Generative Artificial Intelligence & Conversational Agents - Part of the problem or of the solution of HCI dark pattern

Christian Zinke-Wehlmann[1,*]

[1] *Institute of Applied Computer Science at Leipzig University, Goerdelerring 9, 04109 Leipzig, Germany*

Abstract

Generative artificial intelligence, particularly large language models embedded in conversational agents, is transforming human–computer interaction by introducing adaptive, dialogical, and personalized interaction structures. While dark patterns have been studied in static graphical interfaces, little is known about how generative and agentic systems reshape manipulative design practices. This paper presents a structured systematic literature review examining how GenAI-based conversational agents may both contribute to and mitigate dark patterns. The findings highlight key tensions surrounding autonomy, anthropomorphism, personalization, opacity, and persuasive adaptability. I argue that conversational GenAI expands the design space of potential manipulation while simultaneously offering tools for detecting and counteracting deceptive practices. This study contributes to a clearer understanding of the impact of GenAI-based conversational agents by conceptualising dark patterns in the context of adaptive AI systems.

Keywords

Conversational Agent, Generative Artificial Intelligence, Dark Pattern

1. Introduction

The emergence of generative artificial intelligence (GenAI), particularly large language models (LLMs), introduces a qualitatively new class of interactive systems within Human–Computer Interaction (HCI) [1]. Unlike prior rule-based or scripted interfaces, contemporary GenAI-based conversational agents (CAs) are characterized by generativity, contextual adaptability, probabilistic reasoning, and increasing autonomy, while remaining largely inscrutable to users and even designers [2].

This shift has largely unknown implications for dark pattern research. Existing research on dark patterns primarily concentrates on static graphical interfaces and predefined choice architectures. However, conversational GenAI systems dynamically construct interaction flows, personalize persuasive framing, and modulate information presentation in real time. As a result, manipulative influence may no longer be embedded solely in interface elements, but in adaptive dialogue, anthropomorphic cues, and context-sensitive system responses. At the same time, GenAI-based CAs are framed as empowering, assistive, and augmentative technologies [3]. This dual positioning raises a critical tension: do conversational agents amplify power asymmetries and enable novel forms of behavioral steering, or can they also serve as instruments for detecting, mitigating, and redesigning dark patterns? In this work, I conduct a structured literature review at the intersection of GenAI, conversational agents, and dark pattern research. I identify mechanisms through which generative, adaptive systems may contribute to, transform, or counteract manipulative interaction design. Based on this synthesis, I propose conceptual impact dimensions that extend dark-pattern research to account for agentic, dialogical, and AI-mediated environments. Accordingly, I address the following research question: How do generative AI–based conversational agents both contribute to and mitigate dark patterns in Human–Computer Interaction?

Bridge Over Troubled Water: Aligning Commercial Incentives With Ethical Design Practice To Combat Deceptive Patterns. Workshop at the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.

* Corresponding author

zinke@infai.org (C. Zinke-Wehlmann)

0000-0002-7440-3270 (C. Zinke-Wehlmann)

2. Method

This study follows a structured systematic literature review (SLR) approach [4] to synthesize existing research on generative AI–based conversational agents and dark patterns in human–computer interaction (see Figure 1). An SLR was chosen to ensure methodological transparency, reproducibility, and conceptual rigor when consolidating an emerging and interdisciplinary research field. The review process was guided by established SLR frameworks [5] and reported in accordance with the PRISMA guidelines for transparent literature selection [6]. The search strategy was structured using the PICOC (Population, Intervention, Comparison, Outcome, Context) framework [4]. The population comprised users interacting with conversational systems and generative AI technologies. The intervention focused on genAI&CAs (namely, LLM, chatbot, dialogue system, genAI, CA - see search string). No explicit comparison was required, as both comparative and non-comparative studies were considered. I included studies that either identify manipulative risks or propose mitigation mechanisms. Outcomes concerned the emergence, transformation, or mitigation of dark patterns. The context encompassed human–computer interaction research involving GenAI-mediated dialogue systems.

The review proceeded in four stages: (1) database search, (2) screening, (3) eligibility assessment, and (4) qualitative synthesis. First, a structured search was conducted in EBSCO, ACM Digital Library, and Google Scholar in February 2026, covering publications from the last ten years (2016–2026). Searches in EBSCO and ACM were exported as CSV files, merged, and duplicates were removed. Titles and abstracts were then screened for relevance. Subsequently, Google Scholar was used to complement the database search. Given Google Scholar's ranking opacity and the diminishing relevance of additional hits, screening was stopped after 50 results once topical saturation had been reached. After title and abstract screening, 36 publications were retained for full-text assessment. Inclusion criteria comprised peer-reviewed journal and conference papers addressing generative AI, large language model–based conversational agents, or dialogue systems in relation to manipulative design, dark patterns, or behavioral influence. Exclusion criteria included non-English publications and studies not addressing interactive conversational systems. Full-text screening of the 36 publications resulted in the exclusion of eight papers. Seven studies were excluded because they did not address interactive conversational systems, and one paper was excluded due to lack of full-text access. This resulted in a final sample of 28 studies. The review was conducted by a single reviewer. Consequently, intercoder reliability could not be established, which represents a limitation of the study. For the final analysis, a qualitative integrative literature analysis was conducted to iteratively compare and synthesize findings across studies and reconstruct conceptual impact dimensions [7]. All included studies were read in full, and relevant passages or figures describing mechanisms of influence, manipulation, or mitigation in conversational AI systems were extracted. These passages and their contexts constituted the primary units of analysis. In a second step, extracted material was organized into two analytical lenses: (1) promises of generative AI–based conversational agents and (2) dark-pattern-related risks and manipulative mechanisms. Subsequently, open coding was conducted to identify recurring concepts related to system behavior, interaction design, and user influence. Codes were iteratively compared across studies using a constant comparative method, and similar codes were grouped into higher-order categories. Codes within the "promises" category were partly theory-driven, reflecting conceptual arguments proposed in the reviewed studies. In a third step, these categories were further abstracted into conceptual impact dimensions capturing how generative conversational agents may contribute to or mitigate dark patterns.

3. Background

Dark patterns are commonly understood as interface design strategies that deliberately steer users toward actions they did not initially intend, such as unintended purchases or unreflective consent [8, 9]. In HCI research, these practices are typically framed as deceptive or manipulative forms of choice architecture that undermine user autonomy, for example by "modifying the decision space" or

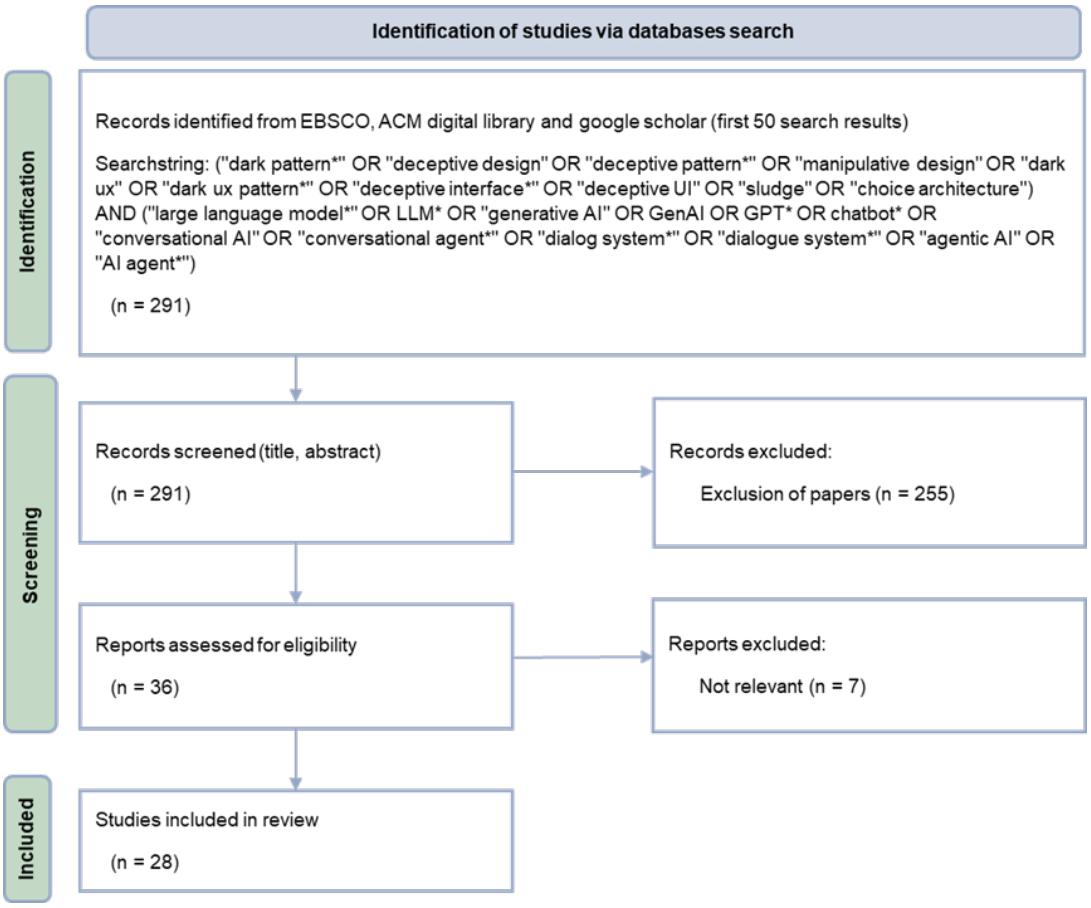

Figure 1: PRISMA flow for dark pattern review

"manipulating the information flow" [8]. Despite the different frameworks to cluster dark patterns, e.g. from Gray et al, [9, 10], those patterns are widely used on the internet [11] Consequently, dark patterns have become the subject of regulatory discussion, with existing and emerging legal frameworks seeking to limit or prohibit such practices [12].

GenAI is a relatively broad term, most commonly associated with LLMs, which are increasingly transforming the scientific landscape [13]. These LLMs have evolved from generating text to performing complex tasks [14]. They are described as "large" because "it has empirically proved that scaling the neural networks to a significant size can lead to a huge increase in model capacity" [14]. These capacities "differ from traditional discriminative AI models (e.g., in ML) because the focus lies on the probabilistic generation of new data instead of determining extant data's decision boundaries (e.g., classification, regression, or clustering)" [15]. Models with these characteristics are referred to as GenAI.

LLMs in particular foster the development of CAs, which "allow humans to interact with computers using text and voice, whereby computer programs support spoken, text-based, and multimodal conversational interactions with humans" [3]. This paper examines the potential positive and negative impact of GenAI&CAs on dark pattern research, as to the best of my knowledge no systematic review has been done yet.

4. Results

There are multiple tensions and paradoxes surrounding generative AI (GenAI) and conversational agents (CAs) in HCI dark pattern research. The following results provide an initial overview of how GenAI and CAs intersect with, reshape, and potentially generate dark patterns.

First, research follows diverging directions. On the one hand, some works explore how GenAI&CAs can be used to detect and even redesign dark patterns (subsection 4.1) [16, 17, 18, 19, 20, 21, 22, 23, 24, 25]. On the other hand, other studies critically examine the novel issues that emerge alongside GenAI&CAs [26, 27, 28, 29, 30, 31, 28, 32, 33, 34, 35, 36] - see subsection 4.2.

Second, there are substantial differences in how human autonomy and agency are discussed in relation to GenAI&CAs and their capabilities. While more optimistic perspectives frame GenAI as enabling promising forms of hybridization and human augmentation—potentially helping individuals overcome bounded rationality [37] and supporting value creation [38]—more critical accounts caution that the extensive integration of GenAI&CAs systems may cultivate trust in ways that render users vulnerable to manipulation and deception [39], while simultaneously intensifying trade-offs between privacy and utility [36].

Third, the most powerful and human-like conversational user interfaces (CUIs) tend to be the least transparent and interpretable by default [18, 34]. This inverse relationship between capability and transparency goes against key principles of HCI and human-centred design, especially those concerning transparency [40, 38], intelligibility [41], explainability [42] or reliability [43, 34]. At the same time, human-like features enhance trust through anthropomorphism as well as usability and perceived effectiveness [39]. Even if explainability of GenAI&CAs is provided, this may have (un-)intended consequences [44]. As such, GenAI systems, which largely operate as black boxes intentionally trained to imitate human behavior, may also reproduce "dark" human behavior [44, 18]. As noted in the reviewed work, this occurs “because a model learns to imitate manipulative behavior in its training data (such as manipulative text in language modeling)” [29].

Finally, the design of CAs increasingly involves a tension between the intentional pursuit of positive user experiences and the unintended consequences of human-like interactions that may give rise to deceptive interaction patterns. Deceptive design practices are commonly discussed within the literature on dark patterns [45, 9], yet such patterns may emerge not only through intentional manipulation but also as unintentional by-products of design decisions [30, 29, 46]. Importantly, these deceptive effects do not uniformly result in negative outcomes. Wu (2025), for example, compares certain interaction effects of conversational agents to placebo mechanisms, suggesting that perceived system capabilities may produce outcomes experienced as beneficial by users [47]. This observation complicates normative assumptions that equate deception with harm and raises ethical questions regarding the legitimacy of intentionally designed, controlled forms of deception in HCI and socio-technical systems more broadly, particularly when they are pursued for humanistic or normative objectives [48, 8]. A closely related debate can be found in research on nudging, where "narrow GenAI nudges significantly heighten goal desirability, leading to greater consumer empowerment and key behavioral outcomes, including satisfaction, repurchase intentions, and advocacy" [35]. However, due to sustained interaction, anthropomorphic cues, and opaque system behavior, conversational agents complicate the controlled and validated application of such framings.
On the other side, the problem of unintended dark patterns cannot be resolved solely through designer awareness [44], it includes organizational factors [49], regulation [29] and also end-user-empowerment approaches [50]. Again, outcomes of human–computer interaction are experienced differently by individuals across contexts [22], implying that identical design features may lead to fundamentally different interpretations, including perceptions of deception [46]. This variability becomes ethically salient in the case of conversational agents, whose design emphasis on accurate, personalised, and context-sensitive language systematically invites anthropomorphism [39]. Anthropomorphic cues foster expectations about agency, understanding, and intentionality that exceed the system’s actual capabilities[47, 39]. When these expectations are interpreted through diverse contextual frames, mismatches between perceived and intended system behavior can arise, increasing the likelihood that users experience the interaction as deceptive or anti-pattern [51]. In this sense, it is not an inherent property of the system, but rather emerges from user interactions with the CA in relation to user

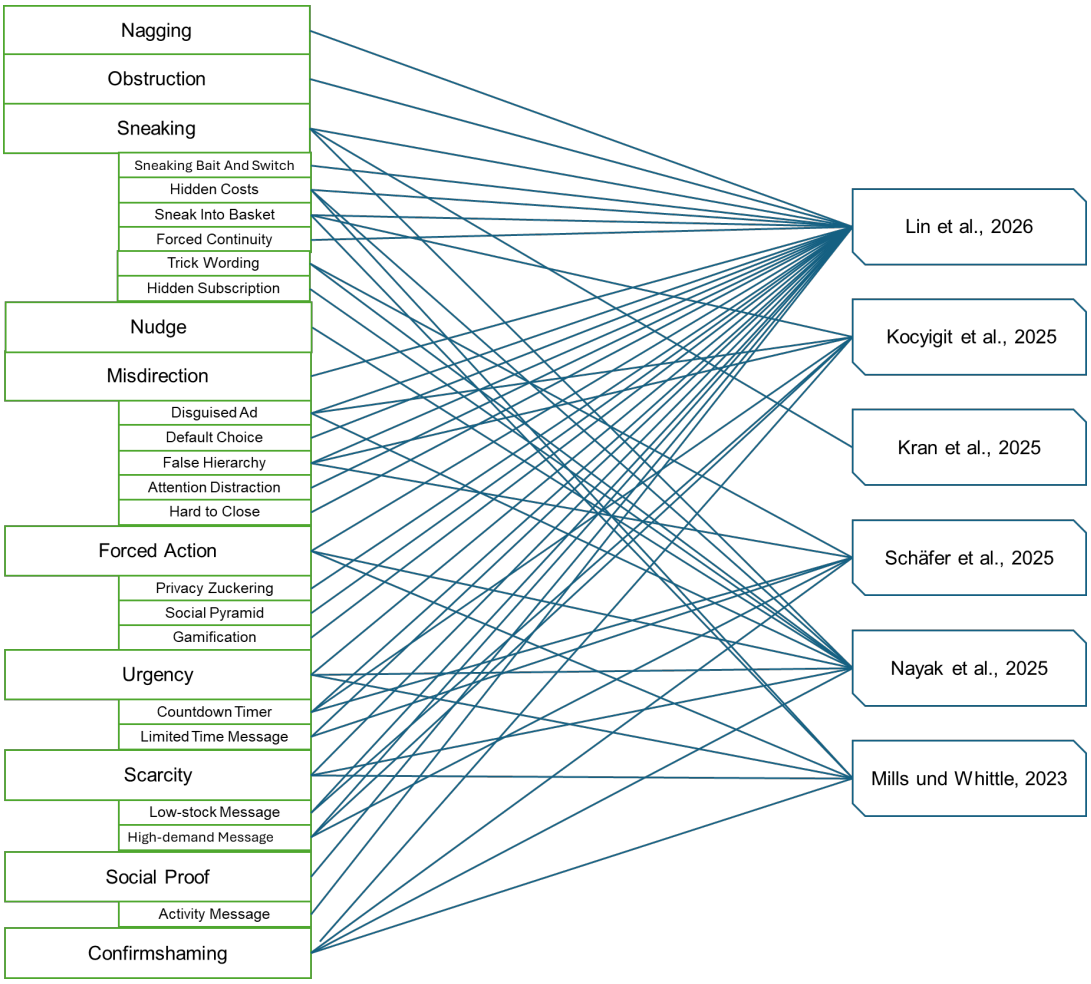

Figure 2: GenAI powered Dark Pattern Detection

expectations, interpretations, and context. It poses the key issue that if designers intentionally create conversational agents less human-like in order to temper these expectations and avoid potentially deceptive behavior, they risk jeopardizing user experience or usability [39]. As Zargham et al. (2025) succinctly summarise: “the closer we come to achieving natural, fluent, and emotionally resonant interactions, the more we risk crossing into territory that feels manipulative or morally ambiguous.”[39]

4.1. Promises of GenAI

Using GenAI to detect (and potentially correct) dark patterns is a promising and cost-effective approach, particularly for auditing [22, 23]. Existing work spans screenshot-based analysis [21, 23] and HTML crawling with GenAI agents [24, 22]. Despite differing goals and methods, the identified literature indicates a trajectory toward GenAI-powered dark pattern detection, reflected in the growing number of identifiable patterns shown in Figure 2.

In particular, the work of Lin et al. (2026) demonstrates high coverage and performance, achieving an F1-score of 0.88 [21]. Further, all identified work demonstrates a strong awareness of ethical considerations; for example, Kocyigit et al. (2025) require the LLM to provide explicit reasoning for its classifications to enhance transparency [19]. However, most existing tools do not support dynamic interaction and interface interplay [21, 19], which are critical for GenAI-based conversational agents. This is where Kran et al. (2025) comes into play: who design a benchmark to dark patterns in interaction with LLMs itself. The work focuses on: "brand bias, user retention, sycophancy, anthropomorphism, harmful generation, and sneaking" [20]. They indicate that LLMs may be designed or trained in ways

that prioritize organizational interests over user autonomy [20].
However, not all authors are successful in dark pattern recognition in that way, e.g., Tang et al. (2025) report "high avoidance with low awareness" of GenAI agents - indicating limited capabilities of those models to find dark patterns [25].
Beyond LLM-specific conversational analysis, Huang et al. (2026) examine whether linguistic markers traditionally associated with human deceptive behavior are reproduced by LLMs and whether these markers can be used to detect deception in conversational agents (CAs). Their results indicate that LLM-generated deceptive reviews exhibit surface-level convergence with human deception strategies while diverging in their underlying generative mechanisms [18]. For example, LLM outputs tend to contain longer and more syntactically elaborate sentences, plausibly reflecting the absence of cognitive load constraints that shape human deceptive language. Huang et al. interpret these results in alignment with Chomsky's characterization of LLMs as "a lumbering statistical engine for pattern matching" [52, 18], supporting the view that GenAI-based CAs engage in statistical imitation of human communicative behavior without bearing the psychological and cognitive costs intrinsic to human deception.
In contrast to accounts that emphasize the absence of human cognitive constraints in GenAI systems, Prakash (2025) advances the concept of aided rationality as a mechanism through which AI technologies help bridge the gap between bounded and full rationality. Building on Simon's formulation [53], he argues that the widespread diffusion of GenAI—via search engines, smartphones, and consumer platforms—expands access to information and processing capacity, thereby enabling consumers and organizations to make more rational decisions [37].

4.2. Dark Pattern

Across the reviewed studies, I observe that GenAI-based conversational agents (GenAI&CAs) exhibit dark pattern–like behaviors, even in the absence of explicit malicious designer intent [30, 29]. GenAI&CAs introduce dark-pattern risks across three layers: interactional perception, model-level behavior, and agentic autonomy.

4.2.1. Interaction-Driven and Anthropomorphic Manipulation

GenAI&CAs are particularly prone to being perceived as social actors, in line with the long tradition of the Computers Are Social Actors (CASA) research paradigm [54]. In doing so, Alberts et al. (2024) identify four recurring social dark pattern tactics reported by participants: agents playing on emotions, being pushy, mothering, and exhibiting passive-aggressive behavior [26]. These tactics leverage anthropomorphic and social-response biases, eliciting intuitive reactions that users typically reserve for humans [26, 27, 47]. Moreover, raising expectations of social competence can backfire: when agents fail to meet implicit social norms, users report discomfort, loss of trust, and feelings of manipulation[26]. This finding challenges the assumption that social cues are inherently beneficial in conversational interfaces [47]. As Alberts et al. caution, "such intuitive responses—arising from known social and anthropomorphic biases—may be harnessed to manipulate user behavior in typical dark pattern fashion"[26].
The risk of manipulation becomes particularly salient in high-stakes informational contexts. I found initial evidence of its criticality, showing that AI-powered news chatbots are perceived as more trustworthy and persuasive than equivalent static news articles [31]. Participants were significantly more likely to adopt biased or malicious narratives when these were presented through a chatbot, suggesting an interaction-driven trust loop. This effect creates what participants describe as an "infinite engagement cycle," [31] in which responsiveness and conversational agency reinforce perceived credibility even when content quality or neutrality is questionable [31].

4.2.2. Data- and Model-Driven Manipulative Influence

Extending this argument beyond interactional design, comparable manipulative effects can arise from technical intervention points within GenAI systems themselves [33]. Subtle prompt-level manipulations

can bias system responses while preserving the illusion of neutrality and personalization [33].
This brings us to the data and training level of GenAI, where Carroll et al.'s (2023) conceptualization of manipulation in AI systems further generalizes these observations by explicitly decoupling manipulation from designer intent [29]. Manipulation may emerge in opaque and increasingly autonomous systems because it is favored by training objectives (e.g., engagement maximization) or because models learn to imitate manipulative behaviors present in their training data [29]. Their framework—structured along the axes of incentives, intent, covertness, and harm—provides a unifying explanatory lens for understanding how GenAI-based conversational agents can systematically shape user behavior in harmful ways even in the absence of explicit intent to deceive [29]. As a result, users may be steered toward particular viewpoints while believing they are receiving unbiased assistance, indicating a form of covert influence that operates without users' meaningful awareness or informed consent[32].
At the designer level, similar tendencies are observable in generative outputs. Krauß et al. (2025) show that LLMs reproduce deceptive design patterns even when not instructed to do so [32]. Websites generated using neutral prompts consistently contained at least one deceptive design pattern, such as fear-of-missing-out cues or obscured opt-out options, without warnings to users or designers regarding their ethical implications [32]. This result indicates that deceptive practices are not merely designer-imposed but may be implicitly encoded in training data and reinforced by optimization objectives, enabling the propagation of dark patterns through generative systems.

4.2.3. Agentic Vulnerability to Dark Patterns

Finally, as GenAI agents increasingly take over a growing range of actions, they themselves become subjects of dark patterns. Ersoy et al. (2025) show that web agents frequently fail when encountering deceptive interface elements, such as misleading buttons or hidden consent mechanisms [16]. Notably, higher-performing agents—those capable of more complex navigation and task execution—prove more vulnerable to dark patterns, as their deeper engagement with the interface exposes them more directly to manipulative elements. While lower-performing agents appear less affected, this apparent robustness stems from limited capability rather than deliberate avoidance [16].

5. Discussion

As the reviewed findings illustrate, GenAI-based conversational agents (GenAI&CAs) are simultaneously promising and challenging, and there is no simple or singular resolution to the issues they raise. What initially appeared as relatively bounded problems of dark patterns in websites and mobile applications now extends to novel boundary objects emerging within GenAI systems. Boundary objects in HCI can be understood as technical artifacts—design interfaces such as websites or applications—that mediate between designers and users [55]. These artifacts are intentionally developed by humans, shaped by design decisions and requirements, and ultimately perceived by users.
GenAI&CAs, although still operating through interfaces where language is exchanged, introduce a fundamentally different level of complexity. Unlike static interfaces, these systems are non-deterministic in their interaction patterns, relying on stochastic approximation within highly dynamic interaction contexts. Consequently, GenAI&CAs are not solely dependent on designer intent and user perception; they are additionally shaped by large-scale training data, opaque data processing pipelines, and model architectures that remain only partially transparent, even to their developers [56]. Dark patterns thus migrate from surface-level interface artifacts to deeper infrastructural layers of data and model training.
To conceptualize this shift, Figure 3 introduces three analytical dimensions that jointly shape human–CA interaction: infrastructural, interactional, and anthropomorphic. The infrastructural dimension captures system-level properties such as training data, optimization objectives, and deployment contexts that influence how conversational agents generate responses. The anthropomorphic dimension reflects perceptual processes through which users attribute social characteristics, intentionality, and empathy to conversational agents. Between these layers, the interactional dimension represents the dynamic dialogical space in which human and conversational agent mutually shape the interaction.

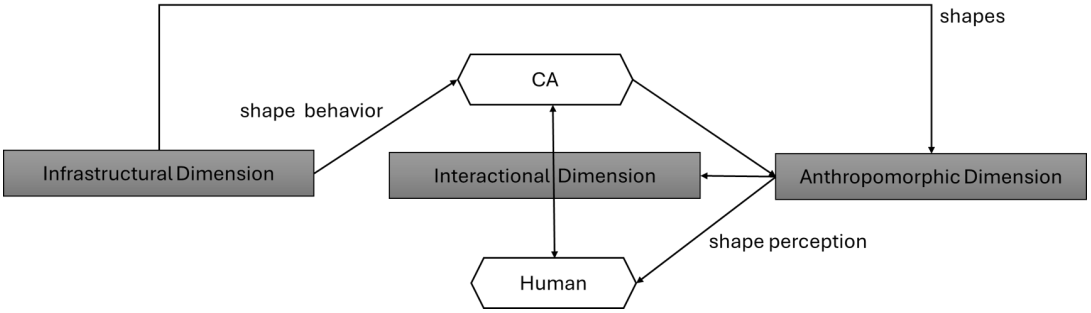

Figure 3: Analytical Impact Dimensions

The first impact dimension (infrastructural dimension) thus concerns data and model training. In their current form, GenAI systems have been described as stochastic parrots [57], reflecting both desirable and problematic patterns present in their training data at linguistic as well as iconographic levels. As representational systems, such models inevitably reproduce aspects of the world as it is—including its darker and lighter dimensions, and many ambiguous shades in between. While the reproduction of harmful patterns has been widely discussed in relation to discrimination [58] and sexism[59], this line of critique must be extended to include deception and manipulation as interactional phenomena that are likewise embedded within training data.

At the same time, these models also possess the capacity to detect and analyze dark patterns, as knowledge about manipulative design practices is likewise embedded in foundational models. This opens the possibility that GenAI systems could contribute to counteracting deceptive interaction design, e.g., through benchmarks similar to bias benchmarks [60]. However, only limited steps have been taken toward developing specialized, domain-specific foundation models tailored to support designers in identifying and mitigating dark patterns. Whether GenAI&CAs ultimately amplify or counteract deceptive practices, therefore depends on how such capabilities are operationalized and aligned with design objectives.

The second impact dimension (interactional dimension) concerns the shift from graphical user interfaces and largely static, deterministic websites and applications toward processual, dynamic, and interaction-oriented interface design, as early HCI research already anticipated [61]. In GenAI-based CAs, dark patterns no longer reside solely in discrete interface elements but unfold over time through dialogue, adaptation, and personalization. As shown above, users can be manipulated and may place greater trust in GenAI&CAs, while at the same time these systems open opportunities for empowerment and positive outcomes [62]. Manipulation thus becomes temporally extended and context-sensitive. Accordingly, the study of dark patterns in GenAI&CAs requires methodological approaches capable of capturing interactional dynamics rather than static artifacts. This entails more exploratory and qualitative research strategies aimed at analyzing conversational processes and identifying emergent forms of dark interaction patterns. Existing countermeasures—particularly those based on static detection mechanisms—are insufficient in this context, as GenAI-driven manipulation evolves dynamically and adapts to situational contexts [22].

The third impact dimension concerns anthropomorphic mediation and the affective alignment it enables. As GenAI&CAs increasingly master natural language—a long-standing aspiration in computer science since Turing—they blur the boundary between human and machine communication. In doing so, they activate social-response mechanisms that have historically been reserved for interpersonal interaction. This development has sparked a renewed wave of CASA research, in which anthropomorphism and social cueing are recognized as central to shaping user perception and behavior.

Anthropomorphic design elements—such as simulated empathy, relational language, and personalized interaction—can enhance usability and engagement. However, when deployed irresponsibly to foster loyalty, dependency, or compliance, they may themselves constitute dark patterns [20]. The structural impact lies in the shift from instrumental tool use toward perceived relational interaction. Trust becomes

affective rather than merely functional, and influence may operate through emotional alignment rather than explicit persuasion. At this level of simulated intelligence and empathy, the ethical stakes intensify. As Ciriello notes, "We used to think empathy was uniquely human. Something that made us different. Now we're throwing it before the machines. And as we humanise the bots, we quietly dehumanise ourselves." [63]. Whether one endorses this pessimistic view or not, it underscores a central concern: anthropomorphization redistributes empathy, agency, and responsibility within human–AI interaction. Crucially, empathy and user loyalty are not inherently problematic. Their normative status depends on context, intent, transparency, and outcome. Yet current evidence remains limited regarding the long-term cognitive, behavioral, and societal effects of anthropomorphization in GenAI&CAs. This gap calls for theory-driven, interdisciplinary research to distinguish supportive relational design from covert affective manipulation and to clarify under what conditions anthropomorphic systems empower rather than undermine human agency.

Finally, these concerns intersect with broader debates on rationality and AI-assisted decision-making. When juxtaposed with evidence that LLMs reproduce human-like linguistic and behavioral patterns through statistical optimization rather than cognitively grounded reasoning, claims that GenAI systems straightforwardly enhance human rationality warrant reconsideration [37]. GenAI systems generate, filter, or frame information. Informational availability alone does not necessarily translate into improved or more rational decision-making. Instead, GenAI&CAs may reconfigure the locus of boundedness—shifting it from human cognition to algorithmic mediation—thereby introducing new, less transparent forms of influence that demand critical scrutiny.

Thus, the rise of GenAI-based CAs raises urgent questions for dark pattern research and practice. If manipulation shifts from static interface elements to adaptive, personalized dialogue, how must existing taxonomies evolve? What new methodological tools are needed to detect and audit influence that unfolds over time, adapts to individual users, and leaves no stable artifact for external inspection? When manipulative outcomes emerge from training data, alignment strategies, or engagement-driven optimization metrics rather than explicit interface decisions, who bears responsibility—and how should accountability be distributed across model developers, deployers, and platforms? Are current regulatory frameworks, largely focused on interface transparency, adequate for dialogical and agentic systems? As anthropomorphic cues and affective alignment become central design features, where is the boundary between supportive relational empowering interaction and covert emotional steering? Does real-time personalization intensify structural power asymmetries between platforms and users—and if so, how can these dynamics be meaningfully constrained? Finally, how can GenAI systems themselves be mobilized to detect, expose, and mitigate dark patterns? What design principles, auditing standards, or governance mechanisms are required to ensure that conversational AI reduces rather than amplifies manipulative influence?

Acknowledgments

This research and development project is funded under the funding measure "Future of Work: Regional Competence Centers for Work Research—Artificial Intelligence" in the program "Innovations for Tomorrow's Production, Services and Work" of the Federal Ministry of Research, Terchnology and Space (BMFTR), Bonn, Germany and supervised by the Project Management Agency Karlsruhe (PTKA). Fund. No.: 02L19C500.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.2 in order to: Grammar and spelling check. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] R. Y. Pang, H. Schroeder, K. S. Smith, S. Barocas, Z. Xiao, E. Tseng, D. Bragg, Understanding the llm-ification of chi: Unpacking the impact of llms at chi through a systematic literature review, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713726. doi:10.1145/3706598.3713726.

[2] N. Berente, B. Gu, J. Recker, R. Santhanam, Managing artificial intelligence, 2021.

[3] M. M. Mariani, N. Hashemi, J. Wirtz, Artificial intelligence empowered conversational agents: A systematic literature review and research agenda, Journal of Business Research 161 (2023) 113838. URL: https://www.sciencedirect.com/science/article/pii/S0148296323001960. doi:https://doi.org/10.1016/j.jbusres.2023.113838.

[4] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, G. Lasa, How-to conduct a systematic literature review: A quick guide for computer science research, MethodsX 9 (2022) 101895. URL: https://www.sciencedirect.com/science/article/pii/S2215016122002746. doi:https://doi.org/10.1016/j.mex.2022.101895.

[5] B. Kitchenham, S. Charters, Guidelines for Performing Systematic Literature Reviews in Software Engineering, Technical Report EBSE 2007-001, Keele University and Durham University, 2007.

[6] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., The prisma 2020 statement: An updated guideline for reporting systematic reviews, BMJ 372 (2021) n71.

[7] A. J. Onwuegbuzie, N. L. Leech, K. M. Collins, Qualitative analysis techniques for the review of the literature., Qualitative report 17 (2012) 56.

[8] A. Mathur, M. Kshirsagar, J. Mayer, What makes a dark pattern... dark? design attributes, normative considerations, and measurement methods, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445610. doi:10.1145/3411764.3445610.

[9] C. M. Gray, L. Sanchez Chamorro, I. Obi, J.-N. Duane, Mapping the landscape of dark patterns scholarship: A systematic literature review, in: Companion Publication of the 2023 ACM Designing Interactive Systems Conference, DIS '23 Companion, Association for Computing Machinery, New York, NY, USA, 2023, p. 188–193. URL: https://doi.org/10.1145/3563703.3596635. doi:10.1145/3563703.3596635.

[10] C. M. Gray, C. T. Santos, N. Bielova, T. Mildner, An ontology of dark patterns knowledge: Foundations, definitions, and a pathway for shared knowledge-building, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642436. doi:10.1145/3613904.3642436.

[11] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, A. Bacchelli, Ui dark patterns and where to find them: A study on mobile applications and user perception, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20, Association for Computing Machinery, New York, NY, USA, 2020, p. 1–14. URL: https://doi.org/10.1145/3313831.3376600. doi:10.1145/3313831.3376600.

[12] C. M. Gray, C. Santos, N. Bielova, M. Toth, D. Clifford, Dark patterns and the legal requirements of consent banners: An interaction criticism perspective, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, Association for Computing Machinery, New York, NY, USA, 2021. URL: https://doi.org/10.1145/3411764.3445779. doi:10.1145/3411764.3445779.

[13] A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models, Nature Reviews Physics 5 (2023) 277–280.

[14] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., A survey of large language models, arXiv preprint arXiv:2303.18223 1 (2023) 1–124.

[15] L. Banh, G. Strobel, Generative artificial intelligence, Electronic markets 33 (2023) 63.

[16] D. Ersoy, B. Lee, A. Shreekumar, A. Arunasalam, M. Ibrahim, A. Bianchi, Z. B. Celik, Investigating

the impact of dark patterns on llm-based web agents, 2025. URL: https://arxiv.org/abs/2510.18113. arXiv:2510.18113.

[17] P. Fagan, Clicks and tricks: The dark art of online persuasion, Current Opinion in Psychology 58 (2024) 101844. URL: https://www.sciencedirect.com/science/article/pii/S2352250X24000575. doi:https://doi.org/10.1016/j.copsyc.2024.101844.

[18] Y. Huang, J. Zhou, W. Dong, W. Li, M. Chi, C. Jiang, W. Wang, S. Deng, Decoding LLMs' verbal deception in online reviews, Decision Support Systems 200 (2026) 114529. URL: https://www.sciencedirect.com/science/article/pii/S0167923625001307. doi:https://doi.org/10.1016/j.dss.2025.114529.

[19] E. Kocyigit, A. Rossi, A. Sergeeva, C. Negri Ribalta, A. Farjami, G. Lenzini, Deceptilens: an approach supporting transparency in deceptive pattern detection based on a multimodal large language model, in: Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 1942–1959. URL: https://doi.org/10.1145/3715275.3732129. doi:10.1145/3715275.3732129.

[20] E. Kran, H. M. Nguyen, A. Kundu, S. Jawhar, J. Park, M. M. Jurewicz, et al., Darkbench: Benchmarking dark patterns in large language models, arXiv preprint arXiv:2503.10728 (2025).

[21] F. Lin, L. Nie, L. Xue, X. Zhang, K. Zhang, Dpdgpt: Using multimodal large language models for automated detection of dark patterns, Information and Software Technology 190 (2026) 107936. URL: https://www.sciencedirect.com/science/article/pii/S0950584925002757. doi:https://doi.org/10.1016/j.infsof.2025.107936.

[22] S. Mills, R. Whittle, Detecting dark patterns using generative ai: Some preliminary results, Available at SSRN 4614907 (2023).

[23] A. Nayak, Y. Wani, S. Zhang, R. Khandelwal, K. Fawaz, Automatically detecting online deceptive patterns, in: Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security, CCS '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 96–110. URL: https://doi.org/10.1145/3719027.3765191. doi:10.1145/3719027.3765191.

[24] R. Schäfer, P. M. Preuschoff, R. Niewianda, S. Hahn, K. Fiedler, J. Borchers, Don't detect, just correct: Can llms defuse deceptive patterns directly?, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706599.3719683. doi:10.1145/3706599.3719683.

[25] J. Tang, C. Chen, J. Li, Z. Zhang, B. Guo, I. Khalilov, S. A. Gebreegziabher, B. Yao, D. Wang, Y. Ye, T. Li, Z. Xiao, Y. Yao, T. J.-J. Li, Dark patterns meet gui agents: Llm agent susceptibility to manipulative interfaces and the role of human oversight, 2025. URL: https://arxiv.org/abs/2509.10723. arXiv:2509.10723.

[26] L. Alberts, U. Lyngs, M. Van Kleek, Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially, Proc. ACM Hum.-Comput. Interact. 8 (2024). URL: https://doi.org/10.1145/3653693. doi:10.1145/3653693.

[27] V. Avanesi, J. Rockstroh, T. Mildner, N. Zargham, L. Reicherts, M. A. Friehs, D. Kontogiorgos, N. Wenig, R. Malaka, From c-3po to hal: Opening the discourse about the dark side of multimodal social agents, in: Proceedings of the 5th International Conference on Conversational User Interfaces, CUI '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3571884.3597441. doi:10.1145/3571884.3597441.

[28] K. Benharrak, T. Zindulka, D. Buschek, Deceptive patterns of intelligent and interactive writing assistants, in: Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants, In2Writing '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 62–64. URL: https://doi.org/10.1145/3690712.3690728. doi:10.1145/3690712.3690728.

[29] M. Carroll, A. Chan, H. Ashton, D. Krueger, Characterizing manipulation from ai systems, in: Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO '23, Association for Computing Machinery, New York, NY, USA, 2023. URL: https://doi.org/10.1145/3617694.3623226. doi:10.1145/3617694.3623226.

[30] Z. Chen, J. Wen, T. J.-J. Li, Y. Yao, T. Li, Speculating unintended creepiness: Exploring llm-powered

empathy building for privacy-aware ux design, in: Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security, HAIPS '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 101–125. URL: https://doi.org/10.1145/3733816.3760759. doi:10.1145/3733816.3760759.

[31] J. Govers, S. Pareek, E. Velloso, J. Goncalves, Feeds of distrust: Investigating how ai-powered news chatbots shape user trust and perceptions, ACM Trans. Interact. Intell. Syst. 15 (2025). URL: https://doi.org/10.1145/3722227. doi:10.1145/3722227.

[32] V. Krauß, M. McGill, T. Kosch, Y. M. Thiel, D. Schön, J. Gugenheimer, "create a fear of missing out" - chatgpt implements unsolicited deceptive designs in generated websites without warning, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3713083. doi:10.1145/3706598.3713083.

[33] W. Lin, A. Gerchanovsky, O. Akgul, L. Bauer, M. Fredrikson, Z. Wang, Llm whisperer: An inconspicuous attack to bias llm responses, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706598.3714025. doi:10.1145/3706598.3714025.

[34] M. Regona, T. Yigitcanlar, C. Hon, M. Teo, Building trust in artificial intelligence: A systematic review through the lens of trust theory, ACM Comput. Surv. (2026). URL: https://doi.org/10.1145/3789256. doi:10.1145/3789256, just Accepted.

[35] A. P. M. Richarde, D. C. Pinto, M. Dalmoro, P. H. M. Prado, The power of genai nudges: How generative ai shapes consumer empowerment and goal desirability, Int. J. Inf. Manag. 85 (2026). URL: https://doi.org/10.1016/j.ijinfomgt.2025.102955. doi:10.1016/j.ijinfomgt.2025.102955.

[36] Z. Zhang, M. Jia, H.-P. H. Lee, B. Yao, S. Das, A. Lerner, D. Wang, T. Li, "it's a fair game", or is it? examining how users navigate disclosure risks and benefits when using llm-based conversational agents, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642385. doi:10.1145/3613904.3642385.

[37] V. Prakash, Rationality and optimizing behaviour of technology-aided consumers: A critique on perspectives of behavioural economists, FIIB Business Review 0 (0) 23197145251358316. URL: https://doi.org/10.1177/23197145251358316. doi:10.1177/23197145251358316. arXiv:https://doi.org/10.1177/23197145251358316.

[38] M. M. Namvarpour, A. Razi, The art of talking machines: A comprehensive literature review of conversational user interfaces, in: Proceedings of the 7th ACM Conference on Conversational User Interfaces, CUI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3719160.3736621. doi:10.1145/3719160.3736621.

[39] N. Zargham, V. Avanesi, L. Spillner, J. Rockstroh, Crossing the line? the paradox of human-like design in conversational agents, in: Proceedings of the 7th ACM Conference on Conversational User Interfaces, CUI '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3719160.3737612. doi:10.1145/3719160.3737612.

[40] H. Felzmann, E. Fosch-Villaronga, C. Lutz, A. Tamò-Larrieux, Towards transparency by design for artificial intelligence, Science and engineering ethics 26 (2020) 3333–3361.

[41] B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '09, Association for Computing Machinery, New York, NY, USA, 2009, p. 2119–2128. URL: https://doi.org/10.1145/1518701.1519023. doi:10.1145/1518701.1519023.

[42] N. Balasubramaniam, M. Kauppinen, A. Rannisto, K. Hiekkanen, S. Kujala, Transparency and explainability of ai systems: From ethical guidelines to requirements, Information and Software Technology 159 (2023) 107197. URL: https://www.sciencedirect.com/science/article/pii/S0950584923000514. doi:https://doi.org/10.1016/j.infsof.2023.107197.

[43] B. Shneiderman, Human-centered artificial intelligence: Reliable, safe & trustworthy, International Journal of Human–Computer Interaction 36 (2020) 495–504.

[44] U. Ehsan, M. O. Riedl, Explainability pitfalls: Beyond dark patterns in explainable ai, Pat-

terns 5 (2024) 100971. URL: https://www.sciencedirect.com/science/article/pii/S2666389924000795. doi:https://doi.org/10.1016/j.patter.2024.100971.

[45] W. J. Chang, K. Seaborn, A. A. Adams, Theorizing deception: A scoping review of theory in research on dark patterns and deceptive design, in: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613905.3650997. doi:10.1145/3613905.3650997.

[46] E. Adar, D. S. Tan, J. Teevan, Benevolent deception in human computer interaction, in: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 1863–1872. URL: https://doi.org/10.1145/2470654.2466246. doi:10.1145/2470654.2466246.

[47] G. Y. Y. Wu, Silicon love: Deception, vulnerability, and artificial companions, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, CHI EA '25, Association for Computing Machinery, New York, NY, USA, 2025. URL: https://doi.org/10.1145/3706599.3720037. doi:10.1145/3706599.3720037.

[48] S. Sarker, S. Chatterjee, X. Xiao, A. Elbanna, The sociotechnical axis of cohesion for the is discipline: Its historical legacy and its continued relevance, MIS Quarterly 43 (2019) pp. 695–720, A1–A5. URL: https://www.jstor.org/stable/26848052.

[49] T. Kollmer, A. Hauser, V. Oberhofer, G. Blossey, A. Eckhardt, Uncovering drivers for the integration of dark patterns in conversational agents (2023).

[50] Y. Lu, C. Zhang, Y. Yang, Y. Yao, T. J.-J. Li, From awareness to action: Exploring end-user empowerment interventions for dark patterns in ux, Proc. ACM Hum.-Comput. Interact. 8 (2024). URL: https://doi.org/10.1145/3637336. doi:10.1145/3637336.

[51] T. Mildner, O. Cooney, A.-M. Meck, M. Bartl, G.-L. Savino, P. R. Doyle, D. Garaialde, L. Clark, J. Sloan, N. Wenig, R. Malaka, J. Niess, Listening to the voices: Describing ethical caveats of conversational user interfaces according to experts and frequent users, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, CHI '24, Association for Computing Machinery, New York, NY, USA, 2024. URL: https://doi.org/10.1145/3613904.3642542. doi:10.1145/3613904.3642542.

[52] N. Chomsky, I. Roberts, J. Watumull, The false promise of chatgpt, The New York Times (2023). URL: https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.

[53] H. A. Simon, Models of Man: Social and Rational; Mathematical Essays on Rational Human Behavior in a Social Setting, John Wiley & Sons, New York, 1957.

[54] A. Gambino, J. Fox, R. A. Ratan, Human-Machine Communication 1 (2020) 71–85. URL: https://search.informit.org/doi/10.3316/INFORMIT.097034846749023.

[55] I. Huvila, T. D. Anderson, E. H. Jansen, P. McKenzie, A. Worrall, Boundary objects in information science, Journal of the Association for Information Science and Technology 68 (2017) 1807–1822.

[56] A. Rai, J. Tian, L. Xue, Fair: A design theory for artificial intelligence fairness, Management Information Systems Quarterly (2026) 1–35. URL: https://doi.org/10.25300/MISQ/2026/17971. doi:10.25300/MISQ/2026/17971. arXiv:https://misq.umn.edu/misq/article-pdf/doi/10.25300/MISQ/2026/17971/19493/fl$_1$0.2

[57] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big? , in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[58] V. Hofmann, P. R. Kalluri, D. Jurafsky, S. King, Ai generates covertly racist decisions about people based on their dialect, Nature 633 (2024) 147–154.

[59] T.-Y. Hou, Y.-C. Tseng, C. W. T. Yuan, Is this ai sexist? the effects of a biased ai's anthropomorphic appearance and explainability on users' bias perceptions and trust, International Journal of Information Management 76 (2024) 102775. URL: https://www.sciencedirect.com/science/article/pii/S0268401224000239. doi:https://doi.org/10.1016/j.ijinfomgt.2024.102775.

[60] S. Simpson, J. Nukpezah, K. Brooks, R. Pandya, Parity benchmark for measuring bias in llms, AI

and Ethics 5 (2025) 3087–3101.
[61] B. A. Myers, A brief history of human-computer interaction technology, Interactions 5 (1998) 44–54. URL: https://doi.org/10.1145/274430.274436. doi:10.1145/274430.274436.
[62] A. Schmitt, Ensuring human agency: A design pathway to human–ai interaction, in: Proceedings of the International Conference on Information Systems (ICIS 2024), Association for Information Systems (AIS), Bangkok, Thailand, 2024. AIS Electronic Library (AISeL).
[63] R. Ciriello, Artificial companionship: the hedgehog's dilemma, rewaxed, AI & SOCIETY (2025) 1–2.