Cybersecurity Awareness of OWASP LLM Risks: A Pre-Study of Digital Media Professionals in Germany

Alexander Lawall [1] and Stephan Böhm [2]
[1]IU International University of Applied Sciences, Erfurt, Thüringen, Germany
[2]RheinMain University of Applied Sciences and Arts, Department of Media Management, Wiesbaden, Germany
e-mail: alexander.lawall@iu.org, stephan.boehm@hs-rm.de

***Abstract*—The growing integration of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) into media workflows introduces new cybersecurity risks that are insufficiently understood by media professionals. Despite the increasing use of AI-supported tools for research, content creation, and automation, empirical evidence regarding awareness of LLM-specific security threats remains limited. This study investigates the cybersecurity awareness and action competence of the surveyed media professionals in Germany regarding five selected OWASP Top 10 for LLM Applications risks. The research addresses two questions: (RQ1) What is the current level of awareness and action competence concerning five selected OWASP Top 10 for LLM Applications risk categories among media professionals? and (RQ2) How do professionals differ in their responses across different LLM-related risk scenarios? The study applies a quantitative survey design grounded in Protection Motivation Theory (PMT) and incorporates vignette-based scenarios aligned with five OWASP LLM risk categories. Preliminary findings from 72 participants indicate a discrepancy between high AI adoption and only moderate confidence in understanding AI-related risks. However, most respondents selected security-conscious responses in the vignette scenarios, particularly regarding sensitive information disclosure and supply chain risks. The preliminary results suggest a need for targeted LLM security training and clearer organizational guidelines for secure AI usage in media organizations, particularly to strengthen self-efficacy in technically complex scenarios.**

Keywords-Generative Artificial Intelligence (GenAI); Large Language Models (LLMs); Cybersecurity Awareness; OWASP LLM Top 10; Protection Motivation Theory (PMT).

I. Introduction

Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) are increasingly transforming workflows in the media industry. Tools, such as ChatGPT, Microsoft Copilot, and Gemini are now widely used for research, text production, translation, summarization, and editorial support. While these systems offer significant efficiency gains, they also introduce new cybersecurity and information integrity risks that extend beyond traditional IT security concerns [1].

Unlike conventional software systems, LLMs interact directly with users through natural language and are increasingly embedded in routine decision-making processes. As a result, risks, such as prompt injection, sensitive information disclosure, insecure output handling, or unsafe third-party integrations may remain unnoticed during everyday work, particularly in journalism and media production environments handling confidential information.

Recent frameworks, such as the Open Worldwide Application Security Project (OWASP) Top 10 for LLM Applications provide structured classifications of the most important LLM-related security risks [2]. However, despite growing technical discussions about LLM security, empirical research on how users perceive and respond to these risks remain limited. Existing studies mainly focus on general AI literacy or broad cybersecurity awareness, while systematic investigations of LLM-specific action competence are largely absent.

To address this research gap, the present study examines the cybersecurity awareness and action competence of media professionals in Germany regarding selected OWASP LLM risks. The study combines Protection Motivation Theory (PMT) with vignette-based scenarios to analyse both perceived risks and practical decision-making behaviour in realistic AI-related situations.

The study addresses the following research questions:

> **RQ1:** What is the current level of awareness and action competence regarding five selected OWASP Top 10 for LLM Applications risk categories among media professionals?
> **RQ2:** How do media professionals differ in their action competence across the five selected OWASP Top 10 for LLM Application risk scenarios presented through vignettes?

To answer these questions, an online survey with integrated vignette scenarios was conducted among media professionals in Germany. The scenarios cover five OWASP-aligned risk categories: Prompt Injection, Sensitive Information Disclosure, Supply Chain Risks, Data & Model Poisoning, and Improper Output Handling. The study contributes to emerging research on human-centred GenAI security and provides practical implications for secure AI adoption in media organizations.

The remainder of this paper is structured as follows. Section II presents the theoretical background and related work, including GenAI and LLM security risks, the OWASP Top 10 for LLM Applications framework, cybersecurity awareness and action competence, and the application of PMT in cybersecurity research. Section III describes the research design and methodology, including the questionnaire design and vignette-based approach. Section IV presents the empirical results of the pre-study. Section V discusses the findings, implications, and limitations of the study. Finally, Section VI concludes the paper and outlines directions for future research.

II. Theoretical Background and Related Work

A. Generative AI and LLM Security Risks

GenAI and LLMs are increasingly integrated into media organizations, supporting content creation, research, translation, personalization, and automation of editorial workflows. Empirical studies consistently report high general awareness of AI technologies among media professionals, with up to 85% of journalists indicating medium to high familiarity with AI-supported tools [1][3].

However, this general awareness does not extend to security-specific implications of LLM usage. Multiple studies demonstrate a persistent gap between recognizing AI adoption and understanding concrete security risks, such as data leakage, manipulation of model outputs, or misuse of generative systems [4][5]. For example, while 85% of surveyed participants acknowledged the use of machine learning systems in digital environments, only 37% demonstrated awareness of associated data security implications [4]. This discrepancy is particularly concerning in media contexts, where professionals routinely process sensitive information, unpublished materials, and confidential sources.

LLM-related security risks further differ from traditional cybersecurity threats in that they are often embedded within routine work practices rather than discrete technical failures. Risks, such as prompt manipulation, insecure reuse of generated output, or unsafe integration of third-party components may remain unnoticed without targeted security knowledge [6]. These characteristics highlight the importance of assessing not only declarative knowledge, but also practical action competence when individuals are confronted with realistic LLM-related risk situations.

B. Five Selected OWASP Top 10 for LLM Risk Categories

The present study focuses on five selected risk categories from the OWASP Top 10 for LLM Applications framework. These provide a structured taxonomy of the most critical security risks associated with Large Language Models, including Prompt Injection, Insecure Output Handling, Training Data Poisoning, Model Denial of Service, and Supply Chain Vulnerabilities [2]. Although the framework has rapidly gained relevance in practitioner communities, empirical research examining human awareness and competence regarding these risk categories remains extremely limited.

A systematic synthesis of recent literature reveals that no identified study has comprehensively assessed awareness or action competence across all five OWASP LLM risk categories among media professionals. Existing research addresses selected risks only indirectly, most prominently insecure output handling in the context of misinformation, deepfakes, and LLM-generated phishing attacks [7][8]. In contrast, risks, such as prompt injection, training data poisoning, and supply chain vulnerabilities are largely absent from empirical investigations involving media-related populations.

This fragmented coverage severely limits comparative analysis across different LLM risk types. Moreover, the recency of the OWASP LLM framework partially explains the lack of direct empirical evidence, as many studies predate its publication or focus on broader AI literacy rather than security-specific vulnerabilities. Consequently, a systematic, OWASP-aligned assessment of awareness and action competence constitutes a clear research gap.

C. Cybersecurity Awareness and Action Competence

Cybersecurity awareness is commonly conceptualized as the extent to which individuals recognize, understand, and recall security threats and appropriate protective measures. However, a substantial body of research demonstrates that awareness alone is insufficient to predict secure behaviour. Instead, action competence, the practical ability to detect, assess, and appropriately respond to security-relevant situations, has emerged as a critical construct.

Among media professionals, awareness is predominantly measured using self-reported surveys or interviews, often without standardized or validated instruments [1][9]. Action competence, where examined, is frequently inferred indirectly through participation in training programs or perceived skills rather than through observation of behavior. Only a small number of studies employ scenario-based or task-oriented approaches, and none systematically map such scenarios to LLM-specific security risks [7][10].

Consistent role-based differences further complicate the picture. Technical staff and managerial roles generally exhibit higher levels of training and competence than journalists and editors, despite the latter being most directly exposed to LLM tools in everyday work [3][11]. This misalignment between exposure and competence suggests that front-line media professionals may underestimate LLM-specific risks or lack the skills necessary to translate awareness into secure action.

D. Protection Motivation Theory in Security Research

PMT offers a well-established framework for explaining how individuals evaluate threats and decide whether to engage in protective behaviours [12][13]. PMT distinguishes between threat appraisal (Perceived Severity and Vulnerability) and coping appraisal (Response Efficacy, Self-Efficacy, and Response Costs), which jointly influence protection motivation and behavioural intention [13].

In cybersecurity research, PMT has been successfully applied to explain a wide range of security-relevant behaviours, including phishing susceptibility, policy compliance, and adoption of protective practices [14]–[18]. The theory is particularly suitable for contexts characterized by time pressure and uncertainty, conditions that are highly characteristic of media production environments.

Despite its extensive use in general cybersecurity contexts, PMT has rarely been applied to risks associated with Generative AI and LLMs, and to date not in studies focusing on media professionals. This represents a significant gap, as PMT provides a theoretically grounded mechanism for integrating awareness, motivation, and action competence within a single

explanatory model. Combining PMT with vignette-based assessments aligned to the five selected OWASP LLM Top 10 risks enables a systematic investigation of current awareness levels (RQ1) and differences in action competence across risk scenarios (RQ2).

III. Research Design and Methodology

A. Questionnaire Design

The questionnaire was designed to examine the use of generative AI systems in the media sector, as well as perceptions of and approaches to security-related risks associated with GenAI and LLM applications. As already presented, the theoretical framework is based on the PMT and the OWASP Top 10 for LLM Applications. The final questionnaire was implemented as an online panel survey in Unipark. The questionnaire was prepared in German. All analyses in the results section of this paper have been translated into English.

The first section of the questionnaire collected demographic and professional characteristics, including age, gender, educational background, media sector, and professional role. Another section addresses the use of AI tools in everyday work. Participants rate both the frequency of AI use and their familiarity with various applications. In addition, specific areas of application, such as research, text creation, translation, social media production, multimedia creation, and technical support were surveyed. This is intended to provide a nuanced picture of the integration of GenAI into various media work processes.

The questionnaire further assessed perceptions of AI-related risks, including misinformation, data breaches, bias, organizational governance, and secure AI usage. The development of the PMT items was guided by the theoretical foundations of Protection Motivation Theory [12][13] and PMT applications in information security research [16]–[18]. As no established and validated scale for measuring GenAI-related security risks in a media context was available at the time of the study, the items were developed on a theory-driven basis. The wording was based on the conceptual definitions of the PMT constructs, Perceived Severity, Perceived Susceptibility, Self-Efficacy, Response Efficacy and Response Costs, as well as on typical operationalisations of these constructs in information security research. The items used in the questionnaire therefore do not constitute direct translations or adaptations of existing scales, but rather context-specific operationalisations for the use of generative AI in media organisations.

The wording was developed by the authors and reviewed for consistency with the theoretical constructs, as well as for comprehensibility within the media context. As the questionnaire was originally developed and administered in German, no translation or back-translation procedures were required. To limit the duration of the survey and minimise the burden on participants, each PMT construct was measured using two indicators that reflect the respective core dimensions of the construct. Each items was measured on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. Scale scores were calculated as arithmetic means of the corresponding items. Because this is a pre-study with a limited sample size, the present analysis does not yet report full-scale validation; reliability and construct validity will be examined in the subsequent full-sample study.

B. Vignette Design for Decision-Making Competence

The core of the study is a vignette-based scenario module that simulates typical work situations from everyday media practice involving generative AI. The aim of this approach is not only to capture attitudes toward AI, but also to examine concrete decision-making in realistic risk situations. The five vignettes used are based on the first five categories of the OWASP Top 10 for LLM Applications. Each vignette describes a brief, everyday work situation in the media sector and asks participants to choose a specific course of action. The response options represent behaviours of varying degrees of safety or risk. The vignettes cover the following risk areas:

- *LLM01 – Prompt Injection:* AI-assisted source evaluation during online research on a political topic. The AI classifies an unknown source as "verified" without justifying its assessment.
- *LLM02 – Sensitive Information Disclosure:* Use of an AI tool to summarize an interview transcript containing sensitive and "off the record" information.
- *LLM03 – Supply Chain Risks:* Installation of an AI plugin with extensive access rights within a content management system.
- *LLM04 – Data & Model Poisoning:* Research using an AI system that provides contradictory or potentially manipulated information from an internal knowledge base.
- *LLM05 – Improper Output Handling:* Automatically generated HTML and script code for an info box intended to be integrated directly into a Content Management System (CMS).

Each vignette was followed by four alternative courses of action representing different levels of risk handling competence. The options ranged from riskiest, clearly insecure behaviour and insufficient mitigation strategies to security-conscious best-practice responses. In some cases, an additional option reflected an overly restrictive avoidance strategy. This design enabled the assessment of participants' ability to identify and select the most appropriate response to the presented GenAI-related security risk. The classification was reviewed by the authors to ensure consistency with the underlying OWASP risk category and the intended learning objective of the scenario.

After each vignette, participants answer additional questions regarding their confidence in their decision, the perceived level of risk, and the realism of the scenario. This is intended to examine not only the actual decision-making process but also respondents' confidence in their judgment and the plausibility of the described situations in their daily work.

Overall, the vignette design expands the questionnaire to include a behaviour-oriented perspective on dealing with GenAI risks. While many previous studies have primarily focused on attitudes or usage intentions, the approach chosen here

provides initial insight into the decision-making competencies required to address security-relevant AI situations in a media context.

C. Note on the Pre-Study Nature of the Survey Data

The following analyses are based on a preliminary study in which an initial portion of the planned total sample was surveyed ($n = 72$). The aim of this pre-study is to review initial descriptive findings and to validate the survey instrument. The results should therefore be regarded as preliminary and do not yet permit any inferential statistical conclusions. Further analyses, in particular correlation analyses, hypotheses testing and the evaluation of statistical relationships between PMT and the vignette study, will be the subject of subsequent work once the full sample is available.

IV. Results

A. Sample Characteristics

Data collection took place from 13.04.2026 to 25.04.2026 using an online questionnaire distributed via an external panel provider. To reach a target group with relevant interests, the questionnaire was specifically targeted at media-savvy individuals in the panel.

A total of 437 people began the survey. Participants were excluded if they did not meet the predefined screening criteria, terminated the questionnaire before completion, or provided incomplete survey responses. The screening process focused on professional activity in the media sector and on employment primarily in Germany. After this screening process, 72 valid datasets remained and were included in the further analysis.

Seventy-two people from the media industry participated in the survey. The average age of 49, with a range from 21 to 76, indicates an experienced, professionally established sample; career starters are hardly represented. The gender distribution is roughly balanced, with approximately 56% female and 44% male participants.

Regarding media sectors, the sample shows a marked imbalance: about 68% of respondents work in print and publishing, while all other sectors, including online/digital/streaming, radio/television, and film and music production, together account for less than one-third and are represented in some cases by only two or three people. This concentration on a single subsector limits the generalizability of the findings to the media industry as a whole, and should be considered when interpreting the results. At the functional level, operational roles are predominant: almost half of the respondents are permanent employees without personnel responsibility, with an additional 18% working as freelancers.

Management positions, ranging from team leadership to executive management, account for about a quarter of the sample, with executive management members constituting the largest management category at 11%. Trainees and working students, on the other hand, are scarcely represented, which confirms the tendency toward experienced professionals already observed in the age distribution. A complete breakdown of all sociodemographic characteristics is presented in Table I.

TABLE I. Sociodemographic Characteristics of the Sample ($n = 72$)

Characteristic	n	**%**
Age (Mean = 49; Min = 21; Max = 76)		
Gender		
Female	40	55.6
Male	32	44.4
Non-binary	0	0.0
No response	0	0.0
Educational Background		
Vocational training	28	38.9
Bachelor's degree	10	13.9
Master's degree	21	29.2
Doctorate	1	1.4
Other	12	16.7
Media Sector		
Print/Publishing	49	68.1
Online/Digital/Streaming	8	11.1
Radio/Television	3	4.2
Film/Video Production	2	2.8
Music/Audio Production	2	2.8
Social Media	2	2.8
Events	2	2.8
Other	4	5.6
Job Level		
Employee (no mgmt. resp.)	35	48.6
Freelancer	13	18.1
Senior management	8	11.1
Team lead	5	6.9
Department head	5	6.9
Other	5	6.9
Apprentice/Trainee	1	1.4
Student worker/Intern	0	0.0

mgmt. resp. = managerial responsibility

B. Use of AI Tools

The following section analyses the extent to which AI tools are known and used in the media industry, as well as the tasks for which they are employed. The focus is on both widely used general-purpose solutions and LLMs, in line with the focus of the following analysis, as well as specialized image-generation tools. This is to paint a nuanced picture of current usage patterns among media professionals.

The awareness and use of selected AI tools among the media professionals surveyed are presented in Table II. ChatGPT stands out in particular: with an awareness rate of 95.8 %, it is by far the best-known of the tools surveyed, and for 27.8% of respondents, using it at least once a day is already part of their daily work routine. Other general-purpose solutions, such as Google Gemini (88.9%) and Microsoft Copilot (84.7%) also have a high level of awareness, but are used intensively far less frequently: only 5.6% and 4.2% of respondents, respectively, use these tools daily or more often. Image-generating AI tools, such as DALL-E, Midjourney, and Stable Diffusion are known to up to half of respondents, but play hardly any role in their daily workflows. Their usage rate of at least once a day is no more than one. 1.4%.

Table III shows the tasks for which the media professionals surveyed use AI tools in the workplace on an at least daily basis. The most common application is research and infor-

TABLE II. AWARENESS AND USAGE OF AI TOOLS ($n = 72$)

Tool	Known (%)	Never used (%)	Daily or more (%)
ChatGPT	95.8	11.1	27.8
Google Gemini	88.9	40.3	5.6
MS Copilot	84.7	48.6	4.2
Perplexity AI	62.5	36.1	6.9
Claude	54.2	34.7	4.2
DALL·E	50.0	30.6	1.4
Midjourney	48.6	31.9	1.4
Stable Diffusion	40.3	29.2	0.0

mation gathering (20.8%), followed by drafting and editing texts as well as summarising and condensing content (13.9% each). Translation and language adaptation and the creation of social media or marketing content are each cited by 11.1% of respondents. Together, these language-based tasks are the most popular AI use cases at the workplace of the surveyed media professionals. By contrast, creative and technical applications play a considerably smaller role: creating or editing image, audio, or video content, planning and concept development, and organising work processes are each mentioned by only 4.2% of respondents. The least common use case is programming, data analysis, or technical support, cited by just 1.4% which corresponds to just one person reporting such a user in the sample.

TABLE III. TASKS FOR WHICH AI TOOLS ARE USED AT WORK ($n = 72$)

Task	Share (%)
Research and information gathering	20.8
Drafting and editing texts	13.9
Summarising and condensing content	13.9
Translation and language adaptation	11.1
Social media or marketing content	11.1
Image, audio, or video creation/editing	4.2
Planning, concept development, ideation	4.2
Organising and structuring work processes	4.2
Programming, data analysis, tech support	1.4

Furthermore, it is noteworthy that more than half of the respondents have never used AI for several areas of work. This is particularly pronounced in programming, data analysis, and technical support (69.4%), the creation of social media or marketing content (62.5%), as well as the creation or editing of image, audio, or video content and the organization of work processes (61.1% each). In the areas of planning, concept development, and idea generation, 55.6% of respondents also stated that they have not yet used AI. This suggests that the potential for AI use in these areas remains largely untapped.

Overall, the findings indicate that AI tools, particularly ChatGPT, are already firmly embedded in the media industry's daily work routines and are used in significant proportions at least daily, though with a clear focus on language-based applications. However, creative, organizational, and technical applications of AI tools have barely been tapped so far. Given the high level of awareness and, in some cases, already routine use, the question arises as to what extent respondents are aware of the specific risks associated with the use of large language models. The following section therefore focuses on the security aspects of LLM use and examines how media professionals assess and deal with them.

C. Risk Perceptions

The first question, in the risk section of the study, asked the respondents to assess their knowledge of the risks associated with using AI tools. Figure 1 shows the distribution of responses. Overall, just under half of the respondents (43.0%) rate their knowledge as rather good or very good, while another third (33.3%) medium, and about a quarter (23.7%) rate it as rather poor or very poor. This means that the majority of respondents (57.0%) have, at best, an average level of risk awareness. Given that tools like ChatGPT are familiar to nearly all respondents and are used daily by almost a third of them, this result is noteworthy: a high level of familiarity with a tool does not necessarily go hand in hand with a strong awareness of the associated risks. This suggests that while the widespread use of AI tools is well established in the media industry, risk awareness has not kept pace to the same extent.

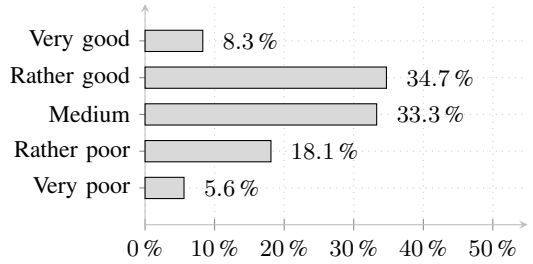

Figure 1. Self-assessed knowledge of AI tool risks ($n = 72$)

The study findings on the respondents' assessments of potential risks associated with the use of AI tools are shown in Table IV. Overall, a clear perception of risk is evident; statements about the immediate consequences of AI errors receive the highest levels of agreement. For example, respondents agree most strongly with the statement that errors in AI systems can have serious negative consequences, followed by concerns regarding the reliability of AI-generated content. The risk of unintended bias is also regarded as relevant. In contrast, agreement on organizational and structural issues is significantly lower. Clear guidelines for dealing with AI risks are perceived only moderately within their organizations, and the feeling of being sufficiently informed about AI risks is also below average. The lowest score is for the statement that time or performance pressure leads to risks being overlooked– which suggests that deliberate disregard of risks is rather rare in everyday work. Overall, a pattern emerges in which risks are generally recognized, but institutional preparedness and personal knowledge are perceived as insufficient.

TABLE IV. PERCEIVED RISKS OF AI TOOL USAGE ($n = 72$, SCALE: 1 = STRONGLY DISAGREE, 5 = STRONGLY AGREE)

Statement	M (SD)
Errors in AI systems can have real negative consequences.	4.06 (1.02)
I have concerns that AI-generated content is not always reliable or accurate.	4.04 (1.08)
Using AI tools can lead to unintended biases or misinterpretations of content.	3.81 (1.07)
In my work environment, potential risks of AI tool usage are taken seriously.	3.72 (1.02)
Using AI tools poses significant risks to the quality of media content.	3.60 (0.99)
My organisation has clear guidelines for managing risks associated with AI tools.	3.19 (1.24)
I feel sufficiently informed about the potential risks and challenges of using AI.	3.08 (1.03)
Time or performance pressure causes me to overlook risks when using AI tools.	2.57 (0.99)

To assess risk perception, the framework of *PMT* was used, which distinguishes between a *threat appraisal* and a *coping appraisal*. Threat appraisal comprises the constructs *Perceived Severity* (SE), i.e., the perceived severity of potential harm caused by AI errors, and *Perceived Susceptibility* (SU), i.e., the subjective assessment of one's own vulnerability to such problems in everyday working life. The coping assessment is operationalised through three constructs: *Self-Efficacy* (SF) captures confidence in one's own ability to recognise and manage AI risks; *Response Efficacy* (RE) measures the perceived effectiveness of protective measures, such as review processes or guidelines; and *Response Costs* (RC) reflects the perceived effort associated with the safe use of AI. Each construct was measured using two items.

Figure 2 shows the mean scores for the five constructs. Overall, a nuanced picture emerges: The threat constructs achieve the highest scores, *Perceived Severity* is the highest at $M = 3.94$, followed by *Perceived Susceptibility* at $M = 3.44$. This suggests that respondents generally perceive AI-related errors as a serious threat. In terms of coping assessment, *Response Efficacy* ($M = 3.85$) achieves a comparably high value, suggesting a positive assessment of the effectiveness of protective measures. In contrast, *Response Costs* ($M = 3.33$) and *Self-Efficacy* ($M = 3.17$) are significantly lower, respondents rate their abilities in dealing with AI risks as rather moderate and perceive the effort required for the safe use of AI as considerable.

As this paper reports results from a preliminary study with a limited sample size, the present analysis focuses on descriptive statistics. Reliability analyses of the PMT scales, including internal consistency indicators and inter-item relationships, can be conducted after completion of the full data collection.

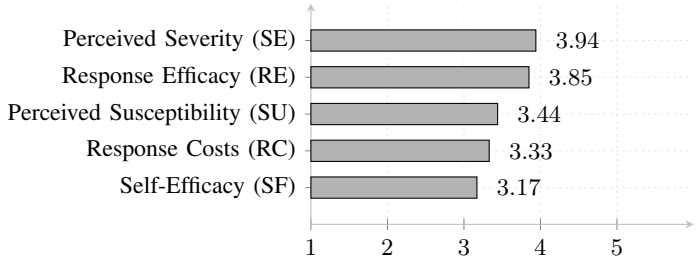

Figure 2. Scale means of PMT constructs ($n = 72$, scale: 1 = strongly disagree, 5 = strongly agree). Values represent the mean across all items per construct.

These analyses will provide additional evidence regarding the measurement quality of the adapted PMT constructs.

A look at the individual items of the threat assessment (Table V) shows that the potential damage to trust caused by AI errors is perceived as particularly significant: Item SE2 achieves the highest score in the entire survey with $M = 4.01$, placing it clearly in the 'agree' range. The perceived consequences for one's own organisation (SE1, $M = 3.88$) are also rated as relevant. The vulnerability scores are somewhat more moderate: whilst SU2 ($M = 3.56$) suggests that AI-related errors are considered realistic in a team environment, the personal impact in day-to-day work (SU1, $M = 3.33$) is somewhat less pronounced.

TABLE V. THREAT APPRAISAL ITEMS ($n = 72$, SCALE: 1 = STRONGLY DISAGREE, 5 = STRONGLY AGREE)

Statement	M (SD)
SE1: AI errors can have severe consequences for my organisation or my work.	3.88 (1.18)
SE2: An AI-related error could significantly damage the trust of users or target audiences.	4.01 (1.00)
SU1: In my daily work, it is realistic that problems caused by AI could arise.	3.33 (1.05)
SU2: In my team or work environment, AI could cause errors or data-related problems.	3.56 (0.96)

In the coping appraisal (Table VI), the items relating to *Response Efficacy* are the most positive: Both careful checks (RE1, $M = 3.88$) and clear guidelines (RE2, $M = 3.83$) are perceived as effective protective measures. The *Self-Efficacy* items, on the other hand, show low scores on the coping scale: respondents do not feel particularly confident in recognising AI risks (SF1, $M = 3.14$), nor in reliably identifying erroneous outputs (SF2, $M = 3.19$). The *Response Costs* paint an ambivalent picture: whilst the general time required for the safe use of AI is perceived as moderate (RC1, $M = 3.56$), agreement with the statement that risks are overlooked under time pressure is somewhat lower (RC2, $M = 3.10$) – suggesting that whilst time pressure is a factor, it is not perceived as a dominant obstacle.

TABLE VI. COPING APPRAISAL ITEMS ($n = 72$, SCALE: 1 = STRONGLY DISAGREE, 5 = STRONGLY AGREE)

Statement	M (SD)
SF1: I know how to identify risks when using AI.	3.14 (1.04)
SF2: I feel capable of identifying problematic or faulty AI outputs.	3.19 (1.06)
RE1: Careful review significantly reduces the risks of using AI.	3.88 (0.95)
RE2: Clear rules or guidelines help to use AI safely in everyday work.	3.83 (0.89)
RC1: Using AI safely takes too much time in everyday work.	3.56 (0.96)
RC2: Under time pressure, it is difficult to consider all risks when using AI.	3.10 (1.10)

D. Vignette Study

The following section presents the results of the vignette study, in which five real-world scenarios relating to safety aspects of LLM deployment were presented. For each vignette, respondents selected one of four courses of action, the order of which was randomised to minimise order effects. The evaluation in this preliminary study is exclusively descriptive; tests for statistical significance and further inferential statistical analyses will be the subject of subsequent studies using the full sample.

1) Vignette LLM01 – Prompt Injection: Table VII shows the distribution of responses for Vignette LLM01 (*Prompt Injection*), in which an AI-assisted source assessment is output as "verified" without any justification. The majority of respondents (68.1%) chose the safest course of action by a clear margin: independently verifying the source and repeating the task. This response aligns with the recommended approach to potential prompt injection risks, as it avoids the uncritical acceptance of a potentially manipulated AI output. Just under a fifth (19.4%) opted to ask the AI for a more detailed answer, a response that does not sufficiently address the underlying risk. 11.1% intended to avoid using AI for research altogether in future, and only 1.4% accepted the AI's assessment without further verification.

TABLE VII. LLM01 – PROMPT INJECTION: RESPONSES TO AI-ASSISTED SOURCE EVALUATION ($n = 72$)

Response option	(%)
Accept the AI assessment without verification.	1.4
Ask the AI to elaborate its answer in more detail.	19.4
Independently verify the source and redo the task.	68.1
Stop using AI for research tasks altogether.	11.1

Respondents rated their decision as fairly certain ($M = 3.86$, $SD = 0.92$) and perceived the associated risk as low ($M = 2.33$, $SD = 0.84$). The situation was rated as somewhat realistic in relation to their own everyday working lives ($M = 3.10$, $SD = 1.08$), which suggests a certain distance from the scenario described.

2) Vignette LLM02 – Sensitive Information Disclosure: Table VIII shows the distribution of responses for Vignette LLM02 (*Sensitive Information Disclosure*), in which an interview transcript containing sensitive and confidential information is to be summarised using an AI tool. The result is particularly clear-cut: at 87.5%, the overwhelming majority of respondents chose the data protection-compliant option, namely the complete removal or anonymisation of sensitive and 'off the record' information before passing it on to the AI tool. This means that support for the safest course of action in this vignette is even more pronounced than in LLM01. The remaining responses are distributed across less secure options: 5.6% stated that they would insert the full transcript and instruct the AI to ignore sensitive data, an approach that offers no effective protection, as the data is still processed. Only 4.2% opted to just remove names, and 2.8% would paste the transcript in its entirety.

TABLE VIII. LLM02 – SENSITIVE INFORMATION DISCLOSURE: RESPONSES TO AI-ASSISTED INTERVIEW TRANSCRIPT SUMMARISATION ($n = 72$)

Response option	(%)
Insert the full transcript into the AI tool.	2.8
Remove names only and leave the rest unchanged.	4.2
Remove or anonymise all sensitive and off-the-record information beforehand.	87.5
Insert the full text and ask the AI to ignore sensitive data.	5.6

Respondents rated their decision as certain ($M = 4.00$, $SD = 0.92$), the highest confidence score of all the vignettes analysed to date. The perceived risk of their decision was rated as low ($M = 2.32$, $SD = 0.98$), which is consistent with the choice of the safest option. Compared to LLM01, the situation was perceived as slightly less realistic in relation to the respondents' own everyday working lives ($M = 2.82$, $SD = 1.13$), which could suggest that dealing with confidential interview transcripts is not part of the everyday working context for all respondents.

3) Vignette LLM03 – Supply Chain Risks: Table IX shows the distribution of responses for Vignette LLM03 (*Supply Chain Risks*), which describes the installation of an AI plugin with extensive access rights in a content management system. Here, too, the picture is clear: 72.2% of respondents chose the safest option and would use only tested and officially approved tools. It is noteworthy that not a single person chose the riskiest option, immediate installation without further testing (0.0%). 16.7% opted to test the plugin in a real project first, which, whilst signalling a degree of caution, does not fundamentally address the risk posed by the extensive access rights. 11.1% would install the plugin but not use it for important content, a strategy that also only mitigates the supply chain risk to a limited extent.

TABLE IX. LLM03 – SUPPLY CHAIN RISKS: RESPONSES TO AI PLUGIN INSTALLATION WITH EXTENSIVE ACCESS RIGHTS ($n = 72$)

Response option	(%)
Install the plugin immediately – it clearly saves time.	0.0
Test the plugin first in a real project.	16.7
Only use reviewed and officially approved tools.	72.2
Install the plugin but avoid using it for important content.	11.1

Respondents rated their decision as certain with the highest confidence value to date ($M = 4.10$, $SD = 1.00$) and assessed the perceived risk as low ($M = 2.18$, $SD = 0.97$), the lowest risk value of all three vignettes evaluated so far. The realism score, at $M = 2.90$ ($SD = 1.08$), falls within the mid-range, similar to LLM02, suggesting that plugin installations with extended access rights are not part of the immediate daily work routine for all respondents.

4) Vignette LLM04 – Data & Model Poisoning: Table X shows the distribution of responses for Vignette LLM04 (*Data & Model Poisoning*), in which an AI system provides contradictory or potentially manipulated information from an internal knowledge database. At 80.6%, the vast majority of respondents chose the most appropriate response: verifying the underlying data and correcting or removing problematic sources. As with LLM03, nobody chose the option of using the content unverified despite time pressure (0.0%), a consistent pattern across several vignettes. 16.7% opted to ask the AI for a more neutral formulation, though this does not address the underlying issue of the potentially biased data source. Only 2.8% assumed that the AI automatically balances different perspectives.

TABLE X. LLM04 – DATA & MODEL POISONING: RESPONSES TO CONTRADICTORY AI OUTPUT FROM AN INTERNAL KNOWLEDGE BASE ($n = 72$)

Response option	(%)
Assume the AI automatically balances different perspectives.	2.8
Use the content anyway due to time pressure.	0.0
Review the underlying data and correct or remove problematic sources.	80.6
Ask the AI to reformulate the content more neutrally.	16.7

Respondents rated their decision as fairly confident ($M = 3.88$, $SD = 0.87$) and assessed the perceived risk as low ($M = 2.47$, $SD = 0.89$). The realism score, at $M = 2.78$ ($SD = 0.95$), is at its lowest level so far, suggesting that scenarios involving manipulated internal knowledge databases are perceived by respondents as being comparatively far removed from their everyday working lives.

5) Vignette LLM05 – Improper Output Handling: Table XI shows the distribution of responses for Vignette LLM05 (*Improper Output Handling*), in which automatically generated HTML and script code is to be integrated directly into a CMS. At 77.8%, the majority of respondents chose the safest option: treating the code as potentially insecure, having it technically reviewed, and only publishing it after approval. Compared to LLM02 and LLM03, however, this proportion is slightly lower, which is also reflected in a higher spread across the remaining options. 11.1% would only skim the code superficially and insert it if no obvious anomalies were found, an approach that does not reliably rule out the risk of embedded malicious code. 6.9% opted to convert it to plain text without further checking, and 4.2% would insert the code directly.

TABLE XI. LLM05 – IMPROPER OUTPUT HANDLING: RESPONSES TO AUTOMATICALLY GENERATED HTML AND SCRIPT CODE FOR CMS INTEGRATION ($n = 72$)

Response option	(%)
Insert the code directly – it was generated for this purpose.	4.2
Skim the code and insert it if nothing seems wrong.	11.1
Treat the code as potentially unsafe, have it technically reviewed, and approve before publication.	77.8
Convert the content to plain text and publish without further review.	6.9

Respondents rated their decision as fairly certain ($M = 3.78$, $SD = 1.01$), the lowest confidence score of all five vignettes, and rated the perceived risk at $M = 2.68$ ($SD = 0.90$), the highest risk score in the entire vignette study. This suggests that, in this vignette, respondents were most likely to perceive the danger but, at the same time, were least certain about the correct response. The realism score, at $M = 2.83$ ($SD = 1.09$), falls within the mid-range, comparable to the other vignettes.

Across all five scenarios, the study reveals a consistent pattern: the majority of media professionals surveyed choose the safest course of action in hypothetical risk scenarios. The proportion choosing the safest response ranges from 68.1% (LLM01) to 87.5% (LLM02) and is well over two-thirds of respondents in all cases. It is particularly striking that in all five vignettes, less than 5.0% chose the riskiest option, which suggests a fundamental awareness of risk within the sample. The accompanying questions also reveal a consistent picture: The confidence values are consistently in the range $M = 3.78$–4.10, which corresponds to a relatively confident assessment of one's own decision. The perceived risk of the chosen course of action was consistently rated as low ($M = 2.18$–2.68), which is consistent with the choice of the safest option. LLM05 recorded the highest risk value ($M = 2.68$) alongside the lowest confidence value ($M = 3.78$), an indication that technically complex scenarios, such as the integration of AI-generated code, are perceived as more difficult to assess. The realism of the scenarios was consistently rated as moderate ($M = 2.78$–3.10), with LLM01 (*Prompt Injection*) being perceived as the scenario most akin to everyday life. Overall, the results suggest that the surveyed media professionals generally reacted in a safety-conscious manner

within the presented scenarios in situations explicitly framed as high-risk. It remains unclear, however, to what extent this behaviour is maintained under real-world time pressure and without the clear context of a test situation, a question that further studies involving the full sample should address.

V. Discussion and Limitations

A. Interpretation of Key Findings

Regarding RQ1 the results of this preliminary study provide a nuanced picture of media professionals' cybersecurity awareness and their ability to respond to LLM-specific risks. AI-powered tools, particularly ChatGPT, are already firmly integrated into everyday work practices: almost all respondents are familiar with such tools, and nearly one-third use them daily. At the same time, security-related competence appears less developed, as more than half of respondents rated their knowledge of AI risks as only average or below average, while self-efficacy within the PMT framework remained moderate (M = 3.17).

About RQ2 and the differences in terms of various risk scenarios the vignette study revealed that clear majorities, ranging from 68.1% (LLM01) to 87.5% (LLM02), selected the most security-compliant response across all five scenarios. The riskiest option was selected by less than 5% for all vignettes. This suggests that respondents demonstrate situational risk awareness when threats are explicitly visible within a structured survey context. However, it remains unclear whether similar behaviour would persist under real-world conditions characterized by time pressure, routine workflows, or less explicit risk framing. Social desirability effects may also have contributed to the observed preference for safer responses. Because respondents were explicitly presented with security-relevant situations in a survey environment, some participants may have recognized the expected secure behaviour and selected the normatively correct option. Consequently, the vignette results should be interpreted as indicators of decision-making competence under controlled conditions rather than direct evidence of actual behaviour in real-world work settings.

The technically most complex scenario, LLM05 (Improper Output Handling), produced the lowest decision confidence (M = 3.78) and the highest perceived risk (M = 2.68). This indicates that respondents perceived technical risks, such as integrating AI-generated code, as more difficult to assess than editorial or content-related risks. The discrepancy between comparatively high perceived risk and lower confidence is consistent with the moderate self-efficacy scores observed in the PMT scales and may additionally reflect the limited technical background of many respondents.

B. Implications for Secure GenAI Usage in Media Organisations

The findings have concrete implications for the secure use of GenAI in media organizations. The central practical problem lies not in respondents' fundamental ignorance of risk, but in a gap between risk recognition and the ability to act, particularly for technically complex risk types. From an organizational perspective, media companies should therefore take action on two levels.

Firstly, clear governance structures and binding guidelines for the use of AI are required. The results show that respondents rate the effectiveness of guidelines and review processes relatively highly (Response Efficacy M = 3.85), but perceive such structures in their organisations only moderately ('My organization has clear guidelines for managing risks associated with AI tools', M = 3.19). There is a concrete need for action here: organizations that establish clear guidelines for the safe use of AI can translate existing motivation for safe use into consistent behaviour.

Secondly, training programs should be specifically designed to strengthen self-efficacy, that is, not only to impart declarative knowledge about risks, but also to develop the practical ability to recognize LLM-specific threats in real-world work situations and respond appropriately. Scenario-based training formats, such as those used in this study, appear particularly suitable for this purpose, as they link cognitive risk assessment with action-oriented decision-making.

C. Theoretical Contributions to PMT-Based Security Research

This study offers an exploratory contribution to PMT-based cybersecurity research by applying the framework to selected GenAI and LLM risks in the media industry. The results confirm the fundamental suitability of the PMT framework for this new area of application: the constructs Perceived Severity, Response Efficacy, Self-Efficacy, and Response Costs demonstrate a plausible differentiation in terms of content and behave in accordance with the theory. It should be noted, however, that the derivation of the two measurement items used in each case was rather pragmatic. In future work, the existing literature should be evaluated more thoroughly to base the measurement on empirically validated items or to derive appropriate item sets systematically using existing approaches.

Furthermore, the combination of PMT constructs with an OWASP-based set of vignettes expands the methodological repertoire of PMT research: whilst previous studies have predominantly used PMT to predict behavioural intentions, the approach chosen here allows for the simultaneous measurement of threat appraisal, coping appraisal and situational competence. This design can serve as a template for future studies investigating the relationship between PMT constructs and actual decision-making behaviour in AI-related security contexts. However, the examination of statistical correlations between PMT constructs and vignette decisions is reserved for the full sample.

D. Limitations: Panel and Self-Selection Bias

The present results must be interpreted in light of several methodological limitations. The data are based on a pilot study with n = 72 participants recruited via a commercial online panel. This approach carries the risk of panel bias: individuals who regularly participate in online surveys may differ systematically from the general population of media professionals in Germany, for example, in terms of their

affinity for technology or their willingness to reflect on their competencies.

Furthermore, a self-selection bias must be assumed: the decision to participate in a study on AI risks is likely to attract a disproportionately high number of people who are already open-minded about the topic or interested in security issues. This could at least partially explain the observed high proportions of security-compliant vignette decisions and overstate the actual risk perception across the media industry as a whole. It should also be noted that the vignette design captures deliberate behaviour in a test situation, not real behaviour in a work context, a problem of social desirability cannot be ruled out.

E. Generalizability Beyond the Media Sector

The generalisability of the findings is limited in several respects. Within the media sector, print, and publishing clearly dominate, accounting for around 68% of the sample, whilst other sub-sectors, such as online/digital, broadcasting, or film production, are represented by only very small sample sizes. Valid conclusions about the media sector as a whole cannot, therefore, be drawn.

An additional limitation concerns the strong representation of print and publishing professionals within the sample. As newsroom structures, workflows, and AI adoption practices differ across media sectors, the findings may primarily reflect the perspectives of professionals working in publishing-related environments rather than the broader media industry.

Beyond the media industry, caution is advised when generalising the results. Media professionals face specific job requirements, in particular, the routine handling of confidential information, source protection, and high-profile content, which are likely to influence their perception of LLM risks and their response to corresponding scenarios. Whether comparable patterns occur in other sectors with intensive LLM use, such as healthcare, legal services or the financial sector, remains an open empirical question. Future studies should therefore include cross-sector and cross-country designs to test the robustness and external validity of the findings presented here.

VI. Conclusion and Future Research

A. Conclusion

This study examined the cybersecurity awareness and action competence of media professionals in Germany regarding selected risks from the OWASP Top 10 for LLM Applications framework. With rapidly increasing integration of GenAI and LLMs into media workflows, the study addressed an important research gap concerning the human-centred security implications of AI usage in professional media contexts.

The findings demonstrate that AI-supported tools, particularly ChatGPT and comparable LLM systems, are already firmly embedded in the daily work practices of many media professionals. At the same time, the results reveal a discrepancy between widespread AI adoption and the level of security-related knowledge and confidence. Although respondents generally perceived AI-related risks as serious and relevant, their self-assessed competence in identifying and handling LLM-specific threats remained only moderate. Nevertheless, the vignette-based analyses showed that the majority of participants selected security-conscious responses across all OWASP-related risk scenarios, indicating a fundamental awareness of secure behaviour when risks are explicitly recognizable.

The study further contributes theoretically by demonstrating the applicability of PMT in the context of GenAI and LLM security. The combination of PMT constructs with OWASP-aligned vignette scenarios provides a structured framework for analysing the relationship between perceived threats, coping mechanisms, and practical decision-making competence in AI-supported environments. In particular, the findings suggest that Perceived Severity and Response Efficacy are comparatively high, whereas Self-Efficacy remains less developed, emphasizing the importance of practical training and organizational support.

From a practical perspective, the results underline the necessity for media organizations to establish clear governance structures, security guidelines, and targeted awareness programs addressing LLM-specific cybersecurity risks. As AI-supported workflows become increasingly integrated into editorial and production processes, strengthening human-centred security competence becomes essential for maintaining organizational resilience and information integrity.

B. Future Work

Future research should extend the study to larger and more diverse samples, including additional media sectors and international contexts. Once the full dataset becomes available, inferential statistical analyses should investigate relationships between PMT constructs, AI usage intensity, and security-related decision-making competence. In particular, future work should analyse whether Perceived Severity, Self-Efficacy, or Organizational Support significantly influence secure behaviour in LLM-related risk situations.

Future studies should complement questionnaire-based assessments with behavioural and experimental research designs. For example, participants could be exposed to realistic AI-supported work tasks involving time pressure, incomplete information, and competing work priorities. Such approaches would allow researchers to observe actual security-related decision-making behaviour and assess whether the security-conscious responses observed in vignette scenarios are maintained under more realistic working conditions.

Moreover, longitudinal and intervention-based studies could evaluate the effectiveness of targeted cybersecurity awareness programs and AI governance strategies within media organizations. Experimental studies simulating real-world time pressure and operational constraints may further improve understanding of how media professionals behave in authentic AI-supported work environments. For example, future research could manipulate available decision time, introduce competing tasks, or simulate breaking-news situations in which participants must evaluate AI-generated content under strict

deadlines. Finally, future research may expand the OWASP-aligned vignette framework to additional LLM risks and compare behavioural patterns across different professional domains beyond the media sector.

ACKNOWLEDGEMENTS

The development of the vignettes, the derivation of scenarios and courses of action from the OWASP Top 10 LLM Risks, and the design of the corresponding questionnaire items were supported by ChatGPT (OpenAI). The analysis of the collected data and the presentation of the results (e.g., generation of tables from survey data in Excel) were carried out with the support of Claude.ai (Anthropic). The authors are solely responsible for all decisions regarding the research design, the interpretation of the results, and the final drafting of the manuscript.

references.bib

REFERENCES

[1] A. S. M. Alshawi, "The Impact of Artificial Intelligence in Digital Media on the Reality of Content Production and Journalists' Perceptions of Their Functions and Roles", *International Journal of Environmental Sciences*, pp. 1006–1016, Sep. 2025, ISSN: 2229-7359. DOI: 10.64252/58fen002.

[2] OWASP Foundation, *OWASP Top 10 for LLM Applications 2025*, Version 2025, released November 18, 2024, Nov. 2024. [Online]. Available: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/ (visited on 02/10/2026).

[3] B. Sarrionandia, S. Peña-Fernández, J. Ángel Pérez-Dasilva, and A. Larrondo-Ureta, "Artificial intelligence training in media: Addressing technical and ethical challenges for journalists and media professionals", *Frontiers in Communication*, vol. 10, p. 1 537 918, 2025.

[4] A. Wieczorek, K. Postrzednik-Lotko, *et al.*, "Machine learning algorithms on social media: Privacy risks, user awareness and security implications", *Safety and Security Advances*, vol. 1, no. 1, pp. 18–43, 2025.

[5] D. Dettman, "A study on the knowledge and perception of artificial intelligence", *Evidence Based Library and Information Practice*, vol. 19, no. 2, pp. 139–141, Jun. 2024. DOI: 10.18438/eblip30436.

[6] A. Milani *et al.*, "When AI is fooled: Hidden risks in llm-assisted grading", *Education Sciences*, vol. 15, no. 11, 2025, ISSN: 2227-7102. DOI: 10.3390/educsci15111419.

[7] M. Weinz, N. Zannone, L. Allodi, and G. Apruzzese, "The impact of emerging phishing threats: Assessing quishing and llm-generated phishing emails against organizations", in *Proceedings of the 20th ACM Asia Conference on Computer and Communications Security*, ser. ASIA CCS '25, Association for Computing Machinery, 2025, pp. 1550–1566, ISBN: 9798400714108. DOI: 10.1145/3708821.3736195.

[8] C. Olea *et al.*, "Evaluating phishing email efficacy", in *Proceedings of the 2025 Computers and People Research Conference*, ser. SIGMIS-CPR '25, Association for Computing Machinery, 2025, ISBN: 9798400714979. DOI: 10.1145/3716489.3728437.

[9] D. Tworzydło, N. Życzyński, C. Olexová, and P. Szuba, "Threats from the web and cybersecurity issues from the perspective of public relations professionals", *Humanities and Social Sciences*, vol. 30, no. 4-part 1, pp. 279–293, 2023.

[10] J. Fu, Y. Lu, Z. Yang, F. Nah, and R. LC, "Cracking aegis: An adversarial llm-based game for raising awareness of vulnerabilities in privacy protection", in *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, 2025, pp. 639–662.

[11] S. E. McGregor, F. Roesner, and K. Caine, "Individual versus organizational computer security and privacy concerns in journalism", *Proceedings on Privacy Enhancing Technologies*, pp. 418–435, 2016.

[12] R. W. Rogers, "A protection motivation theory of fear appeals and attitude change", *Journal of Psychology*, vol. 91, no. 1, pp. 93–114, 1975.

[13] J. E. Maddux and R. W. Rogers, "Protection motivation and self-efficacy: A revised theory of fear appeals and attitude change", *Journal of Experimental Social Psychology*, vol. 19, no. 5, pp. 469–479, 1983.

[14] M. A. Helmiawan *et al.*, "Quantitative analysis of the key factors driving cybersecurity awareness among information systems users", *Repository FTI UNSAP*, vol. 25, no. 1, pp. 1897–1910, 2025.

[15] H.-C. Pham, I. Ulhaq, M. Nkhoma, M. N. Nguyen, and L. Brennan, "Exploring knowledge sharing practices for raising security awareness", in *Australasian Conference on Information Systems 2018*, UTS ePRESS, 2018, pp. 1–7.

[16] T. Herath and H. R. Rao, "Protection motivation and deterrence: A framework for security policy compliance in organisations", *European Journal of Information Systems*, vol. 18, no. 2, pp. 106–125, 2009.

[17] P. Ifinedo, "Understanding information systems security policy compliance: An integration of the theory of planned behavior and the protection motivation theory", *Computers & Security*, vol. 31, no. 1, pp. 83–95, 2012.

[18] A. Vance, M. Siponen, and S. Pahnila, "Motivating IS security compliance: Insights from habit and protection motivation theory", *Information & Management*, vol. 49, no. 3-4, pp. 190–198, 2012.