Fast, Valued, and Trusted with Reservations: Employee Perceptions of Microsoft 365 Copilot in Administrative and Knowledge Work at a Large Non-Profit Welfare Organization

Alexander Loth[a,e,*], David Wester[b], Hannah Cha[d,e], Marc-Oliver Pahl[c]

[a] *Frankfurt University of Applied Sciences, Nibelungenplatz 1, 60318, Frankfurt am Main, Germany*
[b] *Malteser Hilfsdienst e.V., Erna-Scheffler-Str. 2, 51103 Cologne, Germany*
[c] *IMT Atlantique, UMR IRISA, Chaire Cyber CNI, Brest, France*
[d] *Stanford University, Stanford, 94305, CA, USA*
[e] *Microsoft, One Microsoft Way, Redmond, 98052, WA, USA*

Abstract

Enterprise generative AI assistants are being rolled out at scale, yet little is known about how employees in non-profit organizations use them, how far they trust them, and whether they verify what these assistants produce. We report a descriptive field study: an employee survey on a Microsoft 365 Copilot deployment at a large German non-profit welfare organization (N = 311 of 1,093 invited employees; 288 answered all substantive blocks). Respondents work predominantly in administrative and staff functions, so the findings describe administrative and knowledge work rather than the organization as a whole. Respondents report strong perceived speed gains (85% agree at least somewhat that Copilot helps them complete tasks faster) and high overall satisfaction (85%), while trust is more differentiated: 68% rate Copilot as competent, and 37% at least somewhat agree that they trust it in general. Most respondents verify outputs before reuse (57% always), while fewer routinely open the cited sources (18% always). Self-rated knowledge of LLM hallucinations is low for 39% of respondents, while 69% report being at least fairly aware of misinformation risks when using LLMs. Dependence on a single technology vendor is the most pronounced ethical concern (59% at least rather strong). Open answers describe Copilot as a drafting, summarizing and sparring partner and name wrong results, verification effort, prompting and training effort, vendor dependence, overreliance and data protection as main challenges. Read through the lens of trust calibration and automation bias, the results point to verification routines, source-grounded design and rollout training.

*Corresponding author.
Email addresses: alexander.loth@stud.fra-uas.de (Alexander Loth), david.wester@malteser.org (David Wester), hcha417@stanford.edu (Hannah Cha), marc-oliver.pahl@imt-atlantique.fr (Marc-Oliver Pahl)

Keywords: generative AI, Microsoft 365 Copilot, trust calibration, automation bias, verification, misinformation, non-profit organizations

1. Introduction

Generative AI assistants embedded in productivity suites move large language models (LLMs) from experimental use into everyday knowledge work. Microsoft 365 Copilot is one of the most widely deployed such assistants and operates directly on an organization's mail, documents, meetings and chats. Evidence on its use comes mostly from vendor reports and from controlled and field studies in commercial firms (Brynjolfsson, Li, & Raymond, 2025; Dell'Acqua et al., 2023; Dillon, Jaffe, Immorlica, & Stanton, 2025; Noy & Zhang, 2023). Survey evidence from the public sector is emerging (Bright et al., 2025), but non-profit welfare organizations, which handle sensitive beneficiary data and are bound to explicit value commitments, are largely absent from this evidence base.

At the same time, research on automation and on misinformation shows that fluent, confident output can mislead users who do not verify it (Ji et al., 2023; Parasuraman & Manzey, 2010). Readers also cannot reliably tell LLM-generated from human-written news text (Loth, Kappes, & Pahl, 2026a). In an enterprise assistant, this risk shifts from public information ecosystems into internal documents, reports and decisions. Whether employees calibrate their trust and check the sources an assistant cites is therefore both a productivity question and an information integrity question.

This paper reports a descriptive institutional field study: an employee survey on a Microsoft 365 Copilot deployment at Malteser, a large German Catholic welfare organization. We are not aware of published survey evidence on Microsoft 365 Copilot use in a non-profit welfare organization that combines use, value, trust, verification and value-based concerns. Because only aggregated item-level data were used, the study describes perceptions and reported behavior; it does not test a theoretical model. We read the findings through the lens of trust calibration and automation bias (J. D. Lee & See, 2004; Parasuraman & Manzey, 2010). We address four research questions:

RQ1 How intensively and for which tasks do employees use Copilot?

RQ2 How do they rate its value?

RQ3 How far do they trust Copilot, how aware are they of hallucination and misinformation risks, and how deeply do they verify outputs and cited sources?

RQ4 Which barriers, ethical concerns and perceptions of mission fit shape adoption?

The paper makes three contributions. First, it provides survey evidence on Microsoft 365 Copilot use in a non-profit welfare organization, covering use, value, trust, verification and value-based concerns. Second, it suggests that

verification concentrates on the output rather than on the sources the assistant cites, partly because text generation, a frequent use case, often yields no sources to check. Third, it derives implications for training and for assistant design.

2. Related work

2.1. Generative AI and work outcomes

Experimental and field evidence shows sizable productivity effects of generative AI on knowledge work. In a preregistered experiment on professional writing tasks, access to ChatGPT reduced completion time and raised output quality (Noy & Zhang, 2023). In a large customer support deployment, an AI assistant increased productivity, with the largest gains for less experienced workers (Brynjolfsson et al., 2025). A field experiment with management consultants found that AI improved performance on tasks inside the model's capability frontier but lowered correctness on a task outside it, so that the same tool helps or harms depending on the task (Dell'Acqua et al., 2023). A cross-firm field experiment with Microsoft 365 Copilot reported shifts in how employees allocate time across mail, documents and meetings (Dillon et al., 2025). These studies measure outcomes in commercial settings. They say less about how employees themselves judge the assistant, and they do not cover welfare organizations.

2.2. Public-sector and welfare settings

Research on AI in government has mapped the use of AI by governments across the European Union (van Noordt & Misuraca, 2022) and has examined the challenges of AI-supported decision-making in areas such as social welfare, including the role of humans in the loop (Fischer-Abaigar, Kern, Barda, & Kreuter, 2024). For generative AI specifically, a survey of UK public service professionals in education, health, social work and emergency services found that 22% used generative AI themselves, that a majority expressed trust in it, and that only about a third felt there was clear guidance on its use at their workplace (Bright et al., 2025). Welfare organizations operate close to these settings but differ in governance, funding and value commitments; we know of no comparable survey evidence from them.

2.3. Acceptance and task-technology fit

Perceptions of usefulness and ease of use are established predictors of technology acceptance (Davis, 1989), integrated with social influence and facilitating conditions in UTAUT (Venkatesh, Morris, Davis, & Davis, 2003). Task-technology fit theory adds that performance gains depend on the match between a tool's capabilities and the tasks it is used for (Goodhue & Thompson, 1995). Our instrument draws on these traditions for its effectiveness, ease of use, satisfaction and barrier blocks. We use them as a descriptive lens, since our aggregated data do not permit structural tests.

2.4. Trust, automation bias and verification

Trust in an agent rests on perceived ability, benevolence and integrity (Mayer, Davis, & Schoorman, 1995). Research on automation distinguishes appropriate use from misuse through overreliance and disuse through undertrust (Parasuraman & Riley, 1997), and argues that trust should be calibrated to the system's actual reliability (J. D. Lee & See, 2004). Automation bias and complacency describe how users of reliable-seeming automation reduce their monitoring and accept its output without sufficient checking (Parasuraman & Manzey, 2010). This applies to LLM-based systems, where prior work finds that users tend to overestimate the accuracy of LLM responses (Steyvers et al., 2025), and explanations can increase reliance on both correct and incorrect LLM responses (Kim, Vaughan, Liao, Lombrozo, & Russakovsky, 2025). Thus, explanations alone do not reliably prevent overreliance on AI advice, whereas interventions that force users to engage with the decision can (Buçinca, Malaya, & Gajos, 2021). In a survey of 319 knowledge workers, higher confidence in generative AI was associated with less reported critical thinking, while higher self-confidence was associated with more (H.-P. Lee et al., 2025). Users also tend to overestimate their own competence in judging AI advice, which can hinder appropriate reliance (He, Kuiper, & Gadiraju, 2023). Verification behavior, and in particular whether users inspect the sources an assistant cites, is therefore a central behavioral indicator of calibrated trust.

2.5. Hallucination, sycophancy and misinformation

Large language models produce fluent content that is unfaithful to sources or factually wrong (Ji et al., 2023). Models trained with human feedback also tend to agree with users' stated views and to abandon correct answers under pushback (Sharma et al., 2023). Both behaviors turn an assistant into a possible source of misinformation inside an organization. Experts on AI disinformation see different threat profiles in health and in politics and favor domain-specific over uniform countermeasures (Loth, Kappes, & Pahl, 2026c). This argues for studying specific deployment contexts rather than generic use, such as welfare organizations whose work spans health and care.

3. Method

We conducted an anonymous, cross-sectional employee survey and analyzed aggregated item-level frequencies. This section describes the setting, the instrument, the sample, the analysis, and ethics and governance.

3.1. Setting and procedure

The survey was conducted by Malteser's Copilot community team in cooperation with Frankfurt University of Applied Sciences and IMT Atlantique; the invitation also named Microsoft Elevate as a partner. Malteser began rolling out Copilot in March 2024. All 1,093 employees in the target population were invited by email, and fieldwork ran from 31 July to 30 September 2026. Participation

was voluntary and anonymous and took about 15 to 20 minutes. The invitation stated that the results would inform internal decisions, academic research, and the vendor's product and service development. The target population comprised all employees holding a paid Microsoft 365 Copilot licence plus employees who had registered their interest through an internal form.

3.2. Instrument

The German-language questionnaire comprised 75 items in 13 blocks: licensing and tenure, use intensity, task categories (11 items), perceived effectiveness and task-technology fit (5), information quality (4), verification behavior (4), trust (6), ease of use (3), satisfaction (5), barriers and enablers (7), mission fit (4), ethical concerns (4), hallucination knowledge and misinformation risk awareness (2 plus one open item), three open questions, and demographics. Most attitude items used seven-point labeled scales; two verification frequency items used a six-point scale from *never* to *always*. Item wording and the deviations from the design version are documented in the supplementary codebook.

3.3. Sample

Of the 1,093 invited employees, 311 responded (28%), and 288 of them (93%) answered all substantive blocks. Item-level N ranges from 311 to 288 for closed substantive items and from 253 to 280 for demographic items; all shares below use the item-level N as denominator. Most respondents primarily use the paid M365 Copilot version (83%), have used Copilot for more than six months (66%), and received at least one training (68%). Respondents work mainly in administration (39%) and other administrative or staff functions (16%), are 35 to 44 years old (36%), and by self-identified gender are 51% male and 44% female. Taken together, 55% work in administrative functions, whereas only 1% work in nursing, and 54% work in North Rhine-Westphalia.

3.4. Analysis

Respondent-level data are held by Malteser on the survey platform; for this analysis the authors used only aggregated item-level frequencies exported from the survey platform. We therefore report distributions, top-box shares and item-level N, and we do not report correlations, scale reliabilities or group comparisons. Every share was recomputed from the absolute counts with a single rounding step; the script, the frequency table and the codebook are openly available on Zenodo (see Data availability).Open answers were analyzed with reflexive thematic analysis (Braun & Clarke, 2006). The first author coded all answers inductively and iteratively refined the themes; consistent with reflexive thematic analysis, no intercoder agreement was computed (Braun & Clarke, 2019).

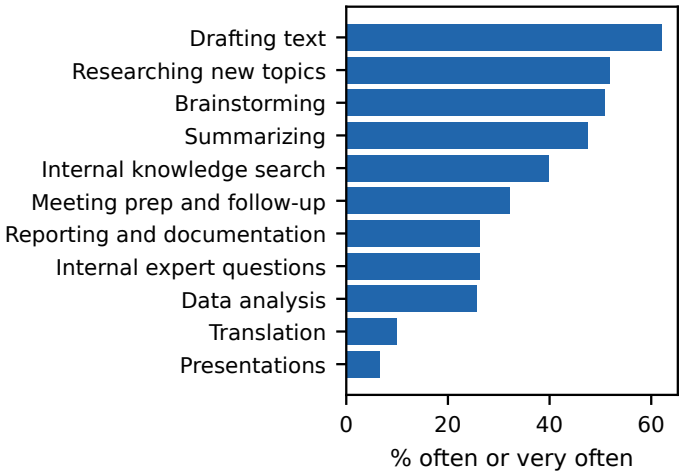


Figure 1: Task categories: share of respondents using Copilot often or very often (n = 301).

3.5. Ethics, data protection and governance

Participation was voluntary and anonymous, and participants were informed about the purpose of the survey, the cooperation partners and the use of the results in the invitation. Data processing followed Malteser's privacy notice for surveys. The academic authors received only aggregated frequencies and the anonymous free-text answers; no personal data were transferred to the academic partners. Formal ethics approval was not sought because the survey was anonymous and voluntary, collected no personal data, and no personal data were transferred to the academic partners.

4. Results

We report results by research question. All shares use the item-level N as denominator.

4.1. RQ1: Use

Copilot is used regularly: 62% use it at least once a day, and 71% spend at least one hour per week with it. Outlook (67%), the standalone Copilot Chat (67%) and Word (55%) are the most common entry points. For 61% of respondents, Copilot accounts for more than three quarters of their AI use. Drafting text is the dominant task (62% often or very often), followed by researching new topics (52%), brainstorming (51%) and summarizing (48%) (Fig. 1).

4.2. RQ2: Perceived value

Effectiveness ratings are high (Fig. 2). Eighty-five percent agree at least somewhat that Copilot helps them work faster, 78% that it improves their overall

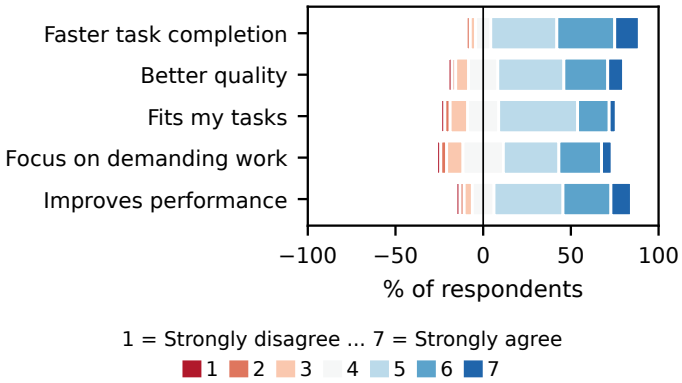


Figure 2: Perceived effectiveness and task-technology fit (n = 300).

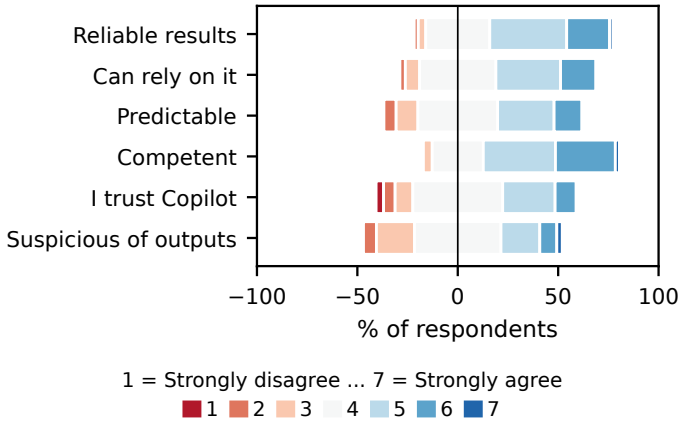


Figure 3: Trust in Copilot (n = 295).

performance, and 72% that it improves quality. Output quality is rated good rather than excellent: 75% rate correctness at least somewhat good, but only 2% very good. Ease of use (82% at least somewhat easy) and overall satisfaction (85%) are high. In open answers, respondents most often describe time savings, drafting from bullet points, summarizing long mail threads and meetings, finding internal content, and using Copilot as a sparring partner.

4.3. RQ3: Trust, risk awareness and verification

Trust is reserved (Fig. 3). Competence (68% at least somewhat agree) and reliability (61%) are rated favorably, but the general statement "I trust Copilot" draws a plurality of *partly, partly* answers (45%); 37% at least somewhat agree with it, and 31% at least somewhat agree that they are suspicious of its outputs.

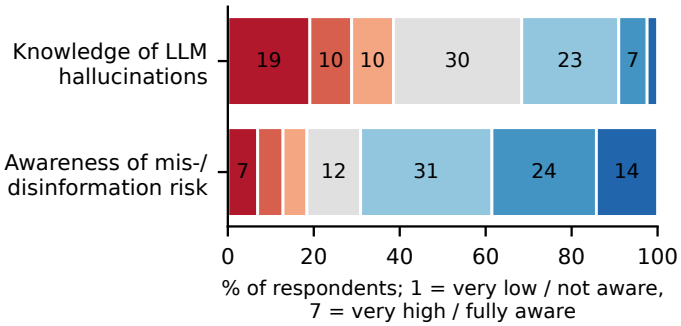


Figure 4: Self-rated knowledge of LLM hallucinations and awareness of mis- or disinformation risk when using LLMs (n = 288). Both items refer to LLMs in general, not to Copilot. Two separate seven-point items, shown as distributions; not linked per respondent.

We report the trust items, the three verification items and the two awareness items as separate distributions. Because the data are aggregated, we cannot link them per respondent: we do not know whether the same people who report low hallucination knowledge also rarely check sources.

Both awareness items refer to large language models in general rather than to Copilot: respondents rated their knowledge of LLM hallucinations and their awareness of the risk of mis- or disinformation when using LLMs. Stated risk awareness exceeds self-rated knowledge (Fig. 4). With symmetric thresholds, 69% are at least fairly aware of the risk (199 of 288), whereas 32% rate their knowledge of LLM hallucinations as at least rather high (91 of 288). Conversely, 39% rate their knowledge as rather low or lower, and 19% as very low or none.

Most respondents verify outputs, and fewer open the cited sources (Fig. 5). Fifty-seven percent always verify outputs before reuse (72% very often or always), and 18% always open the sources Copilot cites (41% very often or always). At the same time, 27% at least somewhat agree that they often adopt Copilot text without major edits. Checking whether a cited document supports a statement is the step that most directly exposes ungrounded output in a retrieval-based assistant, and it is the step respondents report least often.

Open answers give this pattern a concrete form (Table 1). Respondents describe wrong or overgeneralized answers on topics they know, cross-checking as routine, sources that make checking easier, the model's tendency to agree with the user, and colleagues who adopt outputs unchecked. Answers to the question on what respondents would change about Copilot (n = 146) mostly concern features and integration, such as better presentations, image generation and access to internal systems. Answers on the greatest challenge or concern (n = 202) most often name wrong results (17 answers), verification effort (16), prompting and training effort (15), vendor dependence (12), overreliance (11) and data protection (9). A recurring wish within them is that Copilot should flag uncertain or questionable answers and state openly when it does not know something.

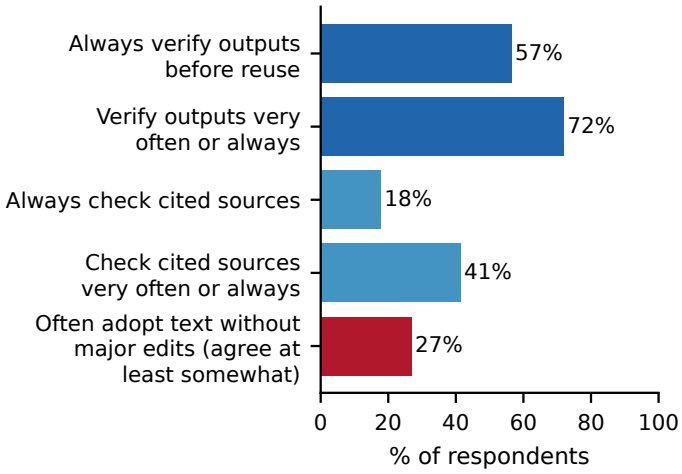


Figure 5: Depth of verification (n = 297). Output verification and source checking use a six-point frequency scale; adoption without edits uses a seven-point agreement scale.

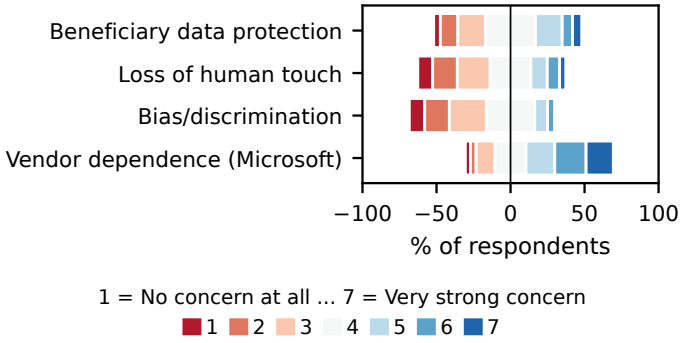


Figure 6: Ethical concerns (n = 288).

4.4. RQ4: Barriers, mission fit and concerns

Clarity on permitted uses is high (72% at least somewhat agree), while sufficient training (49%) and time to learn (45%) are the weakest enablers. Mission fit is rated less uniformly than the absence of conflict suggests: 82% see no conflict between using Copilot and their work at Malteser, 63% agree at least somewhat that Copilot helps them contribute to Malteser's mission, and 61% that it frees time where it matters most for the people they work with, but only 55% agree that its use fits Malteser's values well, with 36% choosing *neither*. Among ethical concerns, dependence on a single technology vendor ranks highest (59% at least rather strong concern, 40% strong or very strong), ahead of beneficiary data protection (31%), loss of the human touch (24%) and bias (15%) (Fig. 6).

Table 1: Illustrative open answers on information integrity (translated from German by the authors). Source question in brackets: M = reasons for the misinformation rating, C = greatest challenge or concern, W = what respondents would change.

Theme	Quote
Wrong output	"On test questions about topics I know well, Copilot sometimes gives wrong or very generalized answers." [M]
Verification	"I am aware that many models, this one included, tend to want to please the user. That is why I almost always cross-check." [M]
Wrong output	"Hallucinations, wrong citations or information" [C]
Uncertainty flag	"The standard flagging of uncertain or questionable answers" [W]
Over-reliance	"Unfortunately, especially in administration, I see many Malteser who use the technology 'blindly'. They type in their work tasks and adopt the results unchecked, so that since the large-scale rollout of MS Copilot a clear deterioration in the quality of work in these areas can be observed." [C]

5. Discussion

Four observations stand out. First, perceived productivity outruns trust: most employees use Copilot daily and value it, yet hold back from general trust. This pattern is consistent with reserved rather than unconditional trust. It does not show that trust is calibrated, because calibration cannot be established without measuring Copilot's actual reliability on respondents' tasks (J. D. Lee & See, 2004), and 27% at least somewhat agree that they often adopt Copilot texts without major edits.

Second, the data suggest that verification focuses on the output more than on its grounding, partly because text generation, a frequent use case, often yields no sources to open. In automation-bias terms, checking that a text reads plausibly is a weaker form of monitoring than checking the evidence behind it (Parasuraman & Manzey, 2010). Fluent model output can bypass readers' source monitoring, so that they rely on cues such as fluency rather than on verified source information (Loth, Kappes, & Pahl, 2026b). Although providing sources can reduce reliance on incorrect LLM responses (Kim et al., 2025), such an intervention is only effective to the extent that users actually engage with the sources. This distinction is also relevant with recent work that argues that output-level audits can fall short of emergent behavior that occur through users' ongoing interactions within personalized systems (Cha, 2026). Copilot provides source references that make the stronger check possible, and one respondent notes that they make checking relatively easy; the open question is how to make opening them routine. Respondents' own suggestions point the same way: several ask Copilot to flag uncertain answers. This applies to retrieval-augmented assistants in general, and training that teaches source checking, that is, opening the cited source and judging whether it supports the claim and whether the reasoning holds, together with design that keeps it effortless, follows directly.

Third, the gap between low self-rated hallucination knowledge and high stated risk awareness, as two separate distributions, indicates that general risk awareness does not by itself translate into knowledge of how hallucinations arise. Whether a targeted awareness campaign on source checking would close this gap is worth discussing; we did not measure the effect of such campaigns. Experts on AI disinformation largely agree on the threats but are split on countermeasures, including media literacy; they argue for layered rather than single-strategy responses (Loth, Schütz, Kappes, & Pahl, 2026). Training is therefore one layer, not a sufficient remedy on its own.

Fourth, dependence on a single technology vendor is the strongest ethical concern in the closed items and is also named in the open answers.

Compared with public-sector survey evidence, the picture is similar in that use is widespread and respondents expect efficiency gains (Bright et al., 2025). Items, populations and the stage of rollout differ, so we do not compare shares directly.

5.1. Limitations

The study covers a single organization and relies on self-reported, cross-sectional data. With a response rate of 28%, self-selection of interested users is likely, in particular because the survey was promoted through the internal Copilot community; perceived value may therefore be overstated. Employees who registered their interest were selected into the population by their own initiative. We could not compare respondents with the invited population, because a breakdown of the 1,093 invited employees was not available. The sample is concentrated in administrative functions (55%) and in North Rhine-Westphalia (54%), with only 1% from nursing, so the findings apply to administrative and knowledge work rather than to the welfare organization as a whole. The invitation disclosed that results would inform internal decisions, academic research, and the vendor's product development, which may have shaped answers. The hallucination knowledge and risk awareness items asked about LLMs rather than Copilot, and several respondents wrote in the free-text answers that they do not know the term LLM; this may have depressed or distorted answers on both items. The analysis used only aggregated item-level frequencies, so we report no correlations, for example between trust and verification; the survey platform allows filtering by answer options, and such an analysis is planned. One negated verification item shows a bimodal pattern suggestive of misreading and is not interpreted. Finally, two authors are affiliated with the vendor of the product studied (see the declaration of competing interests).

6. Conclusion

In administrative and knowledge work at a large German non-profit welfare organization, Microsoft 365 Copilot has become an everyday tool for drafting, researching, brainstorming and summarizing, and employees credit it with clear speed gains and high satisfaction. Their trust is nonetheless reserved, and their

verification appears to concentrate on the output rather than on its sources, partly because many outputs come without sources. Self-rated knowledge of LLM hallucinations lags behind stated awareness of misinformation risk, and dependence on a single vendor is the strongest ethical concern. For organizations, this points to training that teaches source checking as a concrete routine rather than general caution. For assistant design, it points to making cited evidence easier to inspect at the moment an output is reused. Future work should link these perceptions to individual-level response data and to observed use, and replicate the survey in other non-profit settings.

CRediT authorship contribution statement

Alexander Loth: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review and editing. **David Wester:** Conceptualization, Data curation, Investigation, Methodology, Project administration, Resources, Writing – review and editing. **Hannah Cha:** Writing – original draft, Writing – review and editing. **Marc-Oliver Pahl:** Methodology, Supervision, Writing – review and editing.

Declaration of competing interest

Alexander Loth is employed by Microsoft, the vendor of the product studied; the work was conducted as part of his doctoral research at Frankfurt University of Applied Sciences. Hannah Cha is affiliated with Microsoft. Microsoft Elevate, which supports Malteser as a non-profit customer, was named as a cooperation partner in the survey invitation. Beyond the affiliation of two authors, Microsoft was not involved in the study design, data collection, analysis, interpretation or the decision to publish. David Wester is employed by Malteser and was responsible for the survey as its owner; Malteser owns the survey data. The study was carried out in cooperation with Malteser, Frankfurt University of Applied Sciences and IMT Atlantique. Marc-Oliver Pahl declares no competing interests.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics statement

The survey was conducted by Malteser as an anonymous internal employee survey. Participation was voluntary; participants were informed in the invitation about the purpose of the study, the cooperation partners and the use of the results, and consented by participating. Data processing followed Malteser's

privacy notice (section 4, no. 4). Formal ethics approval was not sought because the survey was anonymous and voluntary, collected no personal data, and no personal data were transferred to the academic partners. The academic authors received only aggregated frequencies and anonymous free-text answers.

Data availability

The aggregated, anonymous item-level frequency table, the codebook, the analysis scripts and the figures are openly available on Zenodo at `https://doi.org/10.5281/zenodo.24285326` (CC BY 4.0; code under MIT license). Respondent-level data and the raw free-text answers are owned by Malteser and cannot be shared publicly to protect participant anonymity.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used Claude Opus 5.5 via GitHub Copilot for copy editing. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. doi: 10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. doi: 10.1080/2159676X.2019.1628806

Bright, J., Enock, F., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2025). Generative AI is already widespread in the public sector: Evidence from a survey of UK public sector professionals. *Digital Government: Research and Practice*, *6*(1), 1–13. doi: 10.1145/3700140

Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, *140*(2), 889–942. doi: 10.1093/qje/qjae044

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. doi: 10.1145/3449287

Cha, H. (2026). Identifying harm in personalized, generative ai systems requires user-centered auditing at the interaction level. In *Proceedings of the aaai/acm conference on ai, ethics, and society* (Vol. 9, pp. 617–627).

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319–340. doi: 10.2307/249008

Dell'Acqua, F., McFowland III, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., ... Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality.* SSRN Working Paper. doi: 10.2139/ssrn.4573321

Dillon, E. W., Jaffe, S., Immorlica, N., & Stanton, C. T. (2025). *Shifting work patterns with generative AI* (NBER Working Paper No. 33795). National Bureau of Economic Research. doi: 10.3386/w33795

Fischer-Abaigar, U., Kern, C., Barda, N., & Kreuter, F. (2024). Bridging the gap: Towards an expanded toolkit for AI-driven decision-making in the public sector. *Government Information Quarterly*, *41*(4), 101976. doi: 10.1016/j.giq.2024.101976

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly*, *19*(2), 213–236. doi: 10.2307/249689

He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing about knowing: An illusion of human competence can hinder appropriate reliance on AI systems. In *Proceedings of the 2023 chi conference on human factors in computing systems* (pp. 1–18). doi: 10.1145/3544548.3581025

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. doi: 10.1145/3571730

Kim, S. S., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering appropriate reliance on large language models: The role of explanations, sources, and inconsistencies. In *Proceedings of the 2025 chi conference on human factors in computing systems* (pp. 1–19).

Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 chi conference on human factors in computing systems* (pp. 1–22). doi: 10.1145/3706598.3713778

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

Loth, A., Kappes, M., & Pahl, M.-O. (2026a). Can humans tell? A dual-axis study of human perception of LLM-generated news. In *Companion publication of the 2026 18th acm web science conference* (pp. 34–39). New York, NY, USA: ACM. doi: 10.1145/3795513.3807431

Loth, A., Kappes, M., & Pahl, M.-O. (2026b). Eroding the truth-default: A causal analysis of human susceptibility to foundation model hallucinations and disinformation in the wild. In *Companion proceedings of the acm web conference 2026* (pp. 1125–1132). New York, NY, USA: ACM. doi: 10.1145/3774905.3795832

Loth, A., Kappes, M., & Pahl, M.-O. (2026c). Experts disagree on how to fight AI disinformation, but agree that health and politics need different solutions. *Harvard Kennedy School Misinformation Review*, *7*(4). doi: 10.37016/mr-2020-205

Loth, A., Schütz, M., Kappes, M., & Pahl, M.-O. (2026). The mitigation

paradox: How AI-disinformation experts converge on threats but polarize on solutions. In *Proceedings of the AAAI/ACM conference on AI, ethics, and society* (Vol. 9, pp. 1722–1733). AAAI Press. doi: 10.1609/aies.v9i3.43261

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The Academy of Management Review*, *20*(3), 709–734. doi: 10.2307/258792

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, *381*(6654), 187–192. doi: 10.1126/science.adh2586

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. doi: 10.1518/001872097778543886

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., ... Perez, E. (2023). *Towards understanding sycophancy in language models.* arXiv:2310.13548. doi: 10.48550/arXiv.2310.13548

Steyvers, M., Tejeda, H., Kumar, A., Belem, C., Karny, S., Hu, X., ... Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, *7*(2), 221–231.

van Noordt, C., & Misuraca, G. (2022). Artificial intelligence for the public sector: Results of landscaping the use of AI in government across the European Union. *Government Information Quarterly*, *39*(3), 101714. doi: 10.1016/j.giq.2022.101714

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. doi: 10.2307/30036540