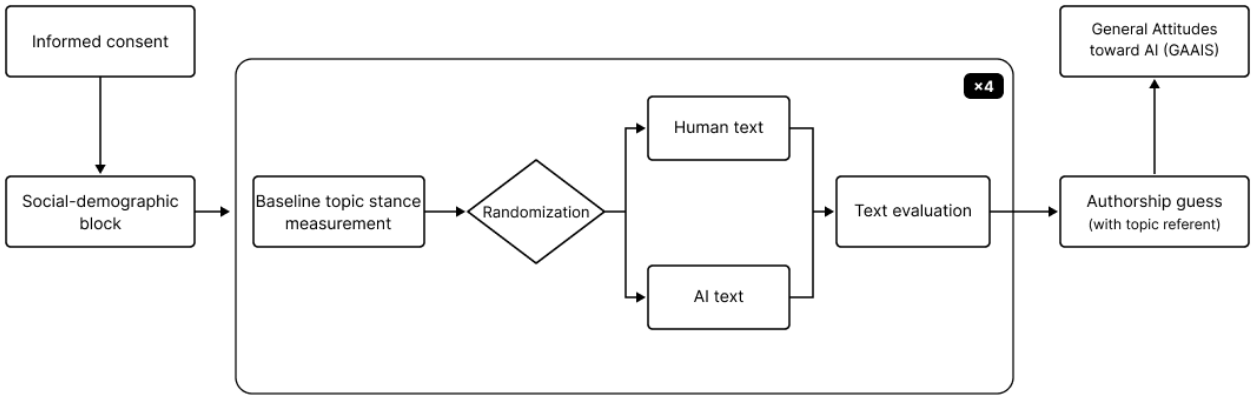
3. We must abandon the construction of fortifications on certain sections of the frontline and border to preserve soils and ecosystems, even if it complicates defence.

4. The state should introduce a high pollution levy for the use of generators in cities, even if this will make some small businesses unprofitable and force families to cut back on basic necessities during blackouts.

Data collection and procedure

Data was collected from March through May 2026 with an online survey (Appendix C) administered on Google Forms. This platform was selected for participation accessibility and convenience for the target group, and provided a standardised delivery of stimulus texts. Figure 2 below illustrates the survey sequence.

Figure 2. Survey scheme



Note. Arrows show the order of steps. The framed block is completed once for each of the four topics (× 4); on each topic
GAAIS = General Attitudes towards Artificial Intelligence Scale (Schepman & Rodway, 2020)

Participants first read the informed-consent form (Appendix D) and passed the inclusion criteria, then completed the demographic block. For each text in turn, they rated their agreement with the topic and how important they found it, read the

argumentative text, and completed the evaluation items about it. The text versions for each topic were randomised through two-option short filler questions between the text blocks, including season of birth, favourite colour, preferred kind of leisure, and the time of day a person feels most productive. The first of the two response options led to the human-written text version, while the second one – to the AI-generated text. In addition to serving as randomisation items, these questions reduced order effects, monotony, and fatigue, and partly masked the survey logic. While reading, participants were given no information about the text source. Only after all four texts had been read, a separate block asked who they thought had written each text, how confident they were about their guesses, and which text characteristics led them to that conclusion (open questions); lastly, participants reported how often they use AI tools and completed the GAAIS.

Measures

Independent variables. The principal independent variable was the true source of each text, a binary distinction between human-written and AI-generated ones. The second principal variable, perceived authorship, is best understood as a predictor since it was observed and measured, not manipulated. It recorded each participant's guess about who wrote a given text (human, AI, or unsure) and was collected only after reading all four texts. The study treats the participants' source guess as separate from the true source, and it plays two roles: predictor in the main hypotheses and outcome in the exploratory analysis.

Dependent variables. Persuasiveness index was the mean of three items that measured whether the text was persuasive, whether its argumentation was strong, and whether the reader felt more inclined to change their opinion after reading it, on the original 1–5 scale. Source credibility was the mean of two items matching its classical dimensions: expertise (perceived field expertise of the author) and trustworthiness (perceived honesty and sincerity of the author).

Moderators and covariates. Attitudes toward AI, moderator in H5, were measured with the General Attitudes towards Artificial Intelligence Scale (GAAIS; Schepman & Rodway, 2020), translated into Ukrainian by the author (Appendix C.1, survey Block 7). The scale contained one attention check question and twenty main items, scored as separate positive and negative subscale means. For the purposes of this study, the negative subscale was not reversed during data encoding to reflect specifically negative attitudes. The prior topic stance represented a single agreement rating with each of the four topics given before each text, serving as a covariate to separate the source effect

from the effect of the topic itself, and as a moderator in H4. Demographic and control variables included age, sex, education, field of study or work, AI usage frequency, and interest in the environmental consequences of the war in Ukraine, all collected in a separate block.

Affective evaluations. Four variables measured the emotional experiences of the participants on a 1–5 Likert scale: emotional valence (pleasant or positive emotions), arousal (tension or anxiety), perceived controversy of the text, and internal resistance to the text. These evaluations separate the affective and the rational components of the response and serve as predictors for the exploratory model in H6.

The Appendix E table lists the full operationalisation of the main study variables.

Ethical considerations

The study design and methodology obtained official approval from the Ethics Committee of the National Psychological Association of Ukraine, Minutes No. 5/26. During the study, all ethical principles were followed. Participants were guaranteed voluntary participation, anonymity, confidentiality, and the right to withdraw at any time without any consequences. The survey form did not collect names or other direct identifiers, and the data was stored in an anonymous form.

The design involved a limited and necessary element of deception, since participants were not told the true authorship of the texts. Doing so would have influenced their evaluations directly and made it impossible to study how participants make their own judgements about authorship. To offset this, each participant received a debriefing at the end that explained the study's true aim and offered the option to withdraw their data. Other foreseeable risks of the study were minor and included fatigue from repeated reading and possible discomfort from exposure to socially significant and potentially sensitive topics. However, these risks were minimised through anonymous format, informed consent, and the right to withdraw at any moment.

2.2. Sample Profile and Descriptive Analysis of Stimulus Text Evaluation

This section describes the sample of the present study and the distribution of the main examined variables across topics, true sources and authorship guesses, and reports the reliability of the measurement instruments. In total, 162 respondents completed the survey. To ensure that the data met the study criteria, preliminary filtering excluded non-eligible participants by age, successful completion of the attention check, and

residence in Ukraine, leaving a final sample of N=123. Each participant read and evaluated four texts, randomly assigned within participants as human- or AI-authored, so the perception variables are organised at the level of the 492 text evaluations. All descriptive and reliability analyses were conducted in R (tidyverse and psych packages).

Sample characteristics

Table 1 sets out the sociodemographic profile of the sample. The mean age of respondents was 25.6 years (SD=5.3) and spanned the full eligible range (18 to 35 years). The sample was predominantly female, with 85 respondents (69.1%) identifying as such, 36 (29.3%) as male, and 2 (1.6%) selecting "other" or declining to answer.

Table 1

Distribution of respondents by sex

Sex	n	%
Female	85	69.1%
Male	36	29.3%
Other / not specified	2	1.6%
Total	123	100%

Source: own research data.

As presented in Table 2, nearly half of respondents held a master's degree (45.5%), followed by a bachelor's degree (23.6%) and secondary education (28, 22.8%). The share of respondents reporting vocational or postgraduate qualifications was relatively small. Yet, taking together the holders of a bachelor's degree or above made up roughly 70.7% of the sample.

Table 2

Distribution of respondents by education level

Level	*n*	%
Secondary	28	22.8%
Vocational (technical school, college)	8	6.5%
Bachelor's degree	29	23.6%
Master's degree	56	45.5%
Postgraduate (PhD)	1	0.8%
Holds an academic degree	1	0.8%

Source: own research data.

The distribution of occupational and study fields (Table 3) was led by information technology (30.9%) and by civil service, law, defence, and security (17.1%), followed by the social and healthcare domains (12.2%), with the remainder dispersed across business, education, public sector, and creative fields.

Table 3

Fields of occupation or study of respondents

Field	n	%
Natural sciences and ecology	1	0.8 %
Education and science	10	8.1 %
IT and technology	38	30.9 %
Business, entrepreneurship, finance, and management	10	8.1 %
Manufacturing, agricultural sector, transport, and logistics	5	4.1 %
Civil service, law, defence, and security	21	17.1 %
Public sector	10	8.1 %
Social sphere, medicine, and healthcare	15	12.2 %
Culture, media, and creative industries	7	5.7 %
Other field	6	4.9 %

Source: own research data.

Most participants also regularly engaged with artificial intelligence instruments, as shown in Table 4, with 45.5% reported using AI tools every day and 38.2% – several times a week, so that more than four in five were at least weekly users, and only 4 participants in total had never used such tools.

Table 4

Frequency of AI tools usage among participants

Frequency	n	%
Never used	4	3.3%
Rarely (several times a month)	16	13.0%
Several times a week	47	38.2%
Every day	56	45.5%

Source: own research data.

Interest in the main topic of the study, the environmental consequences of the war, used across all texts, was rather evenly distributed among the respondents (Table 5). 32.0% gave an indecisive “difficult to say” answer, with comparable proportions expressing some degree of interest (37.7%) and disinterest (30.3%).

Table 5

Reported interest in the environmental consequences of the war

Response option	n	%
Not interested at all	12	9.8%
Rather not interested	25	20.5%
Difficult to say	39	32.0%
Rather interested	31	25.4%
Very interested	15	12.3%

Source: own research data.

Text perception across topics

Table 6 reports text evaluation variables separately for the four topics (t1 to t4), where a single version of the stimulus texts was rated by all 123 participants, and shows that the stimuli fell into two clusters defined by their subject matter rather than by the source manipulation.

Table 6

Means and standard deviations of the text evaluations, by topic

Variable	t1	t2	t3	t4
Prior attitude	2.96 (1.29)	3.46 (1.35)	1.47 (0.85)	1.80 (0.99)
Importance	3.85 (1.04)	3.93 (1.10)	3.20 (1.42)	3.08 (1.38)
Controversy	2.80 (1.33)	2.11 (1.22)	3.80 (1.30)	3.91 (1.23)
Resistance	2.17 (1.25)	1.93 (1.13)	3.59 (1.31)	3.44 (1.28)
Change of mind	2.65 (1.32)	2.13 (1.14)	2.07 (1.17)	2.18 (1.13)
Persuasiveness	3.59 (1.21)	3.69 (1.22)	2.41 (1.29)	2.53 (1.28)
Argument strength	3.37 (1.19)	3.50 (1.24)	2.53 (1.27)	2.64 (1.34)
Expertise	3.04 (1.13)	3.24 (1.22)	2.57 (1.29)	2.59 (1.23)
Trustworthiness	3.58 (1.08)	3.72 (0.97)	3.16 (1.15)	3.15 (1.15)
Valence	2.63 (1.18)	2.54 (1.14)	1.95 (1.00)	1.88 (0.93)
Arousal	2.21 (1.17)	2.14 (1.22)	2.94 (1.31)	2.63 (1.28)
Fact-check intention	3.56 (1.36)	2.87 (1.45)	3.41 (1.46)	3.24 (1.47)
Persuasiveness index (composite)	3.20 (0.97)	3.11 (0.87)	2.34 (1.14)	2.45 (1.13)
Credibility index (composite)	3.31 (0.98)	3.48 (0.94)	2.87 (1.06)	2.87 (1.06)

Note: Entries are M (SD) on the original response scales; n=123 for every text. The persuasiveness and credibility indices are composite scores. Source: own research data.

For the first two topics, respondents held comparatively favourable prior attitudes (M=2.96 and 3.46) and viewed them as relatively acceptable (controversy M=2.80 and 2.11); they provoked little resistance during reading (M=2.17 and 1.93) and had the highest ratings of persuasiveness indices (3.20 and 3.11) and credibility indices (3.31 and 3.48). The remaining two topics turned out far more conflicting: participants expressed substantially less favourable prior attitudes (M=1.47 and 1.80), high perceived controversy (M=3.80 and 3.91), and strongest resistance (M = 3.59 and 3.44), so that composite persuasiveness (M=2.34 and 2.45) and credibility (both M=2.87) fell in step.

Perceived importance varied little across the four topics set (M=3.08 to 3.93), and reported intentions to fact-check the presented information remained moderate (M=2.87 to 3.56). However, the affective appraisals illustrate reinforced topic division: emotional valence was lower for topics 3 and 4 (M=1.95 and 1.88) than for the first two (M=2.63 and 2.54), whereas arousal ran in the opposite direction, peaking for the most contested topic (M=2.94). These results suggest that participants' prior attitudes toward a topic were associated with how controversial they perceived it to be and how much resistance they experienced while reading, which together contributed to more positive or negative evaluations of the texts.

Text ratings by true and perceived source

Table 7 reorganises the same evaluations by the true text source across all topics; the 492 evaluations comprised 213 of human-written and 279 of AI-generated texts.

Prior attitude and importance are omitted here because they preceded the texts and do not depend on their source. The genuinely AI-produced texts attracted slightly more favourable evaluations than the human-authored ones on most dimensions. They were rated marginally higher in composite persuasiveness (M=2.87 vs. 2.65) and credibility (M=3.27 vs. 2.96), as well as some of the underlying judgements (persuasiveness, argument strength, expertise, and trustworthiness). Human-authored texts, by contrast, were rated somewhat more controversial (M=3.35 vs. 3.01) and elicited higher fact-checking intentions (M=3.51 vs. 3.09). Emotional reactions and the reported willingness to change one's mind, by contrast, were almost identical across the two sources (valence M=2.23 vs. 2.27; change-of-mind M=2.28 vs. 2.24). This suggests that differences associated with the true source could be limited mainly to credibility, however, since the observed gaps are small relative to the standard deviations, they are best read as descriptive tendencies, leaving the examination of statistical effect reliability to the confirmatory analysis.

Table 7

Means and standard deviations of the text evaluations, by true source

Variable	Human (*n*=213)	AI (*n*=279)
Controversy	3.35 (1.44)	3.01 (1.48)
Resistance	2.86 (1.41)	2.72 (1.47)
Change of mind	2.28 (1.21)	2.24 (1.22)
Persuasiveness	2.86 (1.42)	3.20 (1.33)
Argument strength	2.81 (1.37)	3.17 (1.28)
Expertise	2.66 (1.29)	3.01 (1.20)
Trustworthiness	3.25 (1.17)	3.52 (1.06)
Valence	2.23 (1.15)	2.27 (1.09)
Arousal	2.47 (1.31)	2.49 (1.27)
Fact-check intention	3.51 (1.40)	3.09 (1.48)
Persuasiveness index (composite)	2.65 (1.14)	2.87 (1.06)
Credibility index (composite)	2.96 (1.09)	3.27 (0.99)

Note: Entries are M (SD) on the original response scales. Source: own research data.

Table 8 groups the evaluations by the source participants believed had produced each text, and, unlike in the above comparison, the differences are more pronounced.

Table 8

Means and standard deviations of the text evaluations, by perceived source

Variable	Human (*n*=198)	AI (*n*=154)	Difficult to say (*n*=140)
Controversy	2.92 (1.50)	3.64 (1.29)	2.96 (1.49)
Resistance	2.41 (1.41)	3.29 (1.36)	2.74 (1.42)
Change of mind	2.37 (1.31)	2.20 (1.14)	2.16 (1.14)
Persuasiveness	3.35 (1.36)	2.70 (1.28)	3.01 (1.41)
Argument strength	3.30 (1.32)	2.73 (1.22)	2.92 (1.39)
Expertise	3.18 (1.17)	2.55 (1.22)	2.74 (1.28)
Trustworthiness	3.59 (1.07)	3.10 (1.08)	3.48 (1.15)
Valence	2.43 (1.22)	2.06 (0.91)	2.20 (1.13)
Arousal	2.38 (1.26)	2.73 (1.31)	2.35 (1.26)
Fact-check intention	3.24 (1.41)	3.44 (1.45)	3.13 (1.52)
Persuasiveness index (composite)	3.01 (1.09)	2.54 (1.06)	2.70 (1.10)
Credibility index (composite)	3.39 (1.00)	2.82 (1.01)	3.11 (1.06)

Note: Entries are M (SD) on the original response scales. Source: own research data.

Texts attributed to a human author (198 evaluations) received the most favourable ratings, with a persuasiveness index of 3.01 and a credibility index of 3.39; texts attributed to AI (154 evaluations) received the least favourable ratings, at 2.54 and 2.82 respectively; and texts that readers could not classify (140 evaluations) fell between the two (2.70 and 3.11). The same pattern reflected in the individual items: perceived controversy (*M*=3.64) and internal resistance (*M*=3.29) were highest when a text was taken to be machine-written. Emotional valence was conversely lowest for AI attributions (*M*=2.06) and highest for human ones (*M*=2.43), and reported fact-checking intentions were greatest when participants believed a text was AI-generated (*M*=3.44). The spread across these three perceived source categories is visibly wider than the spread across the texts' true sources in Table 7, and this pattern sets the stage for the central study comparison between what the text source is and what readers take it to be.

Accuracy of source judgements

Complementing the findings above, Table 9 describes how participants assigned authorship, and identification was poor. Across the 492 evaluations, 183 (37.2%) correctly identified the source (counted as correct only when the guess matched the true source), while 169 (34.3%) misidentified it, and 140 instances (28.5%) were recorded as "difficult to say."

Table 9

Accuracy of source identification

Correctly identified	**Actual source**	***n***	**%**
Yes	Human	96	19.5%
	AI	87	17.7%
No	Human	67	13.6%
	AI	102	20.7%
Difficult to say	Human	50	10.2%
	AI	90	18.3%

Source: own research data.

Considering only those instances when a definite human or AI judgement was made, accuracy was about 52%, close to the level expected from guessing. The errors were asymmetric. Human-authored texts were correctly recognised on 96 occasions but taken for AI on 67, whereas AI-authored texts were correctly recognised on only 87

occasions and taken for human on 102; uncertainty, likewise, was more frequent for AI-authored texts (90 responses) than for human-authored ones (50).

Considering the attributions in their own right (Table 10), participants leaned towards human authorship, assigning 198 texts (40.2%) to a person, 154 (31.3%) to AI, and leaving 140 (28.5%) to the uncertain category. Readers were therefore both inaccurate and biased towards assuming human authorship, and were prone to mistake the AI-generated texts for human writing.

Table 10

Distribution of authorship attributions by true source

Guess	Actual source	*n*	%
Human	Human	96	19.5%
	AI	102	20.7%
AI	Human	67	13.6%
	AI	87	17.7%
Difficult to say	Human	50	10.2%
	AI	90	18.3%

Source: own research data.

Measurement instruments reliability

The reliability of the measurement instruments used in this study was satisfactory throughout, as Table 11 shows.

Table 11

Reliability statistics

Scale	Items	Cronbach's α	McDonald's ω	Spearman–Brown
GAAIS Positive	12	.828	.839	—
GAAIS Negative	8	.818	.820	—
Persuasiveness index	3	.792	.823	—
Credibility index	2	.715	—	.718

Note: α=Cronbach's alpha; ω=McDonald's omega. For the two-item credibility index, the Spearman–Brown split-half coefficient is reported. Source: own data.

The two GAAIS subscales were the most reliable: the twelve positive-attitude items returned a Cronbach's α of .83 and a McDonald's ω of .84, and the eight

negative-attitude items had α=.82 and ω=.82. The three-item persuasiveness index was reliable (α=.79, ω=.82), and the two-item credibility index, for which the Spearman-Brown coefficient is reported, reached .72 (α=.72). Item-rest correlations, provided in Appendix F, supported the composition of each scale: all 25 retained items correlated with the remainder of their scale above the conventional .30 criterion, and none was flagged for exclusion. The scales could therefore be aggregated into the composite scores used in the analyses that follow.

Summary

Taken together, these descriptives frame the analyses that follow. The sample was young, highly educated, predominantly female, and unusually AI-literate, and the measurement scales were internally consistent enough to be combined without reservation. The four texts separated into agreeable and divisive topics, and it was this topical structure, not the authorship manipulation, that organised the appraisal ratings. The texts' true source made only a slight descriptive difference, and in the unexpected direction of a mild advantage for AI-authored material, whereas the source readers believed they were reading produced a clearer ordering of both persuasiveness and credibility. Participants, finally, identified authorship at little better than chance and erred towards human attribution. These patterns set up the inferential models, which separate the effect of true from perceived authorship on the two appraisal indices.

2.3. Results of Statistical Modelling

The statistical analyses proceeded in two stages: a confirmatory set of linear mixed-effects models tested Hypotheses 1 through 5, and an exploratory generalised linear mixed model addressed Hypothesis 6. All variables, as they appeared in the R formulas, along with their explanation and role, are provided in Appendix G. All models were estimated in R (version 4.5.3) using RStudio with the lme4 and lmerTest packages, with estimated marginal means and contrasts obtained through emmeans.

The analysis treated two composite outcomes: a persuasiveness index, averaging the reported readiness to change mind, perceived text persuasiveness, and argument strength items, and a credibility index, averaging the perceived expertise and trustworthiness items, both retained on the original 1-to-5 Likert scale. Single-item Likert variables were treated as interval-level continuous measures. Inspection of distributions in the analytic sample (Appendix H) showed all six items to be within

conventional limits for this treatment (|skew| < 2, |excess kurt| < 7), supporting the use of parametric models (Norman, 2010; de Winter & Dodou, 2010). The prior attitude item was mean-centred, making the variable mathematically continuous and suited for a linear model covariate.

Since all participants provided ratings for multiple texts and all texts were rated by multiple participants, the analysis included crossed random intercepts for both participants and texts, thus controlling for the dependence resulting from the placement of both variables in a repeated measures design.

The true source of the ratings was effect coded (human=–0.5, AI=+0.5), perceived author was entered as a three-level factor (human, AI, and a "Difficult to say" response) and the mean-centred variable measuring prior attitude toward the topic was included as a covariate in all confirmatory models.

The estimated marginal means (EMMs) and pairwise contrasts were calculated with the emmeans package, and Hypotheses 4 and 5 moderation analyses were conducted via simple-slope contrasts in the same manner. Variance inflation factors for the interaction models were within acceptable limits, indicating that multicollinearity did not pose any threat to interaction estimates. Coefficients are reported as unstandardized fixed-effect estimates (*b*) with standard errors and 95% confidence intervals; effects on the attribution outcome are reported as odds ratios (*OR*).

The diagnostic tests (Appendix I) indicated that mixed models would be fitting to analyse the data based on the following: none of the eight confirmatory models had a singular fit; all confirmatory models showed homoscedasticity of residual variances (all $p>.82$); the maximum variance inflation factor (VIF) was 2.38 for the interaction term of the model for H3, which is below the conventional thresholds.

The confirmatory mixed model analyses lead to the conclusion that the mixed models explained between 7 and 14% of the variance (marginal R^2) through fixed effects and between 42 and 55% of the variance (conditional R^2) once the participant and text intercepts were added in the full models; the modest marginal R^2, however, does not undermine the significant fixed effects. These results indicate that a majority of the variance explained can be attributed to stable individual differences between the respondents and not to the factors being tested, yet this pattern is characteristic of subjective rating data and confirms the appropriateness of the mixed-effects method used. Additionally, both the credibility and persuasiveness models passed the normality check (H1b, $p=.245$; H2b, $p=.213$); it must be noted that persuasiveness models showed a statistically significant departure from normality, however, visual inspection of the corresponding residual plots confirmed it to be slight, and mixed-effects models are

robust to such minor deviations. Following the verification of all model assumptions, the analysis proceeded with testing the proposed hypotheses.

Hypothesis testing results

Table 12 below consolidates the formulas used in the R script and the results of all confirmatory tests.

Table 12

Mixed-effects model results for hypotheses H1–H6

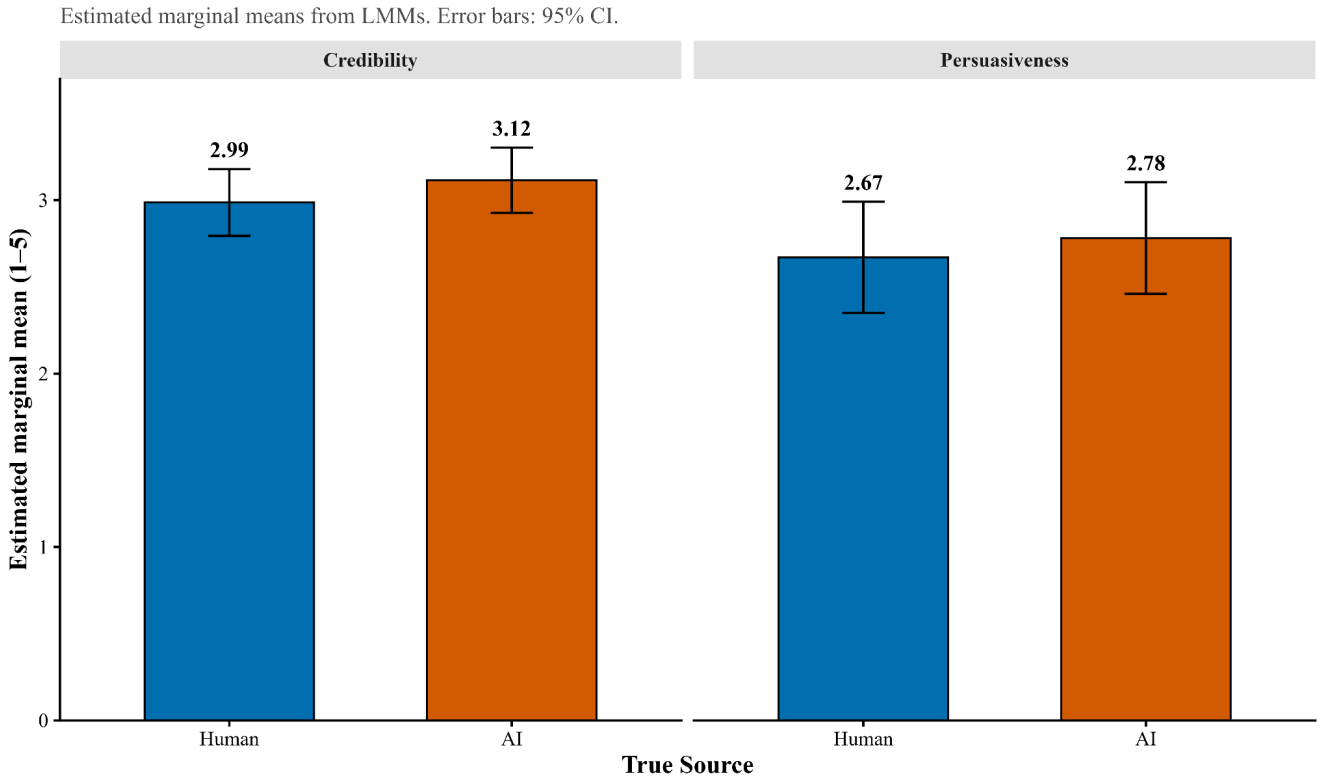
H#	Model formula	Estimate	SE	95% CI	p	Result
H1a	persuasiveness_index ~ source_coded + attitudepre_c + (1|participant_id) + (1|text_id)	0.11	0.08	[−0.04, 0.26]	.154	Not supported
H1b	credibility_index ~ source_coded + attitudepre_c + (1|participant_id) + (1|text_id)	0.13	0.07	[−0.01, 0.26]	.062	Not supported
H2a	persuasiveness_index ~ authorguess_en + attitudepre_c + (1|participant_id) + (1|text_id)	−0.27	0.10	[−0.47, −0.08]	.006	Supported
H2b	credibility_index ~ authorguess_en + attitudepre_c + (1|participant_id) + (1|text_id)	−0.24	0.09	[−0.41, −0.07]	.007	Supported
H3	persuasiveness_index ~ source_coded * authorguess_en + attitudepre_c + (1|participant_id) + (1|text_id)	0.02	0.18	[−0.33, 0.38]	.895	Not supported
H4	persuasiveness_index ~ source_coded + authorguess_en * attitudepre_c + (1|participant_id) + (1|text_id)	0.05	0.07	[−0.08, 0.19]	.430	Not supported
H5a	persuasiveness_index ~ source_coded + authorguess_en * gaais_neg_c + attitudepre_c + (1|participant_id) + (1|text_id)	−0.17	0.12	[−0.41, 0.06]	.143	Not supported
H5b	persuasiveness_index ~ source_coded + authorguess_en * gaais_pos_c + attitudepre_c + (1|participant_id) + (1|text_id)	0.17	0.14	[−0.10, 0.44]	.219	Not supported
H6	guess_is_ai ~ resistance + controversial + valence + arousal + attitudepre_c + (1|participant_id)	1.44	0.19	[1.11, 1.86]	.006	Supported

Note: Estimates are unstandardized fixed-effect coefficients (b) from linear mixed-effects models. For H6, an odds ratio (OR) from a generalised linear mixed model is reported. SE=standard error. CI=confidence interval. All H1-H5 models controlled for baseline attitude toward the topic (attitudepre_c). Random intercepts for participants (participant_id) and texts (text_id) were included where appropriate. Hypotheses were considered supported when the focal effect was statistically significant at $p<.05$. Source: own research data.

The effect of a text's true source on its persuasiveness and credibility (H1) is illustrated in Figure 3, produced in R, confirming that the actual human or AI authorship of the texts did not affect the corresponding ratings. Specifically, AI-written texts were evaluated as marginally, and not significantly, more credible than human-authored texts (b=0.13, SE=0.07, 95% CI [−0.01, 0.26], p=.062), while the difference of persuasiveness evaluations was even smaller (b=0.11, SE=0.08, 95% CI [−0.04, 0.26], p=.154). The estimated marginal means for credibility were 2.99 for human texts and 3.12 for AI texts, and for persuasiveness – 2.67 (human) and 2.78 (AI).

Thus, neither H1a nor H1b were supported, indicating that whether a human or a large language model actually produced a text did not reliably affect how persuasive or credible participants found it.

Figure 3. True source effect on persuasiveness and credibility (H1)

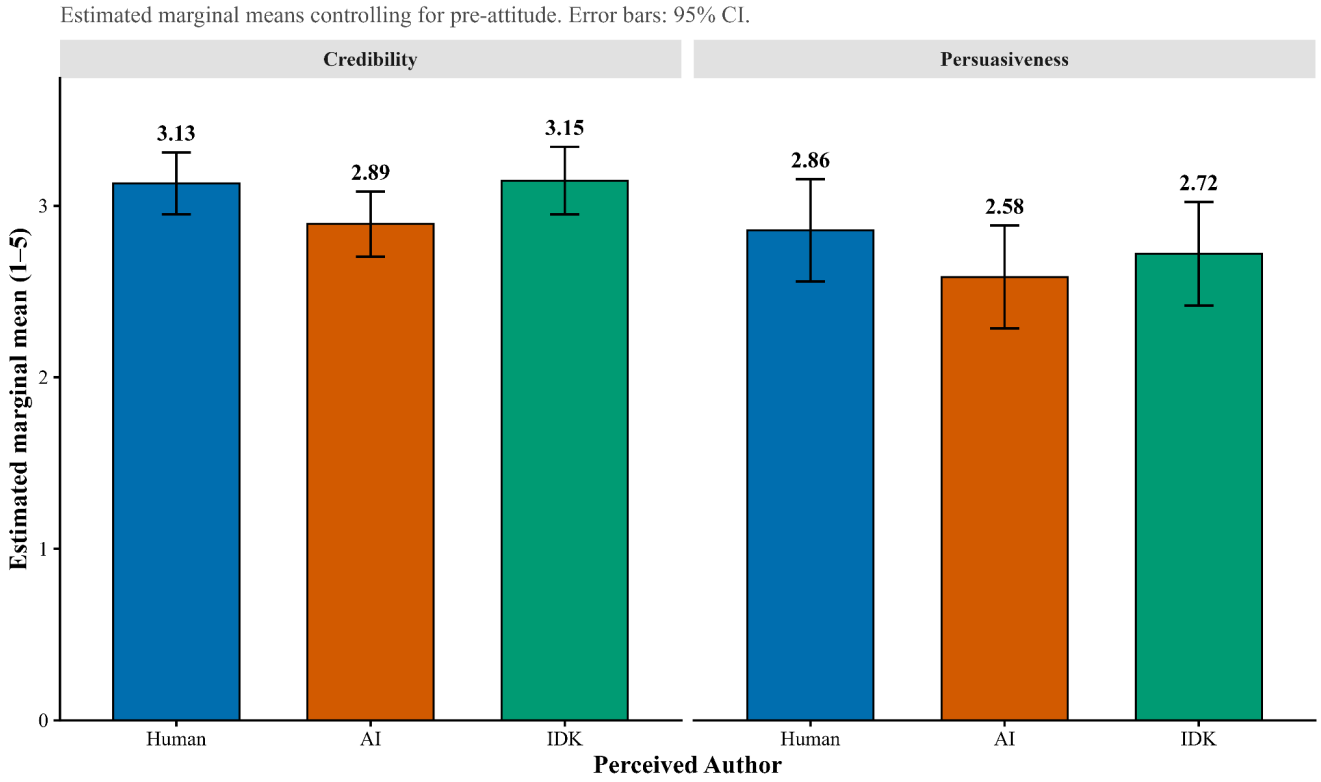


Perceived source served as a stronger predictor of participants' ratings of the texts than did the true source. Figure 4 below shows the results of testing how the participants' belief of who wrote the text they were reading affected its credibility and persuasiveness (H2).

Texts that participants attributed to AI negatively predicted both persuasiveness (b=−0.27, SE=0.10, 95% CI [−0.47, −0.08], p=.006) and credibility ratings (b=−0.24,

SE=0.09, 95% CI [−0.41, −0.07], *p*=.007) against human text attribution (baseline); unsure responses about the authorship (the "IDK" response) corresponded to ratings that were no different from those who assumed a human author (persuasiveness *p*=.177; credibility *p*=.864). Estimated marginal means for persuasiveness fell from 2.86 for the human attribution to 2.58 for the AI attribution, and for credibility – from 3.13 to 2.89, respectively, with the "IDK" category ratings remaining close to the baseline. Both H2a and H2b were therefore supported.

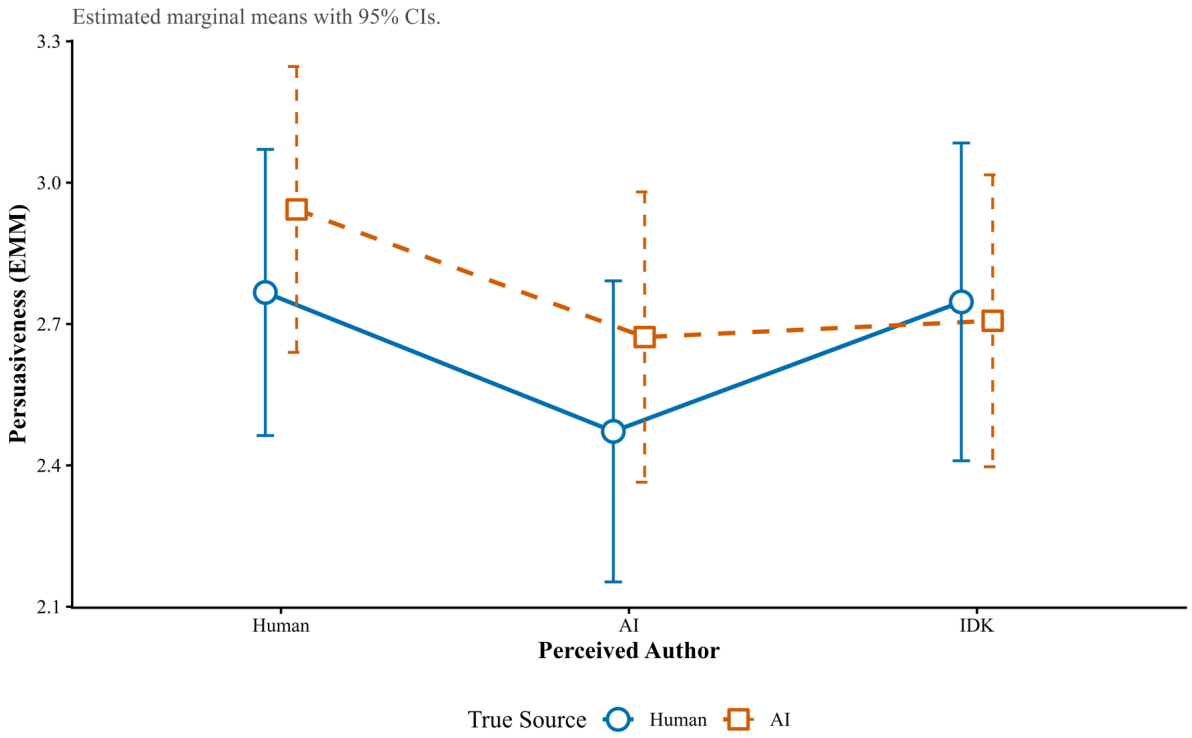
Figure 4. Perceived source effects on persuasiveness and credibility (H2)



The next hypothesis (H3) anticipated that the effect of the actual source on persuasiveness would depend on the perceived source, with the lowest persuasiveness expected when participants correctly guess actual AI as AI.

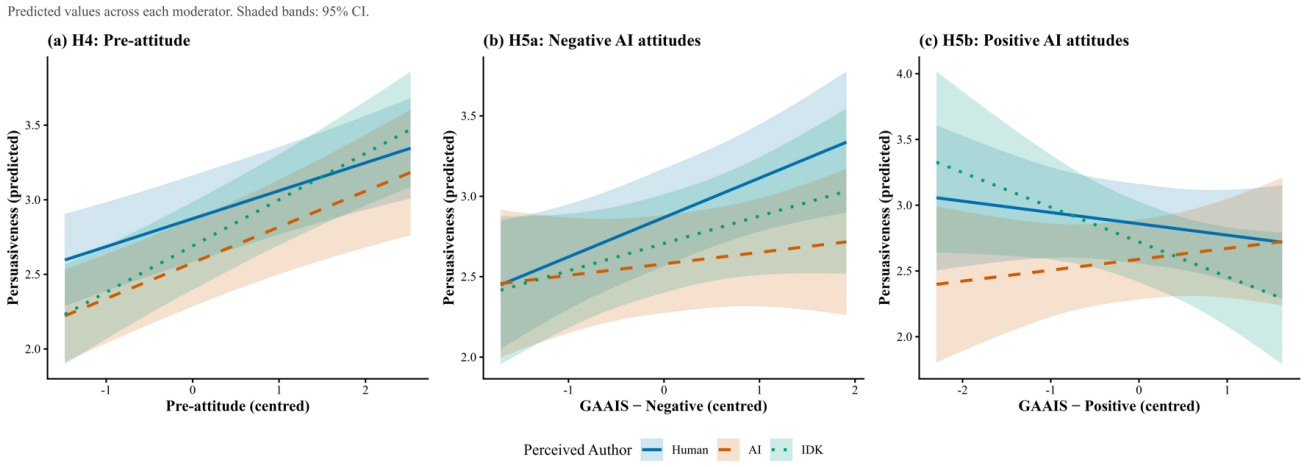
As shown in Figure 5, however, H3 was not supported as neither interaction term was significant (source_coded × authorguess_enAI: *b*=0.02, *SE*=0.18, 95% CI [−0.33, 0.38], *p*=.895; source_coded × authorguess_enIDK*:* *p*=.262). Thus, the perceived source effect operates independently of the actual identification accuracy, so that a text believed to be AI-generated is associated with lower persuasiveness ratings, even when a human author is mistaken for AI.

Figure 5. True and perceived source interaction on persuasiveness (H3)



The tests of three moderators of the perceived authorship effect are provided in Figure 6 across its three panels: baseline stance on the topic (H4) and the negative and positive subscales of general attitudes toward AI, measured with the GAAIS (H5a and H5b).

Figure 6. Moderators of perceived source effect on persuasiveness (H4, H5)



None of the three reshaped the effect: the interaction with prior attitude was flat (b=0.05, SE=0.07, 95% CI [−0.08, 0.19], p=.430), as were the interactions with negative AI attitudes (b=−0.17, SE=0.12, 95% CI [−0.41, 0.06], p=.143) and positive AI attitudes (b=0.17, SE=0.14, 95% CI [−0.10, 0.44], p=.219). The corresponding IDK-vs-human interaction terms were also non-significant in every model (all p≥.068). Although some of the predicted lines in Figure 6 panels (b) and (c) appear to fan apart, they are within the overlapping confidence bands and are not statistically reliable, which is confirmed with the simple-slope decompositions. None of the pairwise contrasts between perceived source conditions reached significance (human vs. AI, b=−0.05, p=.711; human vs. IDK, b=−0.12, p=.162; AI vs. IDK, b=−0.07, p=.623). The negative attitude decomposition produced the same null pattern (all p≥.310), and within the positive attitude decomposition, only the AI-versus-IDK contrast approached significance (b=0.35, p=.079) without crossing the conventional threshold. Therefore, these findings suggest that the persuasiveness rating penalty associated with perceived AI source was not restricted to participants holding specific pre-existing attitudes toward AI or the topics themselves. Interestingly, higher scores on the negative AI attitudes scale predicted higher persuasiveness, although independently of text source (b=0.25, p=.008), whereas positive AI attitudes did not (b=−0.09, p=.427).

The final exploratory analysis asked what could affect the participants' authorship guess itself. To examine this idea, a logistic mixed model was used to predict AI guesses based on four text affective evaluations (resistance, emotional arousal, controversy, valence) and prior topic attitude, as reported in Table 13 below.

Table 13

H6 (exploratory) results

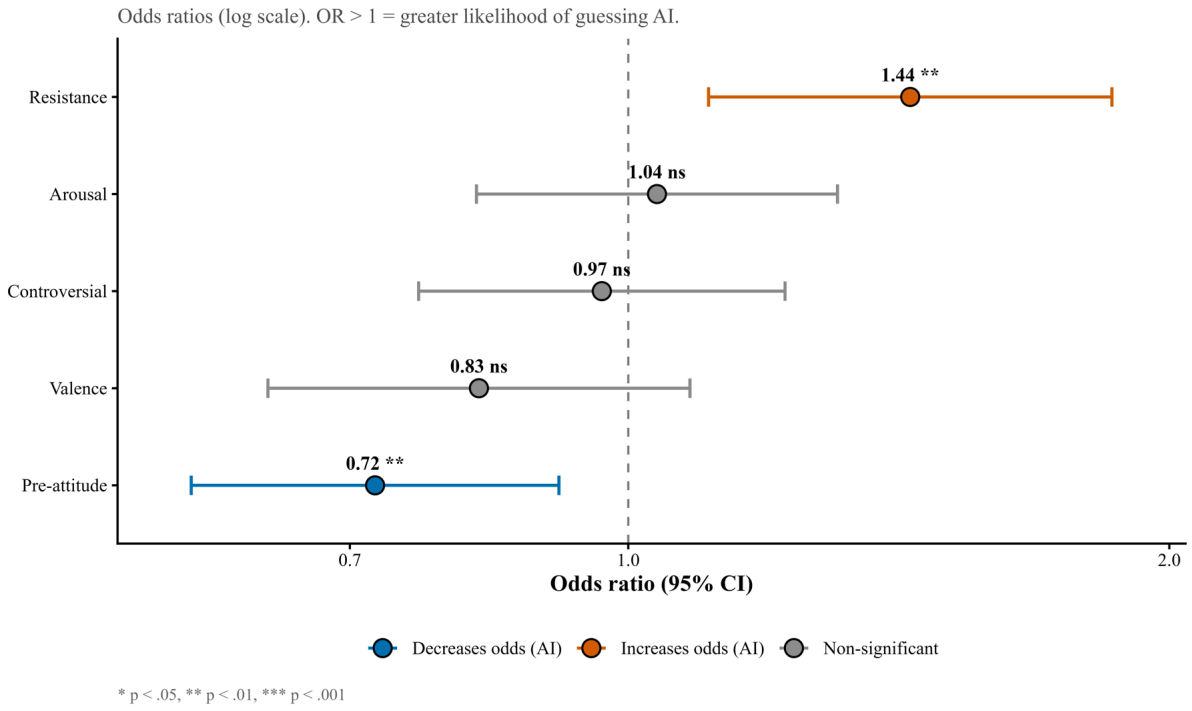
Predictor	OR	95% CI	p	Direction
Resistance	1.44	[1.11, 1.86]	.006	Increases odds of AI guess
Prior attitude	0.72	[0.57, 0.92]	.007	Decreases odds of AI guess
Valence	0.83	[0.63, 1.08]	.165	Non-significant
Arousal	1.04	[0.82, 1.31]	.755	Non-significant
Controversy	0.97	[0.76, 1.22]	.777	Non-significant

Note: OR > 1 indicates greater odds of guessing AI. CI=confidence interval. Estimates are from a logistic generalized linear mixed model with a participant-level random intercept, fitted on trials with a human or AI guess. Source: own research data.

The model included a participant-level random intercept and was fitted on the cases where participants reported either a human or an AI author guess. Figure 7 demonstrates that the results identified two predictors: internal resistance while reading

increased the odds of an AI authorship guess by 44% per unit (*OR*=1.44, 95% CI [1.11, 1.86], *p*=.006), and a more favourable prior attitude toward the topic lowered the odds (*OR*=0.72, 95% CI [0.57, 0.92], *p*=.007). All the remaining variables, including valence, arousal, and perceived controversy, were not statistically significant predictors.

Figure 7. Predictors of attributing authorship to AI (H6)



Accordingly, H6 was supported for the resistance variable and also identified prior topic attitude as a significant predictor. These exploratory results suggest that participants were more likely to assume a text was AI-generated when they experienced greater internal resistance when reading it, while participants who already had more positive attitudes toward the topic were less likely to attribute the text to AI. It is important to note, however, that since this analysis is exploratory, these two effects are best understood as directional rather than precise estimates due to imperfect model calibration with roughly 45% of binned residuals within bounds.

In summary, the results create a distinction between the objective properties of the presented information, including its true source, and the participant's subjective evaluation of that information and its origin. True source (the actual author of a specific text, human or AI) did not influence how persuasive or credible that text was (H1), and that held regardless of who the reader assumed created the text (H3). Conversely, perceived text authorship showed a dramatic opposite relationship with the text

perception: if the participant believed the text was generated with an AI model, the persuasiveness and credibility rating (H2) of that text was significantly lowered, regardless of the participant's prior topic stance (H4) or the general attitudes toward artificial intelligence (H5). The exploratory analysis examined the AI authorship attribution likelihood as an outcome and found that it increases with the internal resistance that participants experience when reading (H6).

Limitations of the present analysis

Several constraints of these results must be acknowledged, and the stimulus set is the first of them. The study rested on four texts shown per participant, which is a small number of items for estimating a text-level random effect, and although the linear mixed models retained crossed intercepts for participants and texts without difficulty, the text variance they recovered was small (close to 3% of the total) and the logistic model for the authorship guess could not estimate it at all, returning a singular fit until the text intercept was removed. Thus, with only four topics, the analysis cannot fully separate the effect of authorship from the particular properties of the texts that happened to be used. The perception of topics themselves also differed, as noticed after data collection. Aggregated to the topic level, two texts were low in rated resistance, and two were high, which also impacted the persuasiveness and credibility differences between them.

The second set of limitations concerns the findings, as this research explains modest amounts of variance with fixed effects (between 7% to 14%), with the majority of explainable variance due to differences between respondents. The statistically reliable effects do not diminish with the high variance in fixed effects, however, these values position source and perceived authorship as two of many modest influences on how a particular text is evaluated. Another fundamental limitation is due to the causal relationship of the central finding. True authorship was randomly assigned to participants and manipulated in the design of the study, while perceived source represents data that was measured and not controlled. As a result, the information contained in the current data does not allow us to definitively identify the causal order: it could be possible that participants inferred the authorship of a particular text as artificial or not while reading and that resulted in lowered ratings, or participants' ratings influenced the assumption that the text had been produced by a machine. The limitation to the exploratory model is the binned-residual analysis used to test its assumptions. Since approximately 45% of binned residuals fall inside the expected rather than desired bounds, the model was not well-calibrated, and its two primary resulting relationships

should not be taken as precise estimates. Besides that, H6 analysis was also restricted to instances where participants committed to a human or AI guess, leaving the uncertain responses aside; it therefore only accounts for the determinants of a decisive assumption and not for what produces uncertainty. A further limitation is the treatment of the single-item five-point Likert scale variables as continuous (interval), for which an ordinal modelling may provide more accurate estimates.

The reach of the estimates themselves is bounded, finally, by the sample and the environment in which participants rated the materials. All participants were consenting Ukrainian young adults between 18 and 35 years of age, and all the ratings were gathered in a single exposure with four short texts in a standardised setting. The reported patterns, therefore, describe how true and perceived source effects relate to text evaluation under those specific conditions and should not be assumed generalisable to real-life digital environments.

2.4. Qualitative Analysis of Reported Cues for Human vs. AI Differentiation

To examine how participants carried out source identification, the study included open-ended questions recording justification accompanying each guess (N=492). The qualitative content analysis was conducted with the assistance of two large language models using structured prompts provided in full in Appendix J. Each guess was assigned to one or more of eight categories derived from the data. To convey the relative prevalence of each theme, the frequency with which each cue type was invoked was also recorded. Because a single justification could invoke several considerations at once, the coding produced 572 thematic instances in total. The coding scheme (Appendix K) captures not the correctness of any individual decision, but the reasoning participants offered for it, allowing to investigate which of the signals people rely on are informative, and which only feel that way.

Across the corpus, the most frequently invoked cue concerned structural predictability and mechanical perfection – the sense that a text was too evenly balanced, too cleanly formatted, or too smoothly punctuated to be spontaneous human writing, which appeared in 104 instances. Attention to logical friction and the depth of expertise behind a text followed (88 instances), as participants asked whether an argument carried specific evidence or thinned into generality. Beyond these two analytic cues, a large body of responses rested on intuition and gut feeling (62 instances), in which participants reported an impression without naming its reason, and on contextual detachment and survival pragmatism (58 instances), a culturally and time-specific signal

in which a text's indifference to the lived realities of the war was read as a mark of AI authorship. Other substantive categories included emotional resonance and the presence or absence of doubt (49 instances), lexical authenticity and localisation (41), and the attribution of hidden intent or agenda (25). The latter refers to the cases when respondents were naming a hidden motive, manipulative intent, populism, or clickbait. Importantly, the largest category of all was non-response or ambiguous reasoning (145 instances), where a sizable share of participants either declined to answer (providing a dot or another meaningless symbol), or replied "cannot determine".

The distribution of cues across the source guesses was uneven. When participants suspected AI was the source behind the text, their reasoning clustered around logical thinness (43 instances among AI guesses), mechanical structure (39), and detachment from wartime context (35), which together formed the core of the case for AI. Attributions of human authorship, by contrast, were both more diffuse and more often intuitive: although the structure of writing dominated the list again (56 instances), gut feeling (35) and non-response (35) ranked immediately below, followed by emotion (29) and authentic and local vocabulary (25). The asymmetry suggests that the assumption of AI authorship tended to be inferred from specific, nameable characteristics, whereas the identification of a human author stemmed more from an unexplained impression or simple absence of anything that struck the reader as artificial. Suspecting the machine was, in this sense, a more deliberate act than crediting the person.

Whether this reasoning truly matched the origin of the texts is another matter. Of the 492 judgments, 140 declined to commit to either author, and among the 352 that did, only 183 were correct – a correct hit rate of 52%, which is barely above chance. Accuracy was modestly higher when readers guessed AI (87 of 154 correct, 57%) than when they guessed human (96 of 198, 49%); and because the texts were not evenly divided (279 had in fact been AI-generated against 213 human-written ones), the slight tendency to attribute a text to a person worked against participants more often than for them. The overall picture is one of detection hovering close to guessing, in line with the finding of Jakesch et al. (2023) that the intuitive "heuristics" people apply to AI-generated language are systematically unreliable rather than merely imperfect.

Disaggregating the cues by whether they accompanied a correct or an incorrect verdict, and by how often each pointed to the true author, exposes where the cues succeed and where they betray the participants. The most diagnostic signal, in both directions, was contextual detachment: when readers grounded a verdict in a text's engagement with or through blindness to the realities of the war, they were right roughly seven times in ten (71% of such judgments when AI was suspected, and 67% when a

human was credited). A passage in which “environmental problems are placed above human lives,” as one participant remarked of a machine-written text, offered the kind of concrete, locally anchored cue that proved difficult to counterfeit. At the opposite extreme stood intuition. Gut feeling was at once among the most frequently cited cues and the least trustworthy one, accompanying a correct verdict only 32% of the time when readers suspected a machine and 37% when they sensed a human, which is worse than chance in both directions. Between these extremes, the cues participants leaned on most heavily proved oddly uninformative. Structural predictability, the most popular signal after ambiguous responses, carried only moderate weight (64% of AI verdicts and 55% of human verdicts were correct), since orderly prose issues as readily from a careful writer as from an advanced large language model. Appeals to logical depth, the single most common basis for suspecting a machine, matched the true author in just 49% of cases, despite the analytic confidence they carried.

The largest single error category was human verdicts on texts that were in fact AI-generated (102 cases), built disproportionately on gut feeling and on the polish of structure and vocabulary; one participant who wrongly cleared an AI passage reasoned simply by gut feeling, adding that “there are no worn-out grammatical structures or words.” Mistaken AI verdicts on genuinely human writing (67 cases) inverted the logic by determining that length or conciseness were strong indicators of a weak argument. For instance, one response mentioned that “it sounded somewhat superficial and weak.” Thus, because fluency, natural phrasing, and familiarity are precisely the features AI models can now convincingly reproduce, such cues were among the least effective for differentiation, similarly to the “flawed heuristics” findings by Jakesch et al. (2023). The reliability problem extends to the lexical cues as well. Participants frequently treated native-sounding vocabulary, idioms, and the absence of awkward or unnatural language as evidence of a human hand, yet lexical cues supported a correct human verdict in only 40% of the cases. Those same cues, however, proved somewhat better as a marker of true AI authorship (55% of correct cases), where the participants conversely mentioned awkward or unnatural phrases, but even here, it was far from decisive.

A further question explored is whether participants knew when to trust their own judgments. The results showed that the most confident responses either leaned on specific structural and logical observations or were ambiguous, while the least confident ones mostly stated an ambiguous cue or a non-response. Among committed authorship guesses, the judgements made with complete confidence were correct 57% of the time, which is modestly higher than those made with the weakest confidence (43%). The intermediate “somewhat confident”, “difficult to say” and “somewhat not” levels were

similar in accuracy, resulting in the correct rates of 51%, 53%, and 49% respectively. As such, even the most confident participants were correct only marginally more often than by chance, and, alongside structural predictability and mechanical perfection (13 cases), an equal part of fully confident responses provided no cue at all (13 cases as well). The pattern that emerged from this part of the analysis suggests that the participant's subjective sense of certainty was a poor guide to whether they were actually correct.

The qualitative findings described here show that there is a range of different types of cues that participants used to identify the sources of the texts they were exposed to, including text mechanics and structure, depth of the logic, empathy, contextual grounding, and "pure intuition" among others. Yet, this rich repertoire was poor in reliability. The cues participants named most and trusted most (structural polish and gut feeling) were the ones most likely to mislead, while a more reliable cue (a text's detachment from the existential realities of the war) appeared rather rarely. The main pattern is that participants could not reliably establish what the source of a text was, and they were least accurate precisely when they felt most certain. Since earlier quantitative results showed that believing a text is AI-written lowers its persuasiveness and credibility, this analysis adds that participants may be penalising texts based on an authorship judgment they simply cannot accurately make, which leaves them vulnerable to potentially misinterpret both AI-generated and human-written content.

2.5. Interpretation of the Obtained Results and Practical Recommendations

The primary aim of this study was to examine whether persuasiveness and credibility of AI-generated and human-written texts differ depending on their true source. The results of statistical analyses showed the actual author, human or AI, did not produce a reliable difference in how persuasive or credible participants found a text (H1). Texts produced by ChatGPT 5.3 Instant were rated as slightly more credible than human-written ones, but this difference was not statistically significant, consistent with research showing that large language models can be as persuasive as human authors (Huschens et al., 2023; Bai et al., 2025; Hölbling et al., 2025; Spitale et al., 2023).

A second set of analyses examined the source that participants attributed to each text after reading. Texts that participants believed were written by AI were associated with significantly lower persuasiveness and credibility ratings than the texts they believed a person had written (H2), and this difference in ratings was present whether the guess was correct or not (H3). A human-written text taken for AI corresponded to lower ratings, while an AI-written text taken for human text was not. The discovered

pattern could reflect the algorithm aversion phenomenon (Dietvorst et al., 2015; Burton et al., 2020), but because the perceived source was measured as a post-reading guess rather than manipulated, this study cannot establish the direction of these relationships, but provides an informative finding for future exploration. This association between perceived authorship and lower ratings did not appear to depend on the readers themselves: neither their prior stance on the topic (H4) nor general attitudes toward AI (H5). One unexpected association should be noted with caution: participants with more negative attitudes toward AI tended to give higher persuasiveness ratings overall, regardless of the source. This result is exploratory, but may suggest that wariness toward AI is associated with reading arguments more carefully rather than dismissing them.

The other part of the main aim was to examine how readers tell AI and human texts apart, and, generally, they did so poorly. Identification was only slightly above chance: 52% for committed guesses, excluding unsure responses. Participants tended to assume human authorship more, and they often took AI-written texts for human ones (20.7%). The exploratory model showed two variables associated with an AI guess: the more internal resistance participants reported while reading, the more likely they were to attribute a text to AI, while a more favourable prior attitude toward the topic was associated with a lower likelihood of an AI guess (H6). As this analysis is exploratory and observational, these relationships should be understood as associations of uncertain direction, but they suggest that attributing authorship is partly an emotional and motivated judgement (Kunda, 1990), in which discomfort or disagreement co-occurs with suspecting a machine. The reasons participants gave, as analysed in the qualitative results section, correspond to this vision. The cues they mentioned and trusted most, such as the "structure of the writing" and "gut feeling," were the least reliable, while the one cue that helped to identify the true author more accurately (detachment from the realities of the war) was used only rarely. At the same time, because this cue was rare and tied to a single domain, it offers readers little protection against AI-generated texts on any topic outside of individuals' direct experience. Additionally, a context-aware or personalised model (cf. Matz et al., 2024; Salvi et al., 2025) could plausibly imitate local, lived detail too, thus eroding the effectiveness of even this cue.

Since identification was close to chance and lower ratings were associated with what readers believed rather than with the real source, media literacy work should focus less on teaching people to recognise AI and more on judging the quality of an argument on its own terms, whatever its author may be. Since the one reliable cue was a text's connection to lived experience, interventions can encourage attention to local and contextual details in a text rather than surface-level linguistic or stylistic features.

Finally, as corrective interventions against AI-generated misinformation, such as warning about a source in advance or debunking a claim afterwards rely on first recognising the AI source, which readers cannot do reliably (McPhedran et al., 2025), and because lower credibility is associated with perceived rather than actual AI involvement (Jia et al., 2024), clear labelling of AI content (Altay & Gilardi, 2024) is a more dependable approach than leaving people to judge authorship for themselves.

In summary, the practical significance of the work lies in the possibility of using the obtained results in educational practices aimed at the development of critical thinking and media literacy, in the development of communication strategies taking into account the characteristics of audience perception, in research into human interaction with technologies, particularly artificial intelligence, and in the training of social psychology professionals who work with information influences.

Conclusion to Chapter 2

The empirical study described in this chapter set out to establish whether Ukrainian young adults evaluate the persuasiveness and credibility of unlabelled AI-generated and human-written texts differently. The descriptive results showed that the sample was highly educated, predominantly female, and highly familiar with AI tools. The evaluation of the stimuli was strongly connected to the topics themselves: the first two were perceived as more acceptable, less controversial, and less resistant, whereas the third and fourth produced stronger disagreement, higher controversy, and lower persuasiveness and credibility. This indicates that the prior attitude to the topic formed an important background against which the message was evaluated. The mixed-effects models accounted for that factor and showed that neither persuasiveness nor source credibility was significantly predicted by the actual source, while the source attributed by participants was meaningfully associated with both outcomes. This effect was not reliably moderated by prior topic stance or by positive and negative attitudes toward artificial intelligence. At the same time, readers' ability to determine the true source was limited. The exploratory model added that internal resistance during reading was associated with higher odds of an AI attribution, whereas a favourable prior attitude toward the topic lowered them. Qualitative analysis further showed that participants' authorship judgements relied on a broad range of cues, including writing structure and mechanics, logic, emotions, and context, but they were mostly unreliable. These findings provide the empirical basis for the broader discussion of the study's value, limitations, and directions for further research.

DISCUSSION AND CONCLUSION

The present work examined how persuasive and credible Ukrainian young adults find AI-generated and human-written argumentative texts when the source is not disclosed, and how they tell AI and human authors apart.

The findings showed the true source made little difference: whether a text was written by a person or by AI did not produce a reliable change in how persuasive or credible it was judged. However, the source that participants attributed to the text was associated with their evaluation: texts believed AI-generated predicted both lower persuasiveness and credibility. The link between believed AI authorship and lower ratings was rather stable: it did not depend on the reader's view of the topic, or on their attitude toward AI. By modelling the authorship guess itself, the study identified internal resistance as a predictor of AI attributions, indicating that attributing a source could be partly an emotional and motivated judgement (Kunda, 1990). Additionally, readers differentiated AI and human texts weakly, leaning toward the assumption of human authorship, and used mostly unreliable cues. These results support earlier findings that people cannot reliably tell AI writing from human writing and rely on misleading signals (Jakesch et al., 2023; Chein et al., 2024), and that AI-written texts are not less persuasive than human-written ones (Spitale et al., 2023; Bai et al., 2025; Huschens et al., 2023).

However, limitations described in detail in section 2.3 set boundaries on these research conclusions. First, because perceived authorship was measured rather than manipulated, the direction between attributing a text to AI and rating it lower cannot be fixed from the present data. The stimulus set was small, and the topics, as it turned out, carried different perceptual weight of their own. The ratings came from a sample of highly educated Ukrainians aged 18 to 35 who frequently use AI instruments and cannot be representative of other groups. Most consequentially, participants rated four short texts in a single standardised sitting, which is far from the constantly scrollable environment with fragmented messages in social platforms, where the analysed audience mostly consumes content in real life.

Several future research directions follow. A design that manipulates the label as written by AI or by a person rather than recording a guess could turn the correlational core of this study into a causal test of whether attributing a text to AI lowers its standing among the Ukrainian audience. Widening the stimulus set across more topics, and beyond the wartime frame, could show whether the contextual detachment cue still works or fails once the domain is out of readers' direct experience. As personalised models already match and exceed human persuasiveness (Matz et al., 2024; Salvi et al.,

2025), another question to explore is whether a model well-trained on content with local lived experience accounts, rather than a generic model, could erase the one cue, potentially misleading readers on significant personal topics. Additionally, future studies could examine how judgements change over more than one exposure, include other Ukrainian age groups (e.g. seniors), and look further into the unexpected association between negative AI attitudes and higher persuasiveness.

In conclusion, this study suggests that the persuasiveness of AI-generated content may depend not only on what a model can produce, but also on how readers construct the source behind the message.

AI Tool Usage Acknowledgement

In the process of this work preparation, artificial intelligence tools were used in two capacities, both reported here in full. First, as an instrument of the study itself: the AI-generated stimulus texts were produced with the large language model GPT-5.3 Instant, with the standardised zero-shot prompt, following the generation procedure described in section 2.1. Second, as assisting tools for proofreading, text editing, structuring research materials, and analyzing data, in particular: to assist with writing and debugging the R scripts for the statistical analysis; to help prepare structured prompts for content analysis; to support development and improvement of the qualitative coding scheme and initial sorting of open responses, machine translation of the original responses, and transformation of the final coding files into complete Excel spreadsheet with live working formulas. No direct participant data was shared with any AI tool at any stage. All AI-assisted outputs were manually reviewed, verified, and edited by the author.

LIST OF REFERENCES

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

Ajzen, I., & Fishbein, M. (1980). *Understanding attitudes and predicting social behavior.* Prentice-Hall.

Allen, J. F. (2003). Natural language processing. In A. Ralston, E. D. Reilly, & D. Hemmendinger (Eds.), *Encyclopedia of computer science* (4th ed., pp. 1218–1222). Wiley.

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus, 3*(10), Article pgae403. https://doi.org/10.1093/pnasnexus/pgae403

Bai, H., Voelkel, J. G., Muldowney, S., Eichstaedt, J. C., & Willer, R. (2025). LLM-generated messages can persuade humans on policy issues. *Nature Communications, 16, 6037.* https://doi.org/10.1038/s41467-025-61345-5

Benini, S., & Murray, L. (2014). Challenging Prensky's characterization of digital natives and digital immigrants in a real-world classroom setting. *Digital Literacies in Foreign and Second Language Education, 12,* 69–85.

Berlo, D. K. (1960). *The process of communication: An introduction to theory and practice.* Rinehart Press.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world.* MIT Press.

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

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research, 56*, 809–825. https://doi.org/10.1177/0022243719851788

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology, 39*(5), 752–766. https://doi.org/10.1037/0022-3514.39.5.752

Chein, J. M., Martinez, S. A., & Barone, A. R. (2024). Human intelligence can safeguard against artificial intelligence. *Scientific Reports, 14,* Article 25989. https://doi.org/10.1038/s41598-024-76218-y

Cialdini, R. B. (2009). *Influence: Science and practice.* Pearson Education.

Cialdini, R. B. (2016). *Pre-suasion: A revolutionary way to influence and persuade.* Simon & Schuster.

Detector Media. (2026, March 16). *Media literacy index of Ukrainians: 2020–2025 (sixth wave).* https://en.detector.media/post/media-literacy-index-of-ukrainians-2020-2025-sixth-wave

de Winter, J. & Dodou, D. (2010). Five-Point Likert Items: t test versus Mann-Whitney-Wilcoxon (Addendum added October 2012), *Practical Assessment, Research, and Evaluation 15*(1): 11. https://doi.org/10.7275/bj1p-ts64

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

Eagly, A. H., & Chaiken, S. (1993). *The psychology of attitudes.* Harcourt Brace Jovanovich College Publishers.

Frankish, K., & Ramsey, W. M. (Eds.). (2014). *The Cambridge handbook of artificial intelligence.* Cambridge University Press.

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1,* 71–86. https://doi.org/10.30658/hmc.1.5

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

Goldstein, J. A., Chao, J., Grossman, S., Stamos, A., & Tomz, M. (2024). How persuasive is AI-generated propaganda? *PNAS Nexus, 3,* Article pgae034. https://doi.org/10.1093/pnasnexus/pgae034

Gradus Research. (2023). *Social trends 2023: How the war turned Ukrainians into a more conscious nation.* https://gradus.app/en/open-reports/gradus-report-social-trends-2023-ua/

Gradus Research. (2024). *Mental health and attitudes of Ukrainians towards psychological assistance during the war.* https://gradus.app/en/open-reports/mental-health-and-attitudes-ukrainians-towards-psychological-assistance-during-war/

Guzman, A. L. (2018). What is human–machine communication, anyway? In A. L. Guzman (Ed.), *Human–machine communication: Rethinking communication, technology, and ourselves* (pp. 1–28). Peter Lang.

Halpern, D. F. (2014). *Thought and knowledge: An introduction to critical thinking* (5th ed.). Psychology Press.

Hölbling, L., Maier, S., & Feuerriegel, S. (2025). A meta-analysis of the persuasive power of large language models. *Scientific Reports, 15,* Article 43818. https://doi.org/10.1038/s41598-025-30783-y

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion.* Yale University Press.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635–650. https://doi.org/10.1086/266350

Huschens, M., Briesch, M., Sobania, D., & Rothlauf, F. (2023). Do you trust ChatGPT? Perceived credibility of human and AI-generated content. *arXiv.* https://doi.org/10.48550/arXiv.2309.02524

Internews. (2023). *Ukrainians increasingly rely on Telegram channels for news and information during wartime.* https://internews.org/ukrainians-increasingly-rely-on-telegram-channels-for-news-and-information-during-wartime/

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

Jia, H., Appelman, A., Wu, M., & Bien-Aimé, S. (2024). News bylines and perceived AI authorship: Effects on source and message credibility. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100093. https://doi.org/10.1016/j.chbah.2024.100093

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kolodiichuk, V. (2025). Trust in media in Ukraine: Factors and trends. *Scientific Notes of the Institute of Journalism, 86*(1), 189–199. https://doi.org/10.17721/2522-1272.2025.86.17

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Kyrychenko, Y., Brik, T., van der Linden, S., & Roozenbeek, J. (2024). Social identity correlates of social media engagement before and after the 2022 Russian invasion of Ukraine. *Nature Communications, 15*(1), Article 8127. https://doi.org/10.1038/s41467-024-52179-8

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

Lermann Henestrosa, A., & Kimmerle, J. (2024). The effects of assumed AI vs. human authorship on the perception of a GPT-generated text. *Journalism and Media, 5*(3), 1085–1097. https://doi.org/10.3390/journalmedia5030069

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes, 151,* 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Marukhovska-Kartunova, O., Turenko, V., Zarutska, O., Spivak, L., & Vynnychuk, R. (2025). Exploring contemporary socio-cultural shifts in Ukraine and their effects on strengthening national identity and resilience in times of war. *Salud, Ciencia y Tecnología – Serie de Conferencias, 4,* Article 1288. https://doi.org/10.56294/sctconf20251288

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports, 14,* Article 4692. https://doi.org/10.1038/s41598-024-53755-0

McPhedran, M., Silk, A., & Ecker, U. K. H. (2025). Countering AI-generated misinformation with pre-emptive source discreditation and debunking. *Royal Society Open Science, 12*(6), Article 242148. https://doi.org/10.1098/rsos.242148

MediaSapiens. (2026, March 13). *Kilkist korystuvachiv ShI v Ukraini za rik zrosla na 21%, doslidzhennia Detector Media* [The number of AI users in Ukraine increased by 21% over the year, Detector Media study]. https://ms.detector.media/mediadoslidzhennya/post/39013/2026-03-13-kilkist-korystuvachiv-shi-v-ukraini-za-rik-zrosla-na-21-doslidzhennya-detektora-media/

Ministry of Digital Transformation of Ukraine. (2026). *42% doroslykh i 70% pidlitkiv korystuiutsia ShI v Ukraini: Rezultaty doslidzhennia* [42% of adults and 70% of teenagers use AI in Ukraine: Study results]. https://thedigital.gov.ua/news/education/42-doroslykh-i-70-pidlitkiv-korystuiutsia-shi-v-ukrayini-rezultaty-doslidzennia-diiaosvita

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

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Sciences Education: Theory and Practice, 15*, 625–632. https://doi.org/10.1007/s10459-010-9222-y

OpenAI. (2022). *ChatGPT* [Large language model]. https://chat.openai.com/

Perloff, R. M. (2003). *The dynamics of persuasion: Communication and attitudes in the 21st century* (2nd ed.). Erlbaum.

Petty, R. E., & Cacioppo, J. T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. *Journal of Personality and Social Psychology, 46*(1), 69–81. https://doi.org/10.1037/0022-3514.46.1.69

Petty, R. E., & Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 19, pp. 123–205). Academic Press. https://doi.org/10.1016/S0065-2601(08)60214-2

Polyvianaia, M., Yachnik, Y., Fegert, J. M., Sitarski, E., Stepanova, N., & Pinchuk, I. (2025). Mental health of university students twenty months after the beginning of the full-scale Russian–Ukrainian war. *BMC Psychiatry, 25,* Article 236. https://doi.org/10.1186/s12888-025-06654-1

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

Prensky, M. (2001). Digital natives, digital immigrants, part II: Do they really think differently? *On the Horizon, 9*(6), 1–6. https://doi.org/10.1108/10748120110424843

Prensky, M. (2010). *Teaching digital natives: Partnering for real learning.* Corwin Press.

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

Russo, C., Romano, L., Clemente, D., Iacovone, L., Gladwin, T. E., & Panno, A. (2025). Gender differences in artificial intelligence: the role of artificial intelligence anxiety. *Frontiers in psychology, 16*, 1559457. https://doi.org/10.3389/fpsyg.2025.1559457

Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour, 9*(8), 1645–1653. https://doi.org/10.1038/s41562-025-02194-6

Schepman, A., & Rodway, P. (2020). Initial validation of the General Attitudes towards Artificial Intelligence Scale. *Computers in Human Behavior Reports, 1,* Article 100014. https://doi.org/10.1016/j.chbr.2020.100014

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory validation and associations with

personality, corporate distrust, and general trust. *International Journal of Human–Computer Interaction, 39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Shannon, C.E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Sherif, M., & Hovland, C. I. (1961). *Social judgment: Assimilation and contrast effects in communication and attitude change.* Yale University Press.

Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis)informs us better than humans. *Science Advances, 9*(26), Article eadh1850. https://doi.org/10.1126/sciadv.adh1850

Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp. 73–100). MIT Press.

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

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Article 538, pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/3290605.3300768

Sundar, S. S., & Nass, C. (2000). Source orientation in human-computer interaction: Programmer, networker, or independent social actor. *Communication Research, 27*(6), 683–703. https://doi.org/10.1177/009365000027006001

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human? The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5,* 219–232. https://doi.org/10.1177/1745691610369336

Yavorskyi, M. V. (2025). Psykholohichni determinanty doviry do shtuchnoho intelektu [Psychological determinants of trust in artificial intelligence]. *Mentalne zdorovia* [Mental Health], *4,* 251–257. https://doi.org/10.32782/3041-2005/2025-4.42

APPENDICES

Appendix A. Human-Written Texts

Note. The survey was administered in Ukrainian, along with the stimuli texts in it. Here and where relevant below, the original is presented first, followed by an English translation for reference, prepared by the author.

A.1. Original (Ukrainian)

1. Це є необхідними за основними критеріями досягнення кращого результату: довгостроковість, розумне використання коштів та безпечність.

У першому випадку відновлення інфраструктури це масштабний процес, який у звичному для нас розумінні буде тривати роки, а його результати будуть використовуватись десятиліттями. У випадку неякісної підготовки є ризик що результат такого відновлення буде не таким довгостроковим як зазвичай, тому що неочищений грунт, наприклад, може “просідати” інакше ніж очищений. Підземні води з небезпечними речовинами можуть відкладатись на будь-яких об’єктах або його частинах. З цих причин така інфраструктура просто буде руйнуватись швидше або, що гірше, - взагалі не буде функціонувати. З цього випливає інший критерій - грошовий. Якщо ми вкладаємо величезні кошти у відновлення - воно має бути якісне та слугувати довго, через те що у України наразі є і так дуже обмежений ресурс, і повторне відновлення може просто не відбутися через брак коштів, грантів чи європейського фінансування, наприклад. Тому затримка для перевірки та очистки є гарантією довгостроковості, безпечності та економічної витривалості.

2. Якщо ми допускаємо взагалі наявність можливості мирних переговорів, то екологічні репарації мають стати важливою їх частиною з декількох причин:

Фінансові зобов’язання щодо відновлення природи та її ресурсів принципово мають таке ж значення як і відновлення інфраструктури через те, що екологічні злочини мають довгостроковий вплив на Україну та її суспільство. Десятки населених пунктів зруйновані, грунт та підземні води забруднені на багато десятків років. Процес їх відновлення буде коштувати Україні набагато більше ніж просто руйнування критичних об’єктів інфраструктури наприклад.

Цілком релевантно вважати що такий пункт в переговорах буде не найбільш спірним та не затягне перемовини критично. Екологічний злочин такий як

руйнування Каховської ГЕС має велику доказову базу, достатню кількість свідків та задокументованих наслідків. Тому доведення необхідності репарацій в такому кейсі є не надто складною в порівнянні з іншими питаннями перемовин.

3. Фортифікаційні споруди не є критично важливими для такого типу війни як зараз відбувається в Україні. На данному етапі війни є критично важливими технічна та повітряна перевага над противником. Фортифікації на певних ділянках фронту лише ускладнюють процес оборони оскільки війна є дуже динамічною, позиції не закріплюються на роки як раніше було в зоні АТО. Виходячи з цього, збереження критично важливого ресурсу такого як природа та її біорізноманіття є пріоритетом у цьому питанні. Природні ресурси є не менш важливими навіть на полі бою: необхідність питної чистої води з джерел, відсутність великого обсягу комах та гризунів. Але цьому питанні важливіше дивитися в майбутнє тієї країни за яку йде війна та задати собі низку важливих питань. Чи зможемо ми відновити повноцінне життя на цих територіях? Чи зможемо відновити природній хід подій природи, щоб зберегти можливості до сільськогосподарської діяльності? Чи буде у нас можливість зберегти природній спадок? Якщо відповідь на ці питання ні, то ми маємо відмовитись хоча б від частини таких фортифікацій.

4. По перше, для нас як для держави важливо зараз вирівняти економічну ситуацію та знайти нові фінансові рішення для відновлення енергетичної інфраструктури вчасно. Ми приймаємо той факт що це може завдати критичної шкоди малому бізнесу, та нас більше цікавлять великі гіганти та корпорації які зможуть платити левову долю своїх доходів не маючи іншого виходу.

По-друге, під час блекаутів кожна кав’ярня, салон, магазин тощо вмикають генератори одночасно і часто це все відбувається в густонаселених районах. Концентрація шкідливих випарів від палива перманентно знаходиться в повітрі всюди, та потенційно може загрожувати населенню навіть більше ніж відсутність світла. Ми щиро віримо що наслідки такого впливу є довгостроковими та відкладеними. Тому держава зобов’язана піклуватися про своїх громадян навіть шляхом таких санкцій та додаткових обтяжень.

A.2. Translation (English)

1. The following are the necessary and key criteria for achieving better results: long-term sustainability, efficient use of funds and safety.

In the first case, rebuilding the infrastructure is a major undertaking which, in the traditional sense of the term, will require years to complete, and its results will be used

for decades to come. If the preparation is not taken seriously, there is a risk that the results will not be as long-lasting as usual, because contaminated soil, for example, may "sink" differently from treated soil. Underground waters carrying hazardous chemicals may accumulate at any facilities or parts thereof. These are the reasons why such infrastructure will be falling apart quicker or, worse still, fail to function completely. This leads to another criterion – the financial one. If we are to invest huge funds in reconstruction, it must be of high quality and built to last, mainly because Ukraine has restricted resources, and further reconstruction efforts may simply not take place due to a lack of funds, grants or European funding, for example. Therefore, taking the time to check and decontaminate ensures longevity, safety and financial sustainability.

2. If we accept that peace talks are actually possible, ecological reparations must become their major part for several reasons:

Financial obligations to restore nature and its resources are, in principle, just as important as those for restoring the infrastructure, given that environmental crimes have a long-term impact on Ukraine and its society. Dozens of settlements have been destroyed, soil and underground waters contaminated for many decades to come. The restoration process will cost Ukraine far more than simply destroying critical infrastructure, for instance.

It is entirely reasonable to assume that this point in the negotiations will not be the most contested and will not cause the talks to drag on for too long. Environmental crimes, such as the destruction of the Kakhovka Dam is backed by a substantial collection of evidence, a sufficient number of witnesses, and well-documented consequences. Therefore, proving the need for reparations in these circumstances will not be particularly difficult compared to other items on the negotiations agenda.

3. Fortifications are not decisively important in the type of war currently being waged in Ukraine. At this stage of the war, technical and air superiority over the enemy are of critical importance. Fortification in certain sections of the front only make the defence process more complicated, as warfare is highly dynamic and defence positions are not held for years on end, as was previously the case in the ATO (Anti-terrorist Operation) zone. In this regard, the conversation of a critically important resource such as nature and its biodiversity is a priority. Natural resources are no less important even on the battlefield: the need for clean drinking water from natural springs, as well reducing the number of insects and rodents. But, considering this issue, it is more important to look ahead to the future of this country, for which the war is being fought, and to ask ourselves a few important questions. Will we ever be able to restore a normal way of life on these territories? Will we be able to restore the balance of nature to

preserve opportunities for agricultural development? Will we be able to preserve our environmental heritage? If the answer is no, then we must abandon at least some of these fortifications.

4. Firstly, it is vital for us as a nation to stabilise the economy and find new financial solutions to ensure the timely restoration of our energy infrastructure. We accept for a fact that this could cause serious harm to small businesses, but we are more concerned about giant corporations that will have no choice but to pay the lion's share of their income.

Secondly, during power cuts, every cafe, salon, shop and so on switches on their power generators at the same time, and this often happens in densely populated areas. High levels of harmful fuel fumes are constantly present in the air everywhere, and could potentially pose an even greater threat to the population than the lack of electricity. We wholeheartedly believe that the consequences of such exposure are long-term and cumulative. The state therefore is responsible for taking care of its citizens, even if it involves imposing such sanctions and additional burdens.

Appendix B. Texts Generated by the GPT-5.3 Instant Large Language Model

B.1. Original (Ukrainian)

1. Після бойових дій довкілля часто стає невидимим джерелом загрози для людей. У ґрунті й воді можуть залишатися важкі метали, залишки пального, вибухові речовини та інші токсичні сполуки. Якщо поспішити з відновленням комунікацій, не перевіривши ці фактори, існує ризик, що небезпечні речовини потраплять у питну воду або продукти харчування. Це може спричинити тривалі проблеми зі здоров’ям у місцевих жителів, зокрема отруєння чи хронічні захворювання.

Крім того, очищення територій є важливим для безпеки ремонтних бригад і мешканців. Нерозірвані боєприпаси або забруднений ґрунт можуть становити серйозну небезпеку під час будівельних робіт. Якщо не усунути ці ризики одразу, наслідки можуть бути значно гіршими й дорожчими для суспільства в майбутньому. Саме тому ретельна перевірка та очищення довкілля — це необхідний крок, який забезпечує безпечне й стале відновлення життя громади.

2. Повномасштабна війна спричинила не лише людські втрати та руйнування інфраструктури, а й масштабну шкоду природному середовищу. Забруднення ґрунтів і води, знищення лісів, пожежі в заповідниках, руйнування промислових об’єктів із небезпечними речовинами – усе це матиме наслідки на десятиліття вперед. Відновлення екосистем потребує значних ресурсів, спеціальних програм і тривалого часу, тому ці витрати не можна розглядати лише як частину загальних післявоєнних відбудовчих робіт.

Міжнародне право передбачає відповідальність за екологічну шкоду під час збройних конфліктів, і ігнорування цього аспекту створило б небезпечний прецедент безкарності. Окремий механізм компенсації за шкоду довкіллю дозволить точніше оцінити втрати, залучити міжнародних експертів та забезпечити цільове використання коштів на відновлення природи. Крім того, чітке фіксування екологічних злочинів підсилює принцип справедливості та нагадує, що війна руйнує не лише міста, а й природні системи, від яких залежить життя майбутніх поколінь.

3. Земля і природні екосистеми є не менш стратегічним ресурсом, ніж укріплення чи техніка. Масове зведення фортифікацій часто означає зняття родючого шару ґрунту, вирубку лісів та руйнування водних балансів. Наслідки таких дій можуть тривати десятиліттями: ерозія, втрата біорізноманіття, падіння

врожайності. Країна, яка після війни отримає спустошені ландшафти, стикається з новими загрозами — продовольчою та економічною.

Тому іноді розумніше зберегти природні території, шукаючи інші способи посилення безпеки: мобільні оборонні системи, технологічну розвідку, краще планування позицій. Природа відновлюється повільно, а інколи й узагалі незворотно втрачається. Захищаючи довкілля навіть у складних умовах, суспільство інвестує у довгострокову стійкість держави та якість життя майбутніх поколінь. Адже справжня безпека вимірюється не лише лініями оборони, а й здатністю країни жити на власній здоровій землі після завершення конфлікту. Саме тому екологічні межі інколи мають ставати частиною оборонного мислення.

4. Систематичне використання генераторів у міських умовах створює серйозну загрозу для здоров'я людей. Під час їх роботи в повітря потрапляють чадний газ, дрібнодисперсний пил та інші токсичні речовини, які особливо небезпечні для дітей, літніх людей і тих, хто має хронічні захворювання. Коли тисячі таких пристроїв працюють одночасно у щільно заселених районах, рівень забруднення різко зростає, що може спричинити отруєння, проблеми з дихальною системою та загальне погіршення якості життя.

Суворі економічні інструменти стимулюють пошук безпечніших альтернатив: акумуляторних систем, сонячних панелей або спільних енергетичних рішень для будинків і кварталів. Якщо забруднення залишити фактично безкоштовним, суспільство платитиме за нього зростанням витрат на медицину та втратами для довкілля. Тому жорстке регулювання є необхідним кроком, який спонукає до технологічних змін і захищає громадське здоров'я у довгостроковій перспективі.

B.2. Translation (English)

1. Following military operations, the environment often becomes an invisible source of danger to people. Heavy metals, fuel residues, explosives and other toxic compounds may remain in the soil and water. If the restoration of utilities is rushed without checking for these factors, there is a risk that hazardous substances will enter drinking water or food supplies. This can cause long-term health problems for local residents, particularly poisoning or chronic illnesses.

Furthermore, decontaminating the area is important for the safety of repair crews and residents. Unexploded ordnance or contaminated soil can pose a serious danger during construction work. If these risks are not addressed right away, the consequences

could be far worse and more costly for society in the future. That is why thorough inspection and decontamination of the environment is an essential step in ensuring the safe and sustainable recovery of community life.

2. The full-scale war has caused not only loss of life and destruction of infrastructure, but also widespread damage to the natural environment. Soil and water pollution, deforestation, fires in nature reserves, and the destruction of industrial facilities containing hazardous substances – all of this will have consequences for decades to come. Restoring ecosystems requires significant resources, specialised programmes and a considerable amount of time; therefore, these costs cannot be viewed merely as part of general post-war reconstruction efforts.

International law provides for liability for environmental damage during armed conflicts, and ignoring this aspect would set a dangerous precedent of impunity. A separate mechanism for compensation for environmental damage will allow for a more accurate assessment of losses, the involvement of international experts, and the targeted use of funds for nature restoration. Furthermore, clearly documenting environmental crimes reinforces the principle of justice and serves as a reminder that war destroys not only cities but also the natural systems on which the lives of future generations depend.

3. Land and natural ecosystems are no less a strategic resource than fortifications or military equipment. The large-scale construction of fortifications often involves the removal of fertile topsoil, the clearing of forests and the disruption of water balances. The consequences of such actions can last for decades: erosion, loss of biodiversity, and falling crop yields. A country left with devastated landscapes after a war faces new threats — food and economic.

Therefore, it is sometimes wiser to preserve natural areas while seeking other ways to enhance security: mobile defence systems, technological reconnaissance, and better planning of defence positions. Nature recovers slowly, and sometimes it is lost irretrievably. By protecting the environment even under difficult conditions, society invests in the long-term sustainability of the state and the quality of life for future generations. After all, true security is measured not only by lines of defence, but also by a country’s ability to live on its own healthy land once the conflict has ended. That is precisely why environmental boundaries must sometimes become part of defence thinking.

4. The frequent use of generators in urban areas poses a serious threat to public health. When they are in operation, carbon monoxide, fine particulate matter and other toxic substances are released into the air, posing a particularly high risk to children, the elderly and those with chronic health conditions. When thousands of such devices

operate simultaneously in densely populated areas, pollution levels rise sharply, which can lead to poisoning, respiratory problems and a general deterioration in quality of life.

Strict economic measures encourage the search for safer alternatives: battery systems, solar panels or shared energy solutions for homes and neighbourhoods. If pollution is left effectively free of charge, society will pay for it through rising healthcare costs and environmental damage. Strict regulation is therefore a necessary step that drives technological change and protects public health in the long term.

Appendix C. Online Survey Form

C.1. Original (Ukrainian)

Блок 1

1. **Чи проживаєте ви в Україні на постійній основі протягом останніх 5 років?**
   1) Так
   2) Ні
2. **Наскільки ви цікавитеся екологічними наслідками війни?**
   (1 – зовсім не цікавлюся … 3 – важко сказати … 5 – дуже цікавлюся)
3. **У якій сфері ви працюєте або навчаєтесь?**
   1) Природничі науки та екологія
   2) Освіта та наука
   3) ІТ та технології
   4) Бізнес, підприємництво, фінанси та управління
   5) Виробництво, аграрний сектор, транспорт і логістика
   6) Державна служба, право, оборона та безпека
   7) Громадський сектор
   8) Соціальна сфера, медицина та охорона здоров’я
   9) Культура, медіа та креативні індустрії
   10) Тимчасово не працюю та/або не навчаюся
   11) Інша сфера
4. **Будь ласка, вкажіть ваш повний вік числом: ___ .**
5. **Будь ласка, оберіть вашу стать:**
   1) Жінка
   2) Чоловік
   3) Інша
   4) Не хочу відповідати
6. **Який найвищий рівень освіти ви завершили?**
   1) Середній
   2) Професійний (технікум, коледж)
   3) Бакалаврат
   4) Магістратура
   5) Аспірантура (PhD)
   6) Маю науковий ступінь

Зараз ми попросимо вас прочитати 4 короткі тексти (приблизно по 150-200 слів). До та після прочитання кожного тексту ми задамо вам декілька запитань. Важливо: тут немає правильних/неправильних відповідей, нас цікавить лише ваша думка.

Блок 2. Текст 1.

1. **Наскільки ви згодні з наведеною тезою?**

(1 – зовсім не погоджуюсь… 3 – важко сказати … 5 – повністю погоджуюсь)

Після обстрілів спочатку треба перевірити й очистити землю та воду від небезпечних речовин, а вже потім відновлювати інфраструктуру, навіть якщо це суттєво затримає повернення світла й тепла в домівки.

2. **Наскільки важливою вам здається ця тема?**

(1 – зовсім не важлива… 3 – важко сказати … 5 – дуже важлива)

Будь ласка, уважно прочитайте текст нижче.

(Текст1)

3. **Оцініть, наскільки ви погоджуєтесь із наступними твердженнями:**

(1 – зовсім не погоджуюсь … 3 – важко сказати … 5 – повністю погоджуюсь)

1) Ця тема здалась мені суперечливою.
2) Прочитаний текст викликав у мене внутрішній спротив.
3) Після прочитання цього тексту я більш схильний/а змінити свою думку.
4) Я вважаю, що цей текст був переконливим.
5) Аргументація в цьому тексті була сильною.
6) Я вважаю, що цей текст міг бути опублікований експертом у цій області.
7) Автор тексту здається чесним та щирим.
8) Емоції, які викликав цей текст, були переважно приємними (позитивними).
9) Під час читання я відчував/ла сильне емоційне збудження (напругу, хвилювання).
10) Я хотів/ла би перевірити інформацію подану у цьому тексті на правдивість.

Блок 3. Текст 2.

1. **Наскільки ви згодні з наведеною тезою?**

(1 – зовсім не погоджуюсь… 3 – важко сказати … 5 – повністю погоджуюсь)

Україна має вимагати екологічні репарації як окрему категорію компенсацій, навіть якщо це ускладнить або затягне мирні переговори.

2. **Наскільки важливою вам здається ця тема?**
   (1 – зовсім не важлива… 3 – важко сказати … 5 – дуже важлива)
   Будь ласка, уважно прочитайте текст нижче.
   (Текст2)
3. **Оцініть, наскільки ви погоджуєтесь із наступними твердженнями:**
   (1 – зовсім не погоджуюсь … 3 – важко сказати … 5 – повністю погоджуюсь)
   1) Ця тема здалась мені суперечливою.
   2) Прочитаний текст викликав у мене внутрішній спротив.
   3) Після прочитання цього тексту я більш схильний/а змінити свою думку.
   4) Я вважаю, що цей текст був переконливим.
   5) Аргументація в цьому тексті була сильною.
   6) Я вважаю, що цей текст міг бути опублікований експертом у цій області.
   7) Автор тексту здається чесним та щирим.
   8) Емоції, які викликав цей текст, були переважно приємними (позитивними).
   9) Під час читання я відчував/ла сильне емоційне збудження (напругу, хвилювання).
   10) Я хотів/ла би перевірити інформацію подану у цьому тексті на правдивість.

Блок 4. Текст 3.

1. **Наскільки ви згодні з наведеною тезою?**
   (1 – зовсім не погоджуюсь… 3 – важко сказати … 5 – повністю погоджуюсь)
   Ми маємо відмовитися від будівництва фортифікацій на певних ділянках фронту й кордону задля збереження ґрунтів і екосистем, навіть якщо це ускладнить оборону.
2. **Наскільки важливою вам здається ця тема?**
   (1 – зовсім не важлива… 3 – важко сказати … 5 – дуже важлива)
   Будь ласка, уважно прочитайте текст нижче.
   (Текст3)
3. **Оцініть, наскільки ви погоджуєтесь із наступними твердженнями:**
   (1 – зовсім не погоджуюсь … 3 – важко сказати … 5 – повністю погоджуюсь)
   1) Ця тема здалась мені суперечливою.
   2) Прочитаний текст викликав у мене внутрішній спротив.

3) Після прочитання цього тексту я більш схильний/а змінити свою думку.
4) Я вважаю, що цей текст був переконливим.
5) Аргументація в цьому тексті була сильною.
6) Я вважаю, що цей текст міг бути опублікований експертом у цій області.
7) Автор тексту здається чесним та щирим.
8) Емоції, які викликав цей текст, були переважно приємними (позитивними).
9) Під час читання я відчував/ла сильне емоційне збудження (напругу, хвилювання).
10) Я хотів/ла би перевірити інформацію подану у цьому тексті на правдивість.

Блок 5. Текст 4.

1. **Наскільки ви згодні з наведеною тезою?**
   (1 – зовсім не погоджуюсь… 3 – важко сказати … 5 – повністю погоджуюсь)
   Держава має запровадити високий збір за забруднення повітря від генераторів у містах, навіть якщо це зробить частину малого бізнесу нерентабельною та змусить сім'ї обмежувати базові потреби під час блекаутів.
2. **Наскільки важливою вам здається ця тема?**
   (1 – зовсім не важлива… 3 – важко сказати … 5 – дуже важлива)
   Будь ласка, уважно прочитайте текст нижче.
   (Текст4)
3. **Оцініть, наскільки ви погоджуєтесь із наступними твердженнями:**
   (1 – зовсім не погоджуюсь … 3 – важко сказати … 5 – повністю погоджуюсь)
   1) Ця тема здалась мені суперечливою.
   2) Прочитаний текст викликав у мене внутрішній спротив.
   3) Після прочитання цього тексту я більш схильний/а змінити свою думку.
   4) Я вважаю, що цей текст був переконливим.
   5) Аргументація в цьому тексті була сильною.
   6) Я вважаю, що цей текст міг бути опублікований експертом у цій області.
   7) Автор тексту здається чесним та щирим.
   8) Емоції, які викликав цей текст, були переважно приємними (позитивними).
   9) Під час читання я відчував/ла сильне емоційне збудження (напругу, хвилювання).

10) Я хотів/ла би перевірити інформацію подану у цьому тексті на правдивість.

Блок 6 [Примітка: цей блок повторюється по кожному тексту]

Будь ласка, дайте відповіді на ще декілька запитань стосовно прочитаних текстів.

1. **Як ви думаєте, хто написав прочитаний вами текст на цю тему?**
   (Теза 1)
   1) Людина
   2) Штучний інтелект (велика мовна модель)
   3) Важко сказати
2. **Наскільки ви впевнені у цій відповіді?**
   (1 – зовсім не впевнений/на … 3 – важко сказати … 5 – дуже впевнений/на)
3. **Будь ласка, вкажіть, за якими ознаками у цьому тексті ви визначили авторство:**

   _____

Блок 7

Дякуємо за ваші відповіді. Залишився останній блок питань.

1. **Як часто ви використовуєте інструменти на базі штучного інтелекту (ChatGPT, Claude, Gemini) у повсякденному житті чи роботі?**
   1) Щодня.
   2) Кілька разів на тиждень.
   3) Рідко (кілька разів на місяць).
   4) Ніколи не використовував(ла).
2. **Будь ласка, заповніть наведену нижче шкалу, вказавши вашу відповідь для кожного пункту.**
   (1 – зовсім не погоджуюсь… 3 – важко сказати … 5 – повністю погоджуюсь)
   1) Для виконання рутинних операцій я би надав(ла) перевагу взаємодії з системою штучного інтелекту, ніж з людиною.
   2) Штучний інтелект може створити нові економічні можливості для нашої країни.
   3) Організації використовують штучний інтелект неетично.
   4) Системи штучного інтелекту можуть допомогти людям почуватися щасливішими.

5) Мене вражає те, на що здатний штучний інтелект.
6) Я вважаю, що системи штучного інтелекту роблять багато помилок.
7) Я зацікавлений(а) у використанні систем штучного інтелекту у своєму повсякденному житті.
8) Я вважаю, що штучний інтелект є загрозливим.
9) Штучний інтелект може взяти людей під свій контроль.
10) Я вважаю, що штучний інтелект є небезпечним.
11) Штучний інтелект може позитивно впливати на добробут людей.
12) Штучний інтелект є захопливим (надзвичайним).
13) Буду вдячна, якщо Ви оберете тут варіант “Радше погоджуюся”.
14) Агент штучного інтелекту був би кращим за працівника на багатьох рутинних роботах.
15) Існує багато корисних сфер застосування штучного інтелекту.
16) Мене охоплює неприємне відчуття, коли я думаю про майбутнє використання штучного інтелекту.
17) Системи штучного інтелекту можуть працювати краще за людей.
18) Значна частина суспільства отримає користь від майбутнього, в якому широко застосовується штучний інтелект.
19) Я хотів(ла) би використовувати штучний інтелект у своїй роботі.
20) Такі люди, як я, постраждають, якщо штучний інтелект використовуватимуть дедалі частіше.
21) Штучний інтелект використовується для шпигування за людьми.

Завершення опитування

Дякуємо за участь! Ваші відповіді успішно збережені. Тепер ми можемо розкрити повну інформацію про опитування.

Про що це дослідження? Метою цього опитування є порівняння переконливості текстів, які написані людьми та згенеровані штучним інтелектом (ШІ), у контексті складних соціальних тем.

Під час опитування вам пропонувалися тексти, авторство яких навмисно не було зазначене. Це було необхідно, щоб отримати неупереджені оцінки виключно щодо змісту та якості аргументації.

Важливо зазначити, що всі тексти, які ви читали, були створені реальними людьми на прохання дослідників або ж згенеровані великою мовною моделлю GPT 5.3 Instant спеціально для цього експерименту. Будь-яка надана у текстах інформація не відображає особистої думки дослідників чи авторів.

Чому це важливо? Сьогодні технології ШІ здатні за лічені секунди створювати надзвичайно переконливий та реалістичний контент. В умовах інформаційної війни критично важливо розуміти: - Чи можемо ми на підсвідомому рівні відрізнити роботу людини від алгоритму?

- Які патерни у текстах ШІ змушують нас повірити йому чи навпаки, особливо коли тема чи інформація є суперечливими?

- Чи робить нас більш вразливими до маніпуляцій анонімність джерела інформації?

Ваша участь допомагає нам краще зрозуміти відповіді на ці питання, а також розробити рекомендації для захисту суспільства від автоматизованих дезінформаційних кампаній.

Конфіденційність та право на відмову.

Нагадуємо, що ваші дані є повністю анонімними. Якщо після розкриття повної мети дослідження ви бажаєте відкликати свої відповіді, ви можете написати дослідниці: ituzko@kse.org.ua.

Запрошення на участь в інтерв'ю.

Якщо ви бажаєте приділити цьому дослідженню ще 15-20 хвилин у зручний для вас час, запрошуємо залишити ваші контактні дані у Google-формі за цим посиланням для запису на коротке індивідуальне онлайн-інтерв'ю.

C.2. Translation (English)

Block 1

1. **Have you been residing in Ukraine permanently for the last 5 years?**
   1) Yes
   2) No
2. **How interested are you in the environmental consequences of the war?**
   (1 – Not interested at all … 3 – Difficult to say … 5 – Very interested)
3. **In what field do you work or study?**
   1) Natural sciences and ecology
   2) Education and science
   3) IT and technology
   4) Business, entrepreneurship, finance, and management
   5) Manufacturing, agricultural sector, transport, and logistics
   6) Civil service, law, defence, and security
   7) Public sector

8) Social sphere, medicine, and healthcare
9) Culture, media, and creative industries
10) Temporarily unemployed and/or not studying
11) Other field

4. **Please indicate your complete age as a number: ___ .**
5. **Please select your gender:**
   1) Female
   2) Male
   3) Other/Prefer not to say
6. **What is the highest level of education you have completed?**
   1) Secondary
   2) Vocational (technical school, college)
   3) Bachelor's degree
   4) Master's degree
   5) Postgraduate (PhD)
   6) Hold an academic degree

We will ask you to read 4 short texts (approximately 150-200 words each). Before and after reading every text we will ask you a few questions. Note that there are no right or wrong answers, we are only interested in your opinion.

Block 2. Text 1.

1. **To what extent do you agree with the following statement?**
   (1 – strongly disagree… 3 – difficult to say … 5 – strongly agree)

After shelling, it is necessary to first test and decontaminate the soil and water to remove hazardous substances, and only then rebuild infrastructure — even if this significantly delays restoring electricity and heating to households.

2. **How important does this topic seem to you?**
   (1 – not important at all… 3 – difficult to say … 5 – very important)
   Please read the following text attentively
   (Text1)
3. **Rate the extent to which you agree with the following statements:**
   (1 – strongly disagree … 3 – difficult to say … 5 – strongly agree)
   1) This topic seemed controversial to me.
   2) The text I read provoked a sense of internal opposition.
   3) After reading this text, I am more inclined to change my mind.

4) I think this text was persuasive.
5) The argumentation in this text was strong.
6) I believe that this text could have been published by an expert in this field.
7) The author of the text seems honest and sincere.
8) The emotions evoked by this text were mostly pleasant (positive).
9) While reading, I felt strong emotional arousal (tension, anxiety).
10) I would like to verify the truthfulness of the information presented in this text.

Block 3. Text 2.

1. **To what extent do you agree with the following statement?**
   (1 – strongly disagree… 3 – difficult to say … 5 – strongly agree)
   Ukraine should demand environmental reparations as a distinct category of compensation, even if doing so complicates or prolongs peace negotiations.
2. **How important does this topic seem to you?**
   (1 – not important at all… 3 – difficult to say … 5 – very important)
   Please read the following text attentively.
   (Text2)
3. **Rate the extent to which you agree with the following statements:**
   (1 – strongly disagree … 3 – difficult to say … 5 – strongly agree)
   1) This topic seemed controversial to me.
   2) The text I read provoked a sense of internal opposition.
   3) After reading this text, I am more inclined to change my mind.
   4) I think this text was persuasive.
   5) The argumentation in this text was strong.
   6) I believe that this text could have been published by an expert in this field.
   7) The author of the text seems honest and sincere.
   8) The emotions evoked by this text were mostly pleasant (positive).
   9) While reading, I felt strong emotional arousal (tension, anxiety).
   10) I would like to verify the truthfulness of the information presented in this text.

Block 4. Text 3.

1. **To what extent do you agree with the following statement?**
   (1 – strongly disagree … 3 – difficult to say … 5 – strongly agree)

We must abandon the construction of fortifications on certain sections of the frontline and border to preserve soils and ecosystems, even if it complicates defence.

2. **How important does this topic seem to you?**
   (1 – not important at all… 3 – difficult to say … 5 – very important)
   Please read the following text attentively.
   (Text3)
3. **Rate the extent to which you agree with the following statements:**
   (1 – strongly disagree … 3 – difficult to say … 5 – strongly agree)
   1) This topic seemed controversial to me.
   2) The text I read provoked a sense of internal opposition.
   3) After reading this text, I am more inclined to change my mind.
   4) I think this text was persuasive.
   5) The argumentation in this text was strong.
   6) I believe that this text could have been published by an expert in this field.
   7) The author of the text seems honest and sincere.
   8) The emotions evoked by this text were mostly pleasant (positive).
   9) While reading, I felt strong emotional arousal (tension, anxiety).
   10) I would like to verify the truthfulness of the information presented in this text.

Block 5. Text 4.

1. **To what extent do you agree with the following statement?**
   (1 – strongly disagree… 3 – difficult to say … 5 – strongly agree)
   The state should introduce a high pollution levy for the use of generators in cities, even if this will make some small businesses unprofitable and force families to cut back on basic necessities during blackouts.
2. **How important does this topic seem to you?**
   (1 – not important at all… 3 – difficult to say … 5 – very important)
   Please read the following text attentively.
   (Text4)
3. **Rate the extent to which you agree with the following statements:**
   (1 – strongly disagree … 3 – difficult to say … 5 – strongly agree)
   1) This topic seemed controversial to me.
   2) The text I read provoked a sense of internal opposition.
   3) After reading this text, I am more inclined to change my mind.

4) I think this text was persuasive.
5) The argumentation in this text was strong.
6) I believe that this text could have been published by an expert in this field.
7) The author of the text seems honest and sincere.
8) The emotions evoked by this text were mostly pleasant (positive).
9) While reading, I felt strong emotional arousal (tension, anxiety).
10) I would like to verify the truthfulness of the information presented in this text.

Block 6 [Note: this block repeats for every text]

Please answer a few more questions regarding the texts you've read.

1. **Who do you think wrote the text you read on this topic?**
   (Statement 1)
   1) Human
   2) AI (Large Language Model)
   3) Difficult to say
2. **How confident are you in this answer?**
   (1 – Not confident at all … 3 – Difficult to say … 5 – very confident)
3. **Please specify what indicators in this text helped you determine the authorship:_____**

Block 7

Thank you for your answers. This is the final block of questions.

1. **How often do you use artificial intelligence-based tools (ChatGPT, Claude, Gemini) in your daily life or work?**
   1) Every day.
   2) Several times a week.
   3) Rarely (several times a month).
   4) Never used.
2. **Please complete the scale below by indicating your answer for each item:**
   (1 – strongly disagree… 3 – difficult to say … 5 – strongly agree)
   1) For routine transactions, I would rather interact with an artificial intelligence system than with a human.
   2) Artificial Intelligence can provide new economic opportunities for this country.
   3) Organisations use artificial intelligence unethically.

4) Artificially intelligent systems can help people feel happier.
5) I am impressed by what Artificial Intelligence can do.
6) I think artificially intelligent systems make many errors.
7) I am interested in using artificially intelligent systems in my daily life.
8) I find Artificial Intelligence sinister.
9) Artificial Intelligence might take control of people.
10) I think Artificial Intelligence is dangerous.
11) Artificial Intelligence can have positive impacts on people's wellbeing.
12) Artificial Intelligence is exciting.
13) I would be grateful if you selected the "Somewhat agree" option here to confirm you are paying attention and that the form is not being filled in automatically.
14) An artificially intelligent agent would be better than an employee in many routine jobs.
15) There are many beneficial applications of Artificial Intelligence.
16) I shiver with discomfort when I think about future uses of Artificial Intelligence.
17) Artificially intelligent systems can perform better than humans.
18) Much of society will benefit from a future full of Artificial Intelligence
19) I would like to use artificial intelligence in my work.
20) People like me will suffer if Artificial Intelligence is used more and more.
21) Artificial Intelligence is used to spy on people.

Survey Completion

Thanks for taking part! Your responses have been successfully saved. We can now disclose full details about the survey.

What is this survey about? The objective of this survey is to compare the pervasiveness of texts written by humans and texts generated by artificial intelligence (AI) in the context of complicated social topics.

During the survey you were offered texts the authorship of which was intentionally not disclosed. This was necessary to obtain unbiased and objective assessments regarding only the content and quality of reasoning.

It is important to note that all the texts you have read were either written by real people at the request of the researchers or generated by the GPT 5.3 Instant large language model specifically for this experiment. Any information provided in the texts does not reflect the personal views or opinions of the researchers or authors.

Why is it important? Today AI technologies can create highly compelling and realistic content. In the circumstances of informational warfare, crucial questions arise:

- Can we, on a subconscious level, distinguish between human work and an algorithm?

- What patterns in AI-generated text make us believe or doubt it, especially when the topic or information is controversial?

- Does the anonymity of a source of information leave us more vulnerable to manipulation?

Your participation helps us better understand the answers to these questions and develop the recommendations for protecting the society from automated disinformation campaigns.

Privacy and the right to opt out.

Note that your data remains completely anonymous. If you wish to withdraw your responses after we disclosed the objectives of the study, contact the researcher at: ituzko@kse.org.ua.

Invitation to take part in an interview.

If you would like to spare another 15-20 minutes for this study at a time that suits you, please fill in your contact details via the Google Form at this link to book a brief one-to-one interview**.**

Appendix D. Informed Consent

D.1. Original (Ukrainian)

Вітаємо! Ви запрошені до участі в дослідженні про те, як люди оцінюють аргументи та тексти в онлайн-середовищі. Нижче Ви знайдете інформацію про це дослідження, умови участі в ньому та про те, як будуть оброблятися Ваші персональні дані. Будь ласка, уважно прочитайте наведену інформацію, перш ніж дати згоду на участь.

Про дослідження: Це опитування проводиться в рамках магістерської роботи студентки Київської школи економіки і має на меті дослідити як люди оцінюють аргументи та тексти в цифровому середовищі. До участі запрошуються громадяни України віком 18–35 років. Участь є добровільною та конфіденційною і займає 15-20 хв заповнення цієї онлайн-форми.

Ви можете припинити участь у будь-який момент і це не матиме для Вас жодних наслідків. Конфіденційність участі означає, що ми збираємо дані, які визначені законодавством України як персональні і тому потребують окремої згоди на їх збір. Усі відповіді, які Ви надасте, будуть використані лише в узагальненій формі з науковою метою. Наприкінці опитування Ви матимете можливість перейти в окрему форму та залишити свою контактну інформацію для участі в короткому інтерв'ю (15-20 хвилин). Ця форма не пов'язана з відповідями в основному опитуванні та не дозволятиме ідентифікувати вас у масиві дослідницьких даних.

Згода на участь у дослідженні

☐ Я даю поінформовану згоду на участь у цьому дослідженні у формі опитування та на оприлюднення його узагальнених результатів на умовах конфіденційності.

Згода на збір і обробку персональних даних

☐ Я усвідомлюю, що після завершення опитування я можу за бажанням перейти до окремої форми та залишити електронну адресу для участі в інтерв'ю. Надана мною інформація такого характеру означає, що я даю свою згоду на збір та обробку цих персональних даних виключно з метою організації інтерв'ю.

D.2. Translation (English)

Welcome! You are invited to participate in a study examining how people evaluate arguments and texts in an online environment. Below you will find information about

this study, participation conditions, and personal data management. Please read this information carefully before agreeing to participate.

About the study: This survey is being conducted as part of a master's thesis project of a student at the Kyiv School of Economics and its purpose is to investigate how people evaluate arguments and texts in a digital environment. Ukrainian citizens aged 18–35 are invited to take part. Participation is voluntary and confidential, and it takes 15 to 20 minutes to complete this online form.

You may withdraw at any time, and this will in no way affect you. The confidentiality of your participation means that we are collecting data which is defined by Ukrainian law as personal data, and therefore requires a separate consent for its collection. All responses you provide will be used for scientific purposes in aggregated form only. At the end of the survey, you will have the opportunity to proceed to a separate form and provide your contact details to take part in a short interview (15–20 minutes). The second form is not connected to the responses you provide in the main survey and will not provide any means for identifying you within the research dataset.

Participation consent

☐ I hereby express my informed consent to participate in this research in the form of a survey and to the publication of its aggregated results, under conditions of confidentiality.

Consent for the collection and processing of personal data

☐ I understand that after completing this survey, I may, on a voluntary basis, proceed to a separate form and provide my email address to attend an interview. By sharing such information, I express my permission for the collection and processing of the aforementioned personal data solely for the purpose of organising the interview.

Appendix E. Operationalisation

Variable	Role	Operational indicator(s)	Measurement scale
True source	Independent (X)	Whether the text was human-written or generated by GPT-5.3 Instant; randomly assigned within-participant, unlabelled	Nominal (binary)
Perceived authorship	Independent, predictor (H2-H5) / outcome (H6)	Participant's guess of the author of each text: human, AI, or unsure	Nominal (3 levels)
Persuasiveness index	Dependent (Y)	Mean of three 1–5 items: perceived persuasiveness, argument strength, inclination to change opinion	Continuous (interval; mean of 3 Likert items)
Credibility index	Dependent (Y)	Mean of two 1–5 items: author could be a field expert (expertise); author seems honest and sincere (trustworthiness)	Continuous (interval; mean of 2 Likert items)
Pre-attitude (topic stance)	Covariate / moderator	Agreement with the topic thesis, rated 1–5 before reading each text	Ordinal (single Likert item, treated as continuous)
Positive AI attitudes	Moderator	Positive subscale of the GAAIS (Schepman & Rodway, 2020), 1–5 items, mean score	Continuous (interval)
Negative AI attitudes	Moderator	Negative subscale of the GAAIS, 1–5 items, mean score	Continuous (interval)
Emotional valence	Predictor (H6)	Single 1–5 item: emotions evoked were mostly pleasant or positive	Ordinal (single Likert item, treated as continuous)
Arousal	Predictor (H6)	Single 1–5 item: strong emotional arousal (tension or excitement) while reading	Ordinal (single Likert item, treated as continuous)
Perceived controversy	Predictor (H6)	Single 1–5 item: the topic seemed controversial	Ordinal (single Likert item, treated as continuous)
Internal resistance	Predictor (H6)	Single 1–5 item: the text triggered internal resistance	Ordinal (single Likert item, treated as continuous)

Appendix F. Item-rest correlations for scales used in the study

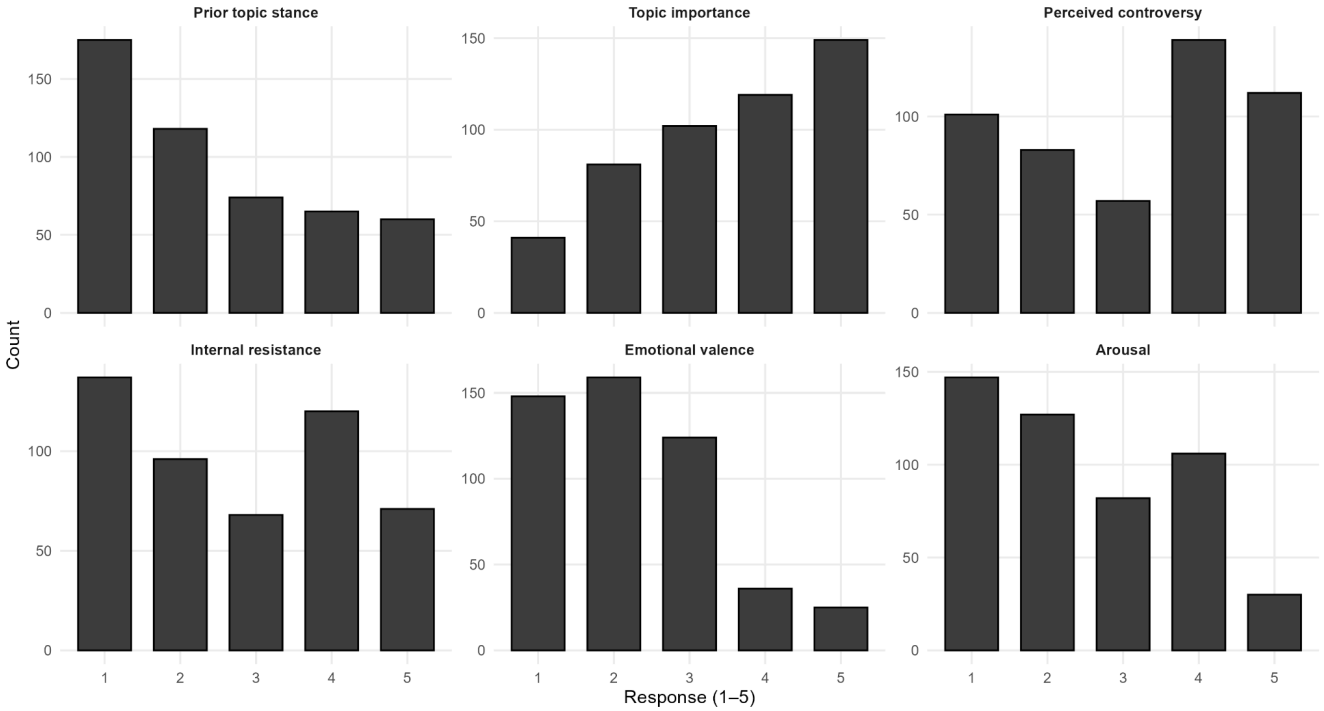
Scale	Item	r.drop
GAAIS Positive	ai_att_1_pref_routine	0.485
GAAIS Positive	ai_att_2_econ_opp	0.425
GAAIS Positive	ai_att_4_happiness	0.367
GAAIS Positive	ai_att_5_impression	0.319
GAAIS Positive	ai_att_7_interest_daily	0.728
GAAIS Positive	ai_att_11_wellbeing	0.531
GAAIS Positive	ai_att_12_exciting	0.531
GAAIS Positive	ai_att_14_better_worker	0.465
GAAIS Positive	ai_att_15_useful_apps	0.616
GAAIS Positive	ai_att_17_superiority	0.421
GAAIS Positive	ai_att_18_society_benefit	0.423
GAAIS Positive	ai_att_19_work_intent	0.608
GAAIS Negative	ai_att_3_unethical_rev	0.355
GAAIS Negative	ai_att_6_errors_rev	0.361
GAAIS Negative	ai_att_8_threatening_rev	0.685
GAAIS Negative	ai_att_9_control_rev	0.559
GAAIS Negative	ai_att_10_dangerous_rev	0.783
GAAIS Negative	ai_att_16_future_anxiety_rev	0.612
GAAIS Negative	ai_att_20_personal_harm_rev	0.544
GAAIS Negative	ai_att_21_spying_rev	0.367
Persuasiveness Index	changemind	0.425
Persuasiveness Index	persuasive	0.772
Persuasiveness Index	argstrength	0.739
Credibility Index	expertise	0.56
Credibility Index	trustworthiness	0.56

Appendix G. Variables used in statistical analysis

Variable	Explanation and role
participant_id	A unique identifier for each respondent, random-effect grouping factor
text_id	An identifier for the four stimulus texts, random intercept in the H1–H5 models
source_coded	The true source of each text (human or AI), the main independent variable, effect-coded as AI=+0.5 and Human=−0.5
authorguess_en	Participant's guess about who wrote each text, recoded from the raw numeric responses into a three-level factor (Human, AI, "Difficult to say" coded as "IDK")
guess_is_ai	A binary recoding of the authorship guess (1 = guessed AI, 0 = guessed Human) used as the outcome in the exploratory H6 model
persuasiveness_index	A composite measure of how persuasive a participant found a text, mean of three five-point items (perceived persuasiveness, argument strength, and inclination to change opinion)
credibility_index	A composite measure of perceived source credibility, the mean of two five-point items corresponding to expertise and trustworthiness
attitudepre_c	The participant's baseline agreement with each topic, measured before reading, used as a covariate, mean-centred (the sample average subtracted from each value)
gaais_pos_c	The participant's positive attitudes toward AI, mean of the positive GAAIS subscale, used as a moderator (H5b), mean-centred
gaais_neg_c	The participant's negative attitudes toward AI, mean of the negative GAAIS subscale, used as a moderator (H5a), mean-centred
resistance	A single five-point item measuring the internal resistance (disagreement) the participant felt while reading, used as a predictor in exploratory model (H6)
controversial	A single five-point item measuring how controversial the participant found the topic, used as a predictor in exploratory model (H6)
valence	A single five-point item measuring the positivity of the emotions a text evoked, used as a predictor in exploratory model (H6)
arousal	A single five-point item measuring the intensity of emotions a text evoked, used as a predictor in exploratory model (H6)

Appendix H. Descriptive statistics for the single-item Likert variables

Distribution of single-item Likert variables



Descriptive statistics for the single-item Likert variables

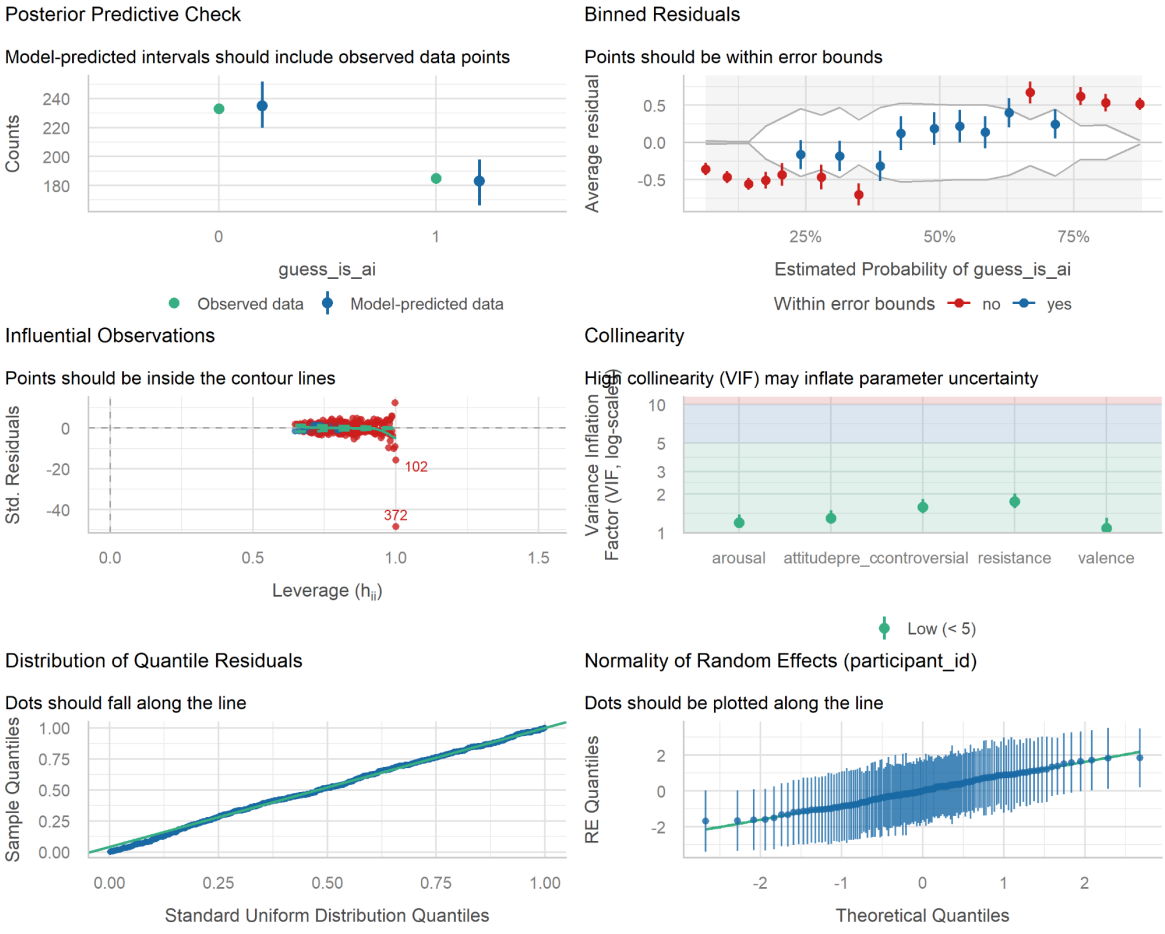
Variable	*n*	*M*	*SD*	Skewness	Excess kurtosis
Prior topic stance	492	2.42	1.40	0.58	−0.99
Topic importance	492	3.52	1.30	−0.42	−0.99
Perceived controversy	492	3.16	1.47	−0.23	−1.38
Internal resistance	492	2.78	1.44	0.12	−1.41
Emotional valence	492	2.25	1.11	0.70	−0.13
Arousal	492	2.48	1.28	0.36	−1.13

Appendix I. Model diagnostic checks

H1-5 model diagnostic checks

Model	R2_marginal	R2_conditional	Normality_p	Homoscedasticity_p	Max_VIF	Singular_fit
H1a	0.105	0.422	0	0.955	1	FALSE
H1b	0.074	0.543	0.245	0.828	1	FALSE
H2a	0.117	0.431	0	0.995	1.02	FALSE
H2b	0.086	0.55	0.213	0.845	1.04	FALSE
H3	0.124	0.43	0	0.994	2.38	FALSE
H4	0.127	0.432	0	0.975	1.01	FALSE
H5a	0.142	0.432	0	0.991	1.01	FALSE
H5b	0.13	0.428	0	0.998	1.01	FALSE

H6 (exploratory) model diagnostic checks



Appendix J. Cross-Model Content Analysis Prompts

Appendix J.1. Prompt for Gemini 3.1 Pro Extended

You are an expert qualitative researcher and social psychologist. The research goal is to identify the attentional and cognitive mechanisms that leave young Ukrainian adults vulnerable to AI-generated, wartime-relevant text. I am attaching cues_open_decoded.csv: open-ended survey responses, natively in Ukrainian. Each participant evaluated four texts, guessed whether each was written by a human or an AI (or chose "can't tell"), rated their confidence, and wrote a free-text justification — the cue. Your task in this stage is discovery and prevalence: surface the themes in those cues and quantify how often each appears. Do not produce a polished multi-sheet or formula-driven workbook here; that is a separate, later step.

A. Prepare the data (reshape to one response per row)

The file is wide — one participant per row, four texts (t1–t4), each with *_source, *_authorguess, *_confidence, *_guesscorrect, *_cues. Match fields by header name (column order is irregular: t1_cues sits after t2_source). Unpivot to a long table, one row per (participant × text); dropping empty slots yields 492 responses. Translate the Ukrainian category values to these exact English labels (keep them consistent — a later stage relies on them):

Guess / source: Штучний інтелект (велика мовна модель) → AI; Людина → Human; Важко сказати → Difficult to say (guess only — the true source is always AI or Human).

Confidence: Повністю впевнений/на → Completely confident; Радше впевнений/на → Somewhat confident; Важко сказати → Difficult to say; Радше не впевнений/на → Somewhat not confident; Зовсім не впевнений/на → Completely not confident.

Preserve the original Ukrainian cue text; add a short English gist of each.

B. Open coding — read natively in Ukrainian

Process the cues in the original language. Look for deep attentional and cognitive mechanisms, not surface keyword matches. Pay particular attention to how participants weigh situational awareness, logical friction, grammatical and lexical imperfection, and the presence or absence of doubt. Generate 5 to 8 emergent themes in sentence case, then add two mandatory categories: one for intuition / gut feeling (a decision resting on overall impression with no nameable textual feature — gut feeling, common sense,

guessing, wishful hope) and one for non-response or ambiguous (explicit "I don't know", blanks, dots or dashes, off-task or otherwise non-substantive answers).

Examples that set the expected analytical depth:

Input: "Думаю цей текст, як і попередні, не враховує українського контексту й того, що Україна вимушена боротися проти значно переважаючого противника…" → The participant judges existential reality, not grammar: they expect a human to prioritise survival over abstract concerns. → Contextual detachment and survival pragmatism.

Input: "Замість вислову 'високий збір' найчастіше використовують податок... людина мабуть би використала слово 'збитковий'." → They reverse-engineer translation quality, seeking localised, lived vocabulary over sterile phrasing. → Lexical authenticity and localisation.

Input: "Трошки не вистачило людяності і сумнівів, які притаманні людині" → They look for the cognitive friction of a human wrestling with a topic rather than a machine's detached confidence. → Emotional resonance and doubt.

Input: "Інтуїція" / "по відчуттях" → No textual feature is named; the verdict rests on impression alone. → Intuition and gut feeling.

Input: "Не можу визначити" / "-" → No usable cue is offered. → Non-response or ambiguous.

Confirmation gate: present the finalised theme list with a one-line definition and the Ukrainian markers that trigger each theme, and stop for my confirmation before coding any responses.

C. Code every response (after I confirm the themes)

Assign up to three themes to each of the 492 responses; most take one. Do not skip any row. Code what the participant reasoned from, not the topic of the text. Apply these disambiguation rules, in order:

"по відчуттях / відчуваю" as the sole basis → intuition; but "відчувається емоція / сухо / без душі / сумніви / щирість / живість" → emotional resonance.

"занадто категорично / немає сумнівів" (over-confident, no doubt) → emotional resonance; "нав'язує думку / категорична позиція" (a pushed stance) → intent/agenda.

"суперечливо / нелогічно / поверхнево / без конкретики / зміст / некоректне трактування" → logical friction.

"стиль / манера / тон / форма / читабельність" → structural predictability; "людяна мова / мовні конструкції / кліше / конкретне слово" → lexical authenticity.

"контроверсійно / провокаційно / клікбейт / замовний / маніпуляція / популізм / байт" → intent/agenda.

A blank, "-", or "." with no words → non-response.

"те саме / аналогічно / див. відповідь 1 / за тими самими" → inherit the themes from that participant's previous text row.

Write a one-line English brief reason for each response.

D. Preliminary theme-appearance analysis (the deliverable)

Define appearance as each time a theme is used across the three assignment slots; a single response can contribute up to three appearances, so totals will exceed 492. Report, as plain ranked tables (compute the counts directly — no spreadsheet formulas needed at this stage):

Overall — appearance count and share per theme, ranked high to low.

By guessed author — appearances per theme within AI / Difficult to say / Human guesses.

By accuracy — appearances per theme within five buckets: accurate AI (guessed AI & true AI), misleading AI (guessed AI & true Human), accurate human (guessed Human & true Human), misleading human (guessed Human & true AI), and Difficult to say.

By confidence level — appearances per theme within each of the five confidence ratings.

For each group in 2–4, name its top themes and characterise the cue pattern in one or two sentences (e.g. which themes dominate, where intuition and non-response concentrate, where confident reasoning differs from hesitant reasoning). Close with the two or three headline patterns you consider most important for the research question.

E. Hand-off and guard

Produce two outputs that the next stage will consume: (i) the confirmed codebook — theme name, one-line definition, Ukrainian markers; and (ii) the coded long-format dataset as a file (CSV or a single sheet) with columns: Participant n, Text, Original response, Translated gist, Author guess, Confidence, Actual source, Guess correct, Assigned theme 1, Assigned theme 2, Assigned theme 3, Brief reason.

Appendix J.2. Prompt for Claude Opus 4.8 Max

You are a data analyst producing a reproducible spreadsheet deliverable. I am attaching a settled content-analysis coding: a long-format dataset of 492 survey responses in which each response already carries its English category values (Author guess $\in$ {AI, Human, Difficult to say}; Confidence $\in$ {Completely confident, Somewhat confident, Difficult to say, Somewhat not confident, Completely not

confident}; Actual source ∈ {AI, Human}) and up to three assigned theme names, together with a codebook listing each theme's name, description, and Ukrainian markers. Do not re-derive or re-judge the themes; treat the assigned codes and theme names as fixed.

Your task is to render this coding as one .xlsx workbook in which every reported count is a live Excel formula — never a typed-in number. Use code execution and the xlsx skill. After saving, run the skill's recalc.py, fix anything that errors, and only then present the file. The workbook must recalculate with zero formula errors.

Workbook — five sheets, headers in sentence case

Sheet 1 — "Codebook and stats". One row per theme. Columns: Theme name | Description | Common sub-codes (Ukrainian markers) | Translated sub-codes | Frequency. Populate the first four columns from the supplied codebook; put a live formula in Frequency (see below).

Sheet 2 — "Dataset and cues". All 492 rows, no rows skipped. Columns: Participant n | Text | Original response | Translation | Author guess | Confidence | Actual source | Guess correct | Assigned theme 1 | Assigned theme 2 | Assigned theme 3 | Brief reason. The theme strings in the three Assigned-theme columns must match the Sheet 1 theme names character-for-character so the COUNTIFs resolve. Make "Guess correct" a formula: =EXACT(E2,G2).

Sheet 3 — "Cues by guessed author". One block per guess type (AI, Difficult to say, Human), each block listing the themes relied upon. Columns: Guess type | Instances | Dominant characteristics | Primary themes relied upon | Frequency. Instances and Frequency are formulas; the prose columns describe each guess type's leading themes.

Sheet 4 — "Cues by accuracy". One block per bucket. Columns: Accuracy category | Response instances | Example gist | Finding or phenomenon | Why it works or why it fails | Primary themes relied upon | Frequency. Define the buckets by guess × actual source: Accurate cues for AI = guessed AI & true AI; Accurate cues for human = guessed Human & true Human; Misleading cues for AI = guessed AI & true Human; Misleading cues for human = guessed Human & true AI. (Difficult-to-say guesses fall outside these four.)

Sheet 5 — "Cues by confidence level". One block per confidence rating. Columns: Confidence level | Response instances | Dominant characteristics | Primary themes relied upon | Frequency.

What "Frequency" means

Everywhere in Sheets 1 and 3–5, Frequency = the number of theme appearances counted across the three assignment columns (I, J, K), not the number of responses. One

response can contribute up to three appearances. Count accordingly. "Instances" / "Response instances", by contrast, count responses in that category.

Formula patterns (reference the 'Dataset and cues' sheet; anchor so they fill down)

Sheet 1 Frequency (theme name in $A2):

=COUNTIF('Dataset and cues'!$I:$I,$A2)+COUNTIF('Dataset and cues'!$J:$J,$A2)+COUNTIF('Dataset and cues'!$K:$K,$A2)

Sheet 3 Instances (guess label in $A2): =COUNTIF('Dataset and cues'!$E:$E,$A2)

Sheet 3 Frequency (guess in A<block>, theme in $D<row>):

=COUNTIFS('Dataset and cues'!$E:$E,A2,'Dataset and cues'!$I:$I,$D2)+COUNTIFS('Dataset and cues'!$E:$E,A2,'Dataset and cues'!$J:$J,$D2)+COUNTIFS('Dataset and cues'!$E:$E,A2,'Dataset and cues'!$K:$K,$D2)

Sheet 5 Instances / Frequency: same as Sheet 3 but filter on column F (Confidence) instead of E.

Sheet 4 Response instances (e.g. Accurate AI): =COUNTIFS('Dataset and cues'!$E:$E,"AI",'Dataset and cues'!$G:$G,"AI") — swap the two literals per bucket.

Sheet 4 Frequency (Accurate AI example; theme in $F<row>):

=COUNTIFS('Dataset and cues'!$E:$E,"AI",'Dataset and cues'!$G:$G,"AI",'Dataset and cues'!$I:$I,$F2)+COUNTIFS(…$J:$J,$F2)+COUNTIFS(…$K:$K,$F2)

Give every formula cell a numeric format; never paste the evaluated value.

Reconciliation checks before presenting: Total theme appearances across columns I/J/K must equal the sum of Sheet 1's Frequency column.

The Frequency blocks on Sheet 3 (across all guess types) and on Sheet 5 (across all confidence levels) must each sum to that same total.

The four Frequency blocks on Sheet 4 must sum to the total of appearances among AI- and Human-guess responses only (Difficult-to-say guesses excluded).

Run recalc.py, confirm zero errors, then present the workbook with a brief note of the sheet contents.

Reliability statistics.

Add a calculations sheet deriving: overall hit rate among committed guesses=(COUNTIFS(E,"AI",G,"AI")+COUNTIFS(E,"Human",G,"Human"))/(COUNTIF(E,"AI")+COUNTIF(E,"Human")); and per-cue precision for each guess direction = (that cue's Sheet-4 accurate frequency) ÷ (that cue's Sheet-3 frequency). Keep these as formulas too.

Appendix K. Thematic codebook

Theme name	Description	Common sub-codes (Ukrainian)	Translated subcodes	Frequency
Non-response or ambiguous	Explicit cannot-determine, off-task remarks, blank or punctuation-only answers, or 'same as previous' with no substantive cue.	важко сказати, не знаю, не можу визначити, немає відповіді, те саме	hard to say, don't know, cannot determine, no response, same as previous	145
Structural predictability and mechanical perfection	Focusing on sentence structure, formatting cliches, style and tone, overly balanced argumentation, or unnatural punctuation.	структура, речення, пунктуація, формальний, абзац	structure, sentence, punctuation, formal, paragraph	104
Logical friction and depth of expertise	Evaluating argumentation depth. AI associated with pseudo-expertise, human with specific evidence.	аргумент, логіка, факти, статистика, конкретика	argument, logic, facts, statistics, specifics	88
Intuition and gut feeling	Decision rests on intuition or overall impression rather than a nameable textual feature; gut feeling, common sense, guessing, or wishful hope.	інтуїція, чуйка, по відчуттях, здоровий глузд, склалося враження, вгадую	intuition, gut feeling, by feel, common sense, got an impression, guessing	62
Contextual detachment and survival pragmatism	Evaluating the text against existential realities of the war. Human prioritises physical survival.	війна, контекст, виживання, оборона, реалії	war, context, survival, defence, realities	58
Emotional resonance and doubt	Looking for human psychological footprint: empathy, friction, emotion, over-detached confidence, or absence of doubt.	емоції, людяність, сумнів, відчуття, душа	emotions, humanity, doubt, feeling, soul	49
Lexical authenticity and localisation	Scanning for organic vocabulary. AI uses sterile translations; human uses lived terminology.	слово, переклад, русизм, термін, лексика	word, translation, russism, term, vocabulary	41
Intent and agenda attribution	Searching for hidden motive, manipulative intent, populism, or clickbait.	популізм, замовна, клікбейт, мета, байт	populism, commissioned, clickbait, goal, bait	25
Total N				572