Volume 6, 2023 ISSN 2638-602X (print)/ISSN 2638-6038 (online)

Human-Machine Communication

ISSN 2638-602X (print)/ISSN 2638-6038 (online)

www.hmcjournal.com

Human-Machine Communication (HMC) is an annual peer-reviewed, open access publication of the Communication and Social Robotics Labs (combotlabs.org), published with support from the Nicholson School of Communication and Media at the University of Central Florida. Human-Machine Communication (Print: ISSN 2638-602X) is published in the spring of each year (Online: ISSN 2638-6038). Institutional, organizational, and individual subscribers are invited to purchase the print edition using the following mailing address:

Human-Machine Communication (HMC)
Communication and Social Robotics Labs
Western Michigan University
1903 W. Michigan Ave.
300 Sprau Tower
Kalamazoo, MI 49008

Print Subscriptions: Regular US rates: Individuals: 1 year, $40.
Libraries and organizations may subscribe for 1 year, $75.
If subscribing outside of the United States, please contact the Editor-in-Chief for current rate. Checks should be made payable to the Communication and Social Robotics Labs.

Human-Machine Communication
Volume 6

Seungahn Nah, University of Florida (U.S.A.)
Gina Neff, University of Cambridge (United Kingdom)
Jochen Peter, University of Amsterdam (Netherlands)
Sharon Ringel, University of Haifa (Israel)
Astrid Rosenthal-von der Pütten, RWTH Aachen University (Germany)
Eleanor Sandry, Curtin University (Australia)
Mauro Sarrica, Sapienza University of Rome (Italy)
Megan Strait, The University of Texas Rio Grande Valley (U.S.A.)
Satomi Sugiyama, Franklin University Switzerland (Switzerland)
Sakari Taipale, University of Jyväskylä (Finland)
David Westerman, North Dakota State University (U.S.A.)

CONTENTS

HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.1

Defining Dialogues: Tracing the Evolution of Human-Machine Communication

Andrew Prahl[1] and Autumn Edwards[2]

1 Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore
2 School of Communication, Communication and Social Robotics Labs, Western Michigan University, MI, USA

Abstract

This introduction to the volume discusses the evolving field of Human-Machine Communication (HMC), drawing on insights from the philosophy of science. We explore critical debates in the field, underscoring the importance of challenging assumptions, embracing interfield work, and fostering dialogue in shaping our understanding of HMC. Moreover, we celebrate the vibrant collaboration between disciplines that drives progress in HMC. This piece serves as an invitation to join the exploration of this collection and contribute to shaping the future of HMC.

Introduction

Introducing Volume 6 of *Human-Machine Communication* provides an opportunity to check how our field is evolving: Is it narrowing or growing in scope, converging toward theoretical uniformity, solving problems, or discovering new challenges? We explore these questions at a critical time for Human-Machine Communication. The public release of LLMs and generative AI tools (e.g., ChatGPT, Co-Pilot, Midjourney) in late 2022 is fueling a surge of interest in the opportunities and risks of machines that communicate. We stand before an exhilarating vista of social and scientific importance. At the same time, we must also heed the call of Hepp et al. (this volume) to maintain a critical and reflexive stance amidst the hype. As we look past the fanfare and aim to place this volume's contents into context, we anchor ourselves in the dialogue exploring the emergence and evolution of

CONTACT Autumn Edwards • autumn.edwards@wmich.edu • School of Communication, Communication and Social Robotics Labs • University of Western Michigan • Kalamazoo, MI

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

scientific fields. Thomas Kuhn's (1962) *The Structure of Scientific Revolutions* is likely the best-known work within the philosophy of science genre, but scores of other authors have explored the reasons new disciplines fail or succeed. Immersing ourselves in this literature, we find much in this volume to suggest that our discipline is evolving in a healthy way. We witness the active debate of core concepts, the use of interfield theories, and new dialogues that redefine relationships in the novel contexts where human-machine communication (HMC) occurs.

Progress Through Challenging Assumptions

In this volume's lead article, van der Groot and Etzrodt ask tough questions about one of the paradigms most central to our discipline: Computers Are Social Actors (CASA). Or, perhaps we have misspoken here. Should we have said Computers *As* Social Actors? This small difference in wording is how Van der Groot and Etzrodt organize their discussion on a much larger and unsettled debate in our field that they express as "Media Equivalence" versus "Media Evocation." Their explication of the differences between these perspectives weaves in many ideas that feel familiar to us in the Human-Machine Communication field but have nevertheless always felt slightly on edge as if they were built on assumptions that were yet to fully bake. For example, are people *really* just "mindlessly" responding to machines like they do humans, or do people interact with careful consideration of the "paradox state" of machines? Alternatively, are people *inspired* by machines as social partners, or are they *deceived* by machines? Van der Groot and Etzrodt offer a fascinating history of these competing views in which they step beyond the bounds of a literature review to chart an etymology of the media equivalence and evocation perspectives. Here, the words of philosopher Stephen Toulmin, who wrote extensively about the philosophy of science, resonate: "The novelty of the conclusion comes, not from the data, but from the inference: by it we are led to look at familiar phenomena in a new way, not at new phenomena in a familiar way" (Toulmin, 1972, p. 20). Van der Groot and Etzrodt's articulation of these perspectives compels us to reexamine our own bedrock assumptions. Van der Groot and Etzrodt rarely present any idea in their piece without a rebuttal to consider—forcing us to name, revisit, challenge, and reflect on our latent presuppositions about the social relations between people and communicative technologies.

It is easy to see such challenging questions as a fracturing force within our young discipline. Must we choose a side in each literature review, operationalization, and conference presentation? Toulmin (1972) would instead suggest that such debates that cut to the core of our field are commendable, "[We] demonstrate our rationality, not by a commitment to fixed ideas, stereotyped procedures, or immutable concepts, but by the manner in which, and the occasions on which, we change those ideas, procedures, and concepts" (p. 5). Van der Groot and Etzrodt's article provides one of those occasions by offering a sparklingly clear partitioning heuristic. Moreover, their article is not only characterized by repeated contrasts; Van der Groot and Etzrodt also offer some paths to resolving these competing views, echoing Toulmin's suggestion that the *way* in which we challenge concepts is important. Toulmin (1972) also cautions, however, that "[when] a set of concepts achieves unchallengeable authority in any field of enquiry, that discipline no longer faces 'scientific' problems" (pp. 189–190). In a similar vein, Kenneth Gergen (1978) reminds us that "the labeling of any given action is forever open to negotiation among interested parties, and the legitimacy of any observation statement is continuously open to change" (p. 1350). This holds particularly true considering that "patterns of human activity are themselves in a continuous state of emergence, *aleatoric* in the sense that they may largely reflect contemporary contingencies" (p. 1353), an issue underlined by HMC researchers (Edwards et al., 2019; Fortunati & Edwards, 2021; Gambino et al., 2020; Lombard & Xu, 2021) that is honored and extended in the present article. Thus, while Van der Groot and Etzrodt's work brings us closer to reconciling competing perspectives, we hope that the spirit of looking at fundamental ideas in new ways persists; in doing so we ensure the continued evolution and health of our field.

Following Van der Groot and Etzrodt's provocative lead article, Marco Dehnert provides a calming refrain in the second article of this volume. Dehnert uses the archipelago metaphor to grasp our field's current state and plot a forward path. Dehnert suggests that rather than rush to congeal around certain beliefs, the HMC community should *embrace* conceptual ambiguity, viewing different subjects, methods, and conceptualizations as a chain of islands ripe for exploration. Dehnert explicates his vision of the HMC field as archipelagic and "made up not of a coherent subject or a cohesive body of literature, but rather entails a variety of islands differing in shape, size, location, and proximity to one another" (p. 32).

Dehnert's archipelago metaphor also emphasizes the diversity in HMC. Many disciplines and methodological approaches have lent their insights to the burgeoning understanding of HMC. Dehnert emphasizes that in HMC, no single theoretical, methodological, or technological island takes precedence over others. Instead, what comes "first" in HMC is an appreciation for the multitude of approaches to our field of study and a realization that conflicts (like those explored in the lead article) are the strength of HMC. What strikes us about Dehnert's analogy is that HMC scholarship's joy lies not in guarding our islands but in navigating between them. In turn, Dehnert's approach inspires us to be bridge-builders, to add to the rich tapestry of our islands' linkages and differences. We are reminded here of D'Agostino's (2012) study of how disciplines expand through fostering the infusion of diverse perspectives and maintaining "incentives for individuals not to converge on some canonical articulation of the abstract and concrete (and hence multiply interpretable) features of their 'code'" (p. 347). This notion resonates with the essence of Dehnert's vision: to nurture diversity and foster dynamic connections. Thus, it is not the destination but the journey itself—the exploration, the rich diversity, the connections—that truly defines the vibrant evolution field of HMC.

With the recent innovations in generative AI and the public release of AI tools, this journey is fast becoming dizzying. Offering reprieve, this volume's third article by Hepp et al. provides a counterpoint to the techno-hype. The message? Not to drown in the hyperbole but to see systems like ChatGPT, LaMDA, and Luminous as harbingers of a new communicative era. Hepp et al. urge us to consider the new communication wave seriously and avoid blind acceptance. Their words echo Van der Groot, Etzrodt, and Dehnert: Don't cling to the familiar, don't submit to the allure of conformity; do embrace the debate. Hepp et al. steer us back to reflexivity, and we remember Toulmin's (1953) words: "One can distinguish, in any science, between the problems which are currently under discussion, and those earlier problems whose solutions have to be taken for granted if we are even to state our current problems" (p. 81). As such, we reflect upon Hepp et al.'s reminder that the study of automated communication is not new (Fortunati & Edwards, 2020, 2022). A long and rich history of research in cybernetics, for example, has confronted and foreseen the problems of machine communicators. But, heeding both Toulmin's words and Hepp et al.'s reflexivity, we stop short of accepting previous "solutions" to these problems. As we confront the reality of automated communication, we should recognize this as an opportunity to reflect, question, and reimagine our understanding of HMC. Hepp et al. effectively join the same hymn with Van der Groot, Etzrodt, and Dehnert: Challenge the status quo. Pay attention to our perspective. Step into the unknown. Here is the lifeblood and future of the HMC field.

Progress Through Interfield Work

As encouraging as the opening articles were in questioning the familiar, there are other indications of the health of the HMC field in this volume. In reading many of this volume's articles, a critical paradigm of scientific advancement comes into focus: interfield work. Interfield work is a merging of theories, methods, and perspectives from different scientific fields. Lindley Darden and Nancy Maull (1977) present a compelling vision in *Interfield Theories,* arguing that interfield theory, and thus interfield work generally, is not just academic cross-pollination but an essential catalyst for scientific progress.

Against this backdrop, we come to the fourth article by Stephens et al. They study an aspect of human-machine communication that is rarely featured but is nevertheless critical: the humans who train machines. "Humans are involved in many steps of a machine learning system's pipeline, but the most common is in labeling data to create a training set for supervised machine learning" (p. 66). Stephens et al. highlight what is easy to forget: machines themselves are the product of human-machine communication. Situated in this early stage of machine development, their work has a compellingly symphonic feel as they unshroud a human-machine relationship that can be both harmonious and discordant. In this dance between human and machine, the partners exchange leading roles via subtleties of feedback and correction.

The dance gets even more fascinating when Stephens et al. venture into the territory of human emotions within these interactions. They reflect on how more involved training processes should consider ". . . how the humans feel and experience this more involved type of interaction" (p. 66). Stephens et al.'s investigation of the emotional landscape of human-machine communication—paired with the efficiencies studied by business scholars and development processes studied in the information systems field—bears a resemblance to Darden and Maull's (1977) insight that "interfield theories are likely to be generated . . . when questions arise about that phenomenon within a field which cannot be answered with the techniques and concepts of that field" (p. 50). Thus, Stephens et al.'s study exemplifies the interfield work that leads to the expansion of scientific horizons. Stephens et al. also offer a clever insight into the position of HMC, specifically in this convergence of disciplines: noting that human-machine communication places greater emphasis on the relationships between humans and machines, as opposed to the narrowing focus on human-machine interaction. This broader perspective offered by HMC underlies Stephens et al.'s findings and shows the benefits of the interfield work that the HMC perspective fosters. This is excellent news for our young discipline. As the pace of advancement in human-machine communication accelerates, interfield work becomes not only a theoretical endeavor but also a practical necessity for HMC.

The nascent importance of interfield work is just as evident in the volume's next article by Concannon, Roberts, and Tomalin. As Concannon et al. point out, the intense development of empathetic agents ("systems capable of responding appropriately to emotional content") is reflected in the recent marketing of many AI "companions" and "assistants," perhaps skewing human expectations of system capabilities. In this context, Concannon et al. signal a different kind of interfield work, merging computational advances with insights from social sciences, effectively extending the reach of the interfield contributions that result from HMC inquiry.

As examined by Concannon et al., empathy in human-machine communication is not about replicating human affective processes but attending appropriately to the emotional expressiveness of an utterance. Interestingly, as articulated in the article, the need to understand empathetic communication is anchored in a wider conversation about how language is used in ways perceived as empathetic, a perspective that draws heavily from interactional linguistics. Although the agents do not possess empathy in the human sense, they can use language to display empathetic concern. Thus, Concannon et al. invite the incorporation of interactional linguistics into the broader study of HMC. This approach reflects Darden and Maull's (1977) suggestion that interfield work flourishes when "A field may investigate the structure of entities or processes, the function of which is investigated in another field" (p. 49). As such, Concannon et al.'s linguistic analysis allows for the investigation of the structure of language in human-machine communication, but their analysis of the function of these words draws upon literature and theory from HMC. Consider, as just one small slice of their rich analysis, Concannon et al.'s examination of the strategies machines use to emulate empathetic communication, some of which (e.g., discussing a "shared" experience) hinge on a distinctly human factor: the suspension of disbelief. However, this suspension of disbelief is by no means a given; it is a highly individual process. If a user cannot suspend their disbelief, the machine's attempt at empathy invariably fails. This nuanced point, though just one facet of Concannon et al.'s study, underscores the complex issues intrinsic to the field of HMC. It highlights the imperative for scholars in HMC to grapple with profoundly human constructs like empathy, and it signals to the broader academic and industrial communities that the creation of empathetic machine agents isn't only the purview of engineers. Scholars from a multitude of fields all have essential roles to play. This collective interfield effort is both the challenge and the opportunity that HMC offers—a testament to the enduring importance of interfield work in HMC.

As we navigate further into this volume, the influence and necessity of interfield work continue to be evident and increasingly significant. The upcoming triptych formed by the sixth, seventh, and eighth studies brings together disparate but complementary academic disciplines to investigate the intricacies of human-machine interactions. We find ourselves witnessing human exclusion from the conversation between physical robots, interacting with virtual bots in an online chat, and engaging in real-life interactions with physical robots. The human experience across these contexts prompts us to observe, infer, and model bot behaviors in our minds, further blurring the lines between human-human and human-machine interactions. In the sixth article, Rosenthal-von der Pütten and Bock interweave robotics and social psychology to study human feelings of social exclusion caused by machines. It is an idea as fascinating as it is unexpected. In exploring this peculiar terrain, the research strays from the well-beaten path of studying how humans can feel *closer* to machines. One of the pioneers of interdisciplinary thought and practice, Julie Thompson Klein, writes about how this creativity is an integral part of pushing boundaries in science; and is characteristic of interdisciplinary (interfield) work (1996). As Rosenthal-von der Pütten and Bock's study unfolds, it is difficult not to recall Klein's (1996) assertion, "Interdisciplinary work is critical in that it exposes the inadequacies of the existing organization of knowledge to accomplish given tasks" (p. 14). Likewise, through an impressive fusion of fields, Rosenthal-von der Pütten and Bock build an experimental environment that probes this uncharted territory, teasing out surprising human emotional responses linked to the quintessential human need for inclusion. When machines engage in exchanges amongst themselves, employing silence or a coded language unfathomable to human listeners, they provoke fundamental questions about our social condition. Can a human feel left out by a discussion held in code? Can a mere machine, a contraption of wires and circuits, trigger feelings of social isolation within us? And as we grapple with these seemingly surreal questions, we also wonder how this perceived exclusion may alter our trust in our tech-based counterparts. By posing these questions, Rosenthal-von der Pütten and Bock set the stage for a deeper exploration into this captivating realm of research that transcends disciplinary boundaries; it is an unconventional investigation that aligns remarkably with Klein's vision of breaching disciplinary silos to pursue new knowledge.

The concept of "communicative anthropomorphization," as proposed by Laaksonen, Laitinen, Koivula, and Sihvonenin in this volume's seventh article, is a compelling contribution to the field of Human-Machine Communication (HMC). Laaksonen et al. redefine anthropomorphization as more than just a design or psychological process; they present it as an intrinsic characteristic of the human-machine communicative process. Klein's (1996) claim that boundaries can be redrawn as interdisciplinary connections are made and solidified aptly captures Laaksonen et al.'s transformative approach that redraws the conceptual boundaries of anthropomorphization. For Laaksonen et al., it's not just about understanding how chatbots are programmed to mimic human conversation or how humans psychologically perceive bots. Rather, it's about converging these and other insights to examine their interaction and mutual influence within the sociotechnical realm of HMC. As Darden and Maull (1977) posited, "Interfield theories explain and make explicit the relationships between different domains of knowledge." (p. 48). In their explication of communicative anthropomorphization, Laaksonen et al. construct an interfield concept that can be built upon and perhaps serve as the foundation for an interfield theory. "Communicative anthropomorphization" gives us a tantalizing glimpse into the field's future, where the boundaries between disciplines dissolve when integrated into the HMC perspective and where knowledge is a vibrant tapestry woven from countless interconnected threads. We should celebrate this progression in HMC.

Following the theme of interfield exploration in HMC research, Stein and Banks, in the volume's eighth article, bring forth a detailed study that deftly weaves elements of parasocial contact theory, social psychology, and media studies, among others, delivering a multidimensional analysis of human-robot interactions. It's a paradigmatic example of how interfield theories are formulated: by explaining and articulating the relationships between different knowledge domains (Darden & Maull, 1977). Recognizing the oft-overlooked influence of mass media representations of robots, Stein and Banks incorporate the concept of parasocial contact—an idea that biases toward dissimilar others, such as humanoid robots, can be alleviated through positive media exemplars. In another instance of interdisciplinary melding, Stein and Banks engage with social psychology literature, bringing insights from intergroup dynamics research to build their rationale. They consider how humans engage with robots as if they were part of an outgroup while at the same time acknowledging the ontological differences that separate them. Stein and Banks's work is reminiscent of the idea expressed by Klein that boundaries can be redrawn as interdisciplinary connections are made and solidified. With this boundary-breaking research, Stein and Banks broaden the horizons of HMC in a way that enhances our understanding of human-machine interactions and opens new avenues for exploration moving forward.

Dynamic Dialogues

With the stage set for discovery, we delve into the last two articles, which present us with a compelling aspect that bolsters the vitality of the discipline and holds an essential role in human-machine communication: dialogue. Not only does dialogue serve as a bridge between paradigms and fields, as demonstrated in the preceding articles, but it also forms the very essence of social existence. Adopting a dialogic perspective is recognizing the interactive, dynamic, and contextual nature of communication and meaning-making processes. This aligns with the tenets of critical realism, introduced by Roy Bhaskar (1975), which suggests the "real" differs from our subjective experiences of it and is largely unobservable. The best approach to science, according to him, is to study how people experience and interact in the world. Bhaskar's view is that scientific progress is synonymous with rethinking existing knowledge in different, more timely contexts. This view of progress is highly pertinent to the discussion in the last two articles in this volume (Vitak et al.; Youk & Park). These articles rejuvenate theories of human interaction born from dialogism and dialectics as relational processes are reimagined in the new contexts of navigation and voice-based assistants.

Transitioning into the latter part of the 20th century, theories like Baxter and Montgomery's and Rawlins' relational dialectics—downstream descendants of dialogism—infused fresh perspectives into stale debates by asserting that relational processes resisted linear explanation. Should we tell our intimate partners everything or hold back? Should we prioritize togetherness or maintain independence? Is it good to be predictable or spontaneous? The dialectics approach answered with an enthusiastic "both/and," contending that each apparently contradictory need or discourse is simultaneously valued, communicatively constructed, and constantly evolving. It highlights the complex, dialogic nature of relational processes, suggesting human activity requires constant negotiation and adaptation. Dialogism's potential to transform social theory and practice call to mind what Kenneth Gergen (1978) terms the "generative capacity" of theory "to challenge the guiding assumptions of the culture, to raise fundamental questions regarding social life, to foster reconsideration of that which is 'taken for granted,' and thereby to furnish new alternatives for social action" (p. 1346). And it's the generative spirit of dialogue that the final two articles deliver as they extend dialogue-based theories into the context of HMC. They shine fresh light on important, dynamic, and fraught contexts of HMC: navigating privacy with voice-based assistants (VBAs; Vitak et al.) and determining blame in navigation technology (Youk & Park). Both articles illustrate how, in the vibrant dialogic sphere, understanding social life and relationships becomes a generative process.

In the ninth article Vitak, Kumar, Liao, and Zimmer employ Communication Privacy Management theory to assess the dynamics of boundary regulation and privacy issues in using VBAs, technologies that "blur boundaries between public and private spaces." We learn, for instance, that VBA users often conceived of privacy as futile and trained their gaze instead on usefulness, a perceived opposition that may be uniquely significant in HMC. Conversely, non-users stressed the need for trustworthy providers and control over access to their information prior to adoption. The results elucidate the potential of dialogue-based theories to highlight the tension-infused experience of engaging in HMC.

Choice and tension also assume a starring role in Youk and Park's final article, exploring why users are more likely to adopt some driving navigation technologies over others. To investigate drivers ' use intentions, Youk and Park orchestrate a beautiful collision between the Theory of Communicative Responsibility and the Technology Adoption Model. Given that human-human interaction involves autonomous entities with the ability to comprehend and formulate messages in complex contexts—and machines only simulate communicative competence—Youk and Park anchor their research squarely in the process of meaning-making, unafraid to challenge theoretical assumptions in a new navigational interaction context. Results highlight, for example, the necessity of considering the purpose and role of the machine in our HMC research alongside the context, an aspect which has been highlighted persuasively in earlier volumes, as well (Gambino & Liu, 2022). This kind of theoretical work marks a pivotal moment for the field of HMC by advancing the dialogue between users and machines in exciting, transformative ways ripe with the generative capacity to "provoke debate, transform social reality, and ultimately reorder social conduct" (Gergen, 1978, p. 1346).

With the final two articles illustrating the potential of dialogism to challenge, clarify, and recast how we understand and explain HMC and the wisdom of Bhaskar's critical realism woven into our explorations, we are reminded of the interfield nature of our discipline that informs related communities of design, policy, and public concern. Returning to relational processes, we imagine them in the context of HMC: Are communicative machines essentially tools or social actors? Should we reveal or conceal our private information when talking to virtual agents? Are agency and responsibility affixed to the human or the machine? Dialogism's reminder is that we may not have to choose. Just as Van der Groot and Eztrodt, and Dehnert remind us in the opening articles, we should instead question the binary status quo and find jewels of comprehension in studying how ordinary people, as well as scholars, grapple with the tensions in human-machine communication (see, e.g., Abendschein et al., 2022; Westerman et al., 2020). Moreover, we should remember that we, as researchers, are part of the dialogue. Hordyk et al., (2013) note that ". . . critical realist data analysis will have a significantly distinctive transforming influence on the researcher" (p. 22). Mutual shaping captures the essence of the intertwined journey of discovery: As we seek to understand HMC, we are not mere observers but are also active participants within the dialogic process.

Moving Forward

Reflecting on this volume of *Human-Machine Communication*, we revel in the progress our field is making. Whether it is Toulmin's (1972) call to question assumptions, Darden and Maull's (1977) interfield collaborations, or Gergen (1978) and Bhaskar's (1975) plea to reimagine interaction, there are numerous indications that the emergent field of HMC is healthy and promises to deliver more theoretical, methodological, and practical breakthroughs going forward. The ever-present interdisciplinary and dialogic nature of the work in this volume is a reminder of the multiplicity of voices and discourses that together constitute the HMC processes and phenomena we study. In fact, it appears to us that a good number of the formative debates (or should we say dialogues?) in the field may be explored as communication-based tensions best understood in their wholeness. As such, this volume, in its wholeness, stands as a testament to the vibrant dialogue between disciplines, co-authors, editorial staff, and all the scholars who reviewed manuscripts for this issue—to whom we are endlessly grateful for nurturing the spirit of collaboration driving progress. With open minds and engaged hearts, we invite you to join us in exploring this collection and shaping the future of Human-Machine Communication.

Author Biographies

Andrew Prahl is an assistant professor in the Wee Kim Wee School of Communication & Information, Nanyang Technological University, Singapore.

https://orcid.org/0000-0003-3675-3007

Autumn Edwards is a professor in the School of Communication, Western Michigan University, Kalamazoo, MI, USA.

https://orcid.org/0000-0002-5963-197X

References

Abendschein, B., Edwards, A., & Edwards, C. (2022). Novelty experience in prolonged interaction: A qualitative study of socially-isolated college students' in-home use of a robot companion animal. *Frontiers in Robotics and AI*, 51. https://doi.org/10.3389/frobt.2022.733078

Bhaskar, R. (1975). *A realist theory of science*. Routledge. https://doi.org/10.2307/2219031

D'Agostino, F. (2012). Disciplinarity and the growth of knowledge. *Social Epistemology*, *26*(3–4), 331–350. https://doi.org/10.1080/02691728.2012.727192

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science*, *44*(1), 43–64. https://doi.org/10.1086/288723

Edwards, A., Edwards, C., Westerman, D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior*, *90*, 308–314. https://doi.org/10.1016/j.chb.2018.08.042

Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication*, *1*, 7–18. https://doi.org/10.30658/hmc.1.1

Fortunati, L., & Edwards, A. (2021). Moving ahead with human-machine communication. *Human-Machine Communication*, *2*, 7–28. https://doi.org/10.30658/hmc.2.1

Fortunati, L., & Edwards, A. (2022). Framing the psycho-social and cultural aspects of human-machine communication. *Human-Machine Communication*, *4*, 7–26. https://doi.org/10.30658/hmc.4.1

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–85. https://doi.org/10.30658/hmc.1.5

Gambino, A., & Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication, 4*, 111–130. https://doi.org/10.30658/hmc.4.6

Gergen, K. J. (1978). Toward generative theory. *Journal of Personality and Social Psychology*, *36*(11), 1344–1360. https://doi.org/10.1037/0022-3514.36.11.1344

Hordyk, S. R., Soltane, S. B., & Hanley, J. (2013). Sometimes you have to go under water to come up: A poetic, critical realist approach to documenting the voices of homeless immigrant women. *Qualitative Social Work*, *13*(2), 203–220. https://doi.org/10.1177/1473325013491448

Klein, J. T. (1996). *Crossing boundaries: Knowledge, disciplinarities, and interdisciplinarities*. University Press of Virginia. http://archive.org/details/crossingboundari0000klei

Kuhn, T. S. (1962). *The structure of scientific revolutions*. University of Chicago Press.

Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication*, *2*, 29–55. https://doi.org/10.30658/hmc.2.2

Toulmin, S. E. (1953). *The philosophy of science*. Hutchinson's University Library.

Toulmin, S. (1972). *Human understanding: The collective use and evolution of concepts*. Princeton University Press.

Westerman, D., Edwards, A. P., Edwards, C., Luo, Z., & Spence, P. R. (2020). I-It, I-Thou, I-Robot: The perceived humanness of AI in human-machine communication. *Communication Studies*, *71*(3), 393–408. https://doi.org/10.1080/10510974.2020.1749683

HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.2

Disentangling Two Fundamental Paradigms in Human-Machine Communication Research: Media Equation and Media Evocation

Margot J. van der Goot[1] and Katrin Etzrodt[2]

1 Amsterdam School of Communication Research/ASCoR, University of Amsterdam, The Netherlands
2 Institute of Media and Communication, Technische Universität Dresden, Germany

Abstract

In this theoretical paper, we delineate two fundamental paradigms in how scholars conceptualize the nature of *machines* in human-machine communication (HMC). In addition to the well-known Media Equation paradigm, we distinguish the Media Evocation paradigm. The Media Equation paradigm entails that people respond to machines *as if* they are humans, whereas the Media Evocation paradigm conceptualizes machines as objects that can evoke reflections about ontological categories. For each paradigm, we present the main propositions, research methodologies, and current challenges. We conclude with theoretical implications on how to integrate the two paradigms, and with a call for mixed-method research that includes innovative data analyses and that takes ontological classifications into account when explaining social responses to machines.

Keywords: CASA, computers are social actors, computers as social actors, human-machine communication, media equation, media evocation, mixed-method research

Introduction

In this theoretical paper, we delineate two fundamental paradigms in how scholars conceptualize the nature of *machines* in human-machine communication (HMC). In addition to the well-known Media Equation paradigm (e.g., Reeves & Nass, 1996), we distinguish the

CONTACT Margot J. van der Goot • m.j.vandergoot@uva.nl • School of Communication Research/ASCoR • University of Amsterdam, P.O. Box 15791 • 1001 NG Amsterdam, The Netherlands

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Media Evocation paradigm (following the concept of *evocative objects* in Turkle's influential work of 1984 and 2007). We argue that these two paradigms fundamentally differ in their explanations for why humans respond socially to machines, in their research questions, and in their research methodologies. The key notion in the Media Equation paradigm is that people respond mindlessly to machines *as if* they are humans. In contrast, the Media Evocation paradigm conceptualizes machines as objects that are *betwixt and between* former diametrical opposites—such as person versus thing— evoking reflection and negotiation processes about the nature of the object but also about ourselves and human identity. Here, social responses are potentially due to the fact that machines *are* a kind of social actors—albeit different ones than human social actors.

For this succinct summary of the two paradigms we were inspired by the observation that articles referring to the CASA framework—which is the most often employed framework to guide HMC research (e.g., Fortunati & Edwards, 2021; Gambino et al., 2020; Spence, 2019)—use the acronym both for "computers *as* social actors" as well as for "computers *are* social actors," usually without addressing this difference.[1] In our interpretation, the main notions of the Media Equation paradigm imply that humans respond to computers or machines *as* social actors, whereas in the Media Evocation paradigm machines *are* (some type of) social actors. The interchangeable use of "as" and "are" in the CASA acronym could be interpreted as an indicator of the fact that thus far HMC research lacked an explicit differentiation between the Media Equation and Media Evocation paradigms.

The current paper aims to present these two paradigms by showcasing their main propositions, associated research methodologies, and current challenges. Herein, we rely on classical works (particularly Nass et al., 1994; Nass et al., 1993; Reeves & Nass, 1996; Turkle, 2005, orig. 1984), as well as recent theoretical and empirical publications within the HMC field. We argue that these classical works have been visionary in drawing our attention to the huge importance of computers in our lives and to the intriguing observation that humans respond socially to these even though they know that they are not communicating with a human. At the same time, HMC researchers have pointed out that the technical developments that now enable us to interact with AI-enabled communicators such as social robots, chatbots, voice assistants that can talk with us, know our name, distinguish our voice, and learn our preferences make it pressing for the HMC research community to revisit our theorizing and decide how to move forward (e.g., Fortunati & Edwards, 2020, 2021; Fox & Gambino, 2021; Gambino et al., 2020; Guzman, 2018; Guzman & Lewis, 2020; Lombard & Xu, 2021; Spence, 2019).

In our conclusions, we note that particularly the Media Equation paradigm gained much momentum in the HMC field, focusing on the notion that people mindlessly respond socially toward machines and typically involving quantitative, experimental research methods. However, current interactions with AI-enabled communicators make it pivotal for HMC scholars to also focus on the Media Evocation paradigm, which emphasizes reflections and negotiations regarding the ontological boundaries concerning (among others) humans and machines. In our conclusions, we also make a plea for more qualitative research, and mixed-method research that includes innovative data analyses and that takes ontological

1. A Google Scholar search listed about 2,170 articles mentioning "Computers *are* Social Actors," whereas about 1,350 articles mentioned "Computers *as* Social Actors" (date of search: February 14, 2022).

classifications into account when explaining social responses to machines. As such, our recommendations echo previous calls for more research looking into how machines blur the ontological boundaries surrounding what constitutes human, machine, and communication (e.g., Edwards & Edwards, 2022; Fortunati & Edwards, 2020, 2021; Guzman & Lewis, 2020) and for more inductive and mixed-method projects (e.g., Gambino et al., 2020; Richards et al., 2022). Our intention is that our comparison of the two paradigms will inspire HMC research to recognize, delineate, and integrate both of these two mindsets that are so fundamental in this field. In doing so, we can increase our understanding of how and in which situations Media Equation and Media Evocation processes are at play, thus gaining a more holistic understanding of our social responses to machines.

Media Equation

Foundational Work

Theoretical Notions

The first paradigm originates in the book *The media equation: How people treat computers, television, and new media like real people and places*, in which Reeves and Nass (1996) reported that they found that individuals' interactions with computers and new media are fundamentally social and natural. They concluded that "media equal real life" and posed that this applies to everyone, applies often, and is highly consequential (p. 5). They emphasized that people have these social responses even though people believe these are not reasonable, and even though they do not think these responses characterize themselves (p. 7). Nass et al. (1993, p. 111) wrote that "[u]sers can be induced to behave *as if* computers were human, even though users know that the machines do not actually possess 'selves' or human motivations," and Kim and Sundar (2012, p. 241) posited that "[e]verybody knows that a personal computer is not human. [. . .] Yet, we respond to it socially." Thus, computers are conceptualized from the perspective of what they are *not,* that is, they are *not* human. Therefore, in the CASA acronym, the adverb "as" (i.e., computers *as* social actors) seems most appropriate for the Media Equation paradigm.

Within this perspective, the paradoxical situation that users constantly exhibit social responses toward computers, while consciously being aware that this behavior may be inappropriate when exhibited toward nonhuman entities, naturally led to a focus on mindless attribution processes (Fortunati & Edwards, 2021, p. 13). People treat computers in some of the ways they treat humans by *mindlessly* applying to them the same social scripts they use in human-human interactions (Nass & Moon, 2000). Reeves and Nass (1996, p. 252) wrote that these automatic responses can be initiated with "minimal cues" (p. 253). Machines with limited cues—for instance words as output—induce individuals to employ human-oriented decision rules that they believe are inappropriate for assessing machine behavior (Nass et al., 1993, p. 111).

Reeves and Nass's (1996) explanation for this phenomenon that people are not evolved to 20th-century technology, and that modern media engage old brains (p. 12). That is, people respond to simulations of social actors and natural objects as if they were in fact social and natural: "absent a significant warning that we've been fooled, our old brains hold sway and we accept media as real people and places" (p. 12). The notion of "ethopoeia" (a direct response to an entity as a human *while knowing* that the entity does not warrant human treatment or attribution, Nass & Moon, 2000, p. 94) was inspired by Langer's (1989/2014, 1992) work on mindfulness and mindlessness. Although the authors of the seminal Media Equation publications acknowledged that "people can be trained to be more mindful of context cues" (Nass & Moon, 2000, p. 98), and that they might be able to think their way out of primitive, automatic responses, they concluded that this strategy makes the process harder and is not typical or usual (Reeves & Nass, 1996, p. 13). They did not see this as a deficiency or dysfunction, but rather as useful and reasonable: people automatically assume reality because throughout evolution there was no reason to do otherwise (Reeves & Nass, 1996, pp. 252–253). In sum, the focus in this paradigm was on the identification of social attitudes, behaviors, and rules that are mindlessly triggered when humans interact with machines (e.g., Nass & Moon, 2000, p. 99).

Research Methodology

In the original CASA experiments (e.g., Nass et al., 1994, 1993; Reeves & Nass, 1996), manifest manipulations were related to behavioral responses, thus inferring mindless processes without asking for self-reports. The research design included the following steps: pick a social science finding about how people respond to each other or to the natural environment (e.g., about politeness); change "human" to "computer" in the theoretical statement; replace one or more humans with computers in the experiment; provide the computer with characteristics associated with humans (e.g., language output, responses based on multiple prior inputs, displaying roles traditionally filled by humans, human-sounding voices), and determine whether the social rule still applies (Nass et al., 1994, p. 72; Reeves & Nass, 1996, pp. 14–15). Thirty-five of such studies led Reeves and Nass (1996, p. 6) to formulate the Media Equation paradigm: media take the place of real people and places. They were very straightforward about not being interested in users' reflections. About their experimental research, they wrote that "these methods do not rely on people's ability to be introspective, and they provide objective data. If we had asked people to comment on whether they were polite to computers [. . .], we would have had nothing to report" (p. 255).

Recent HMC Research

HMC Theorizing

Taking these Media Equation notions as a starting point, recent theoretical contributions focused on how to conceptualize and study humans' interactions with AI-enabled communication technologies such as social robots, chatbots, and virtual agents. Updating the foundational theoretical notions that were introduced in the early 1990s is deemed necessary because people have changed (i.e., they have gained experience with artificial agents), technologies have changed (i.e., recent technologies are much more sophisticated in terms of interactions and anthropomorphic features), and affordances have changed (i.e., what users can do with technologies has developed significantly) (e.g., Fortunati & Edwards, 2021; Fox & Gambino, 2021; Gambino et al., 2020; Lombard & Xu, 2021; Sundar, 2020).

In these theoretical contributions, three main points stand out in particular. First, whereas Reeves and Nass (1996) suggested that automatic responses can be initiated with "minimal cues," authors now propose to further differentiate how single social cues and combinations of these evoke social responses (Gambino et al., 2020; Lombard & Xu, 2021; Sundar, 2020). Social cues have been defined as physical and behavioral features displayed by a social actor, of which a social actor's voice, humanlike appearance, and eye gaze are examples (Lombard & Xu, 2021, pp. 31–32). Importantly, in their Media Are Social Actors (MASA) paradigm, Lombard and Xu (2021) formulated testable propositions on how variations in the quality and quantity of such cues may lead to medium-as-social-actor presence and social responses.

Second, in response to the focus on mindlessness (e.g., Nass & Moon, 2000), Lombard and Xu (2021) proposed mindless and mindful anthropomorphism as two major complementary mechanisms that help to understand people's social responses to technology. Depending on the social cues, individual factors (such as personality and experience with technologies), and contextual factors, mindless or mindful anthropomorphism may be activated. Relatedly, in their work on robots and the Media Equation effect, Złotowski et al. (2018) put forth that anthropomorphism may be the result of a dual process: first, a fast and intuitive (Type 1) process that quickly classifies an object as human-like and results in implicit anthropomorphism, and second, a reflective (Type 2) process that is based on conscious effort and results in explicit anthropomorphism. This ties in with the metatheory of dual processing which increasingly receives attention in HMC research (Koban & Banks, 2023). In addition, Lombard and Xu (2021) suggested that there may be other possible explanations for social responses—that have received less attention in the HMC literature—such as for instance the source orientation explanation, the cognitive load explanation, and folk explanations of social behavior (p. 40).

Third, Gambino et al. (2020) argued that—in addition to the mindless application of human-human social scripts—the mindless application of human-media social scripts may also be at play. This is related to Sundar's (2020) notion of the machine heuristic, which is a mental shortcut whereby people attribute machine characteristics when they make judgments about an interaction (p. 7). Usually positive stereotypes of machines (i.e., they are rule-governed, precise, accurate, objective, neutral, and they do not gossip) as well as usually negative ones (i.e., they are mechanistic, unemotional, cold, and prone to being hacked) form the basis for these heuristics.

HMC Research Methodology

HMC research within the Media Equation paradigm mainly relies on experimental designs. Although these studies typically did not exactly follow the steps that were characteristic for the foundational experiments and are thus not replications of the classical work (with exceptions such as Leichtmann & Nitsch, 2021), some of these did test whether aspects of human-human communication (e.g., correspondence bias and the social desirability effect) also occur in human-robot interactions (e.g., Edwards & Edwards, 2022; Leichtmann & Nitsch, 2021). This type of work has led some authors to conclude that empirical results do not consistently support CASA's predictions (e.g., Fox & Gambino, 2021; Leichtmann & Nitsch, 2021) and that there are differences between how participants judge humans and machines such as robots (e.g., Edwards & Edwards, 2022).

So far, the notions of mindless and mindful processing and dual processing (Koban & Banks, 2023; Lombard & Xu, 2021) have found their way into HMC research in experimental studies in which both mindless and mindful anthropomorphism were included as mediators. A first publication in this line of research was Kim and Sundar's (2012) article in which they challenged Nass and Moon's (2000) notion that anthropomorphism involves the thoughtful, sincere belief that the object has human characteristics and thus cannot explain social responses (p. 93). Kim and Sundar wrote that anthropomorphism could also be automatic and mindless (p. 242), and they thus set out to examine whether the tendency to treat human-like agents as human beings is conscious (mindful) or nonconscious (mindless). Mindless anthropomorphism was measured by asking participants how well the adjectives likeable, sociable, friendly, and personal described the website (with/without a human-like agent), whereas mindful anthropomorphism was assessed by asking participants directly whether they perceived the website as humanlike/machinelike, natural/unnatural, or lifelike/artificial (Powers & Kiesler, 2006). Following this example, recent experiments that investigated the effects of (social cues in) chatbots have also included these two measures as mediators (e.g., Araujo, 2018; Ischen et al., 2020; Zarouali et al., 2021, see also van der Goot, 2022). This is a deviation from the classical Media Equation work, in which self-reports and thus such mediators were deemed unnecessary.

These types of studies are needed to further test Lombard and Xu's (2021) propositions, disentangling the effects of varying social cues on mindless and mindful processing. However, for differentiating these types of processing, the current explicit measures of mindless and mindful anthropomorphism are not uncontested, and researchers aim to move forward by using a combination of methods and measures including behavioral measures, interviews, and open-ended questions; explicit and implicit measures; and psychophysical measures such as fMRI and EEG (e.g., Lombard & Xu, 2021; van der Goot, 2022). More specifically, two-response procedures or conflict-detection procedures in combination with eye-tracking may help to make a clear distinction between Type 1 and Type 2 processes (Koban & Banks, 2023).

Media Evocation

Foundational Work

Theoretical Notions

We called the second paradigm Media Evocation (referring to Turkle's concept *the evocative object)*, which she proposed in her influential book *The second self: Computers and the human spirit* (2005, orig. 1984) and later elaborated on in her book *Evocative objects* (2007). Herein, the computer is conceptualized in terms of its "second nature" as an evocative object: an object that provokes self-reflection (Turkle, 2005, p. 2), fascinates, disturbs equanimity and precipitates thought (p. 19), and a problematic object that defies easy categorization and troubles the mind (p. 4). The computer stands "betwixt and between," in some ways on the edge of mind, thus raising questions about mind itself (p. 29). Thus, in contrast to the absence of a conceptualization of computers in the Media Equation paradigm, computers are conceptualized from the perspective of what they *are.* In this paradigm, users respond socially to machines—and even develop relationships with them—because the machine's evocative and "betwixt and between" nature changes how we think about what a social actor is—and that it does not necessarily have to be a human. Therefore, in the CASA acronym, the adverb "are" (i.e., Computers *are* Social Actors) seems most appropriate for the Media Evocation paradigm.

In contradiction to the propositions in the Media Equation paradigm, it is not the user's behavior, but instead, the machine's state that is paradoxical. Turkle (2005, p. 326) draws on Turner's (1969) work on liminal objects and Douglas's (1966) observations about marginality, referring to ambiguous states, or disorientation of individuals or groups (e.g., adolescents): a threshold state in which they find themselves after disengaging from the prevailing social order or pattern. Similar to those individuals or groups, machines possess neither properties of their previous state (e.g., "thing") nor those of the future one (e.g., "hybrid" or "subject"), inducing reflections and negotiations about these categories and the object itself. Thus, social behavior toward machines is part of a mindful process of reflection, which involves negotiations concerning the nature of the machine, the user, and their relationship. Instead of being fooled by their old brains, users are being inspired by the paradox of the machine to re-think their schemes and paradigms. Hence, computers bring philosophy down to earth, by raising questions about the machine's "life" and "mind," and then by extension, making us wonder what is special about our own (Turkle, 2005, p. 2). Computational objects, poised between the world of the animate and inanimate, being at the same time a thing and a subject, alive and not alive, a physical object and an abstract idea, are experienced as both part of the self and of the external world, evoking questions about life, mind, and human identity.

Research Methodology

Turkle based her notions on her ethnographic work in the 1980s in which she studied computer cultures such as home computer owners, hackers, and artificial intelligence experts, as well as children, by living with them, participating in their lives and rituals, and interviewing them to understand things from the inside (2005, p. 25). For instance, she gave children, in groups and individually, toys—some traditional and some electronic—and observed their spontaneous reactions. She also asked questions in Piaget's style and gave them small tasks to, for example, sort pictures into piles according to whether the objects pictured were "alive" or "not alive" or asked them to draw something alive and not alive (p. 45). This starkly contrasts with the Media Equation paradigm, in which Reeves and Nass (1996) expressed no interest in users' reflection and introspection. Turkle's (2005) description of several groups of people enabled her to show the computer's second nature as a reflective medium and a philosophical provocateur (p. 279). She concluded that we need a new object relations theory. That is, a theory about our connection with objects or things, to help us understand feelings such as attachments to machines and to navigate them responsibly (p. 297).

Recent HMC Research

HMC Theorizing

When Guzman (2018) laid out the foundation for HMC as a research area within communication science, she defined HMC as the creation of meaning among humans and machines (p. 1). She noted that—following earlier work by Blumer (1969), Mead (1967), and Carey

(1989)—communication is also a means through which people learn about their world, form an understanding of Self and Other, and contribute to the shape of society. Thus, questions arise like: "What sorts of relationships emerge when technologies become communicators? How do people understand themselves as the results of their interactions with, for instance, a virtual agent? And what society is being constructed through people's communication with humans as well as machines?" (pp. 3–4). Similarly, just like Turkle (2005) conceptualized computers as philosophical provocateurs, Fortunati and Edwards (2021, p. 15) noted that the blurred boundaries between humans and current-day AI-enabled communicators raise questions such as: "What is a human being? What are our capabilities regarding thinking and doing things? How are these capabilities different from those of communicators such as robots?" And although both humans and machines may be social actors, they are not necessarily seen as the same type of social actor (Edwards & Edwards, 2022, p. 8).

Several scholars worked on conceptualizing the "betwixt and between" nature of these modern communicators. For instance, Etzrodt and Engesser (2021) conceptualized voice-based agents as "personified things," referring to Piaget's fundamental ontological object-subject classification, which they identified as a modification of the "thing" scheme, tending toward "person." Similarly, Guzman (2015, p. 252) referred to such agents as "social things" to emphasize their enhanced social nature, which at the same time is different from social beings. Gunkel (2020, p. 55) referred to Ihde's (1990) "quasi-otherness" to emphasize that some machines like Jibo inhabit a place in between the two ontological classifications "who" or "what," which he substantiated as being between "thing" and "person" in recent publications (Gunkel, 2022). Drawing on Harraway's (e.g., 1991, 2008) ideas of boundary projects and moving ontologies when humans meet other species, Suchman (2011) created the term "subject objects" for humanoid robots, to indicate the simultaneity of both categories during negotiation. In a similar vein, Krummheuer (2015, p. 185) transferred the negotiating act to embodied conversational agents by referring to Braun-Thürmann's (2002) "threshold object" ("Schwellen-Objekt" in German) to indicate that these agents are neither a human nor an artifact, emphasizing the triangulation of the agents' design, the users' interpretation, and the situation itself (p. 183).

HMC Research Methodology

The empirical studies that investigated ontological boundaries and the ontological classification of machines included qualitative interview studies (e.g., Guzman, 2019, 2020; van der Goot, 2022), surveys with open-ended questions (e.g., Edwards, 2018), content analyses of user reviews (e.g., Purington et al., 2017), and surveys that aimed to develop measures that capture the hybrid nature of machines (e.g., Etzrodt, 2022; Etzrodt & Engesser, 2021; Weidmüller, 2022).

These empirical studies showed how users negotiate the nature of machines, and the struggles this implies. For instance, users struggled with how to refer to artificial agents when constantly shifting between the pronouns "she" and "it" (Guzman, 2015; Purington et al., 2017), and the majority of participants in Etzrodt and Engesser's (2021) study classified these agents in the realm of "personified things" but they were highly uncertain about this classification. In addition, Guzman's (2020) analysis of her interviews showed that people differentiate between humans and computers based on origin of being, degree of autonomy, status as tool or tool-user, level of intelligence, emotional capabilities, and inherent flaws, but that the ontological boundaries are getting more and more blurred due to the fact that technologies increasingly emulate human-like qualities such as emotion. Relatedly—when asked to group humans, chimpanzees, and robots—thoughts about naturalness/artificiality, (non-)aliveness, (non-)resemblance to humans' embodiment, intellect, and behavior, or interactivity, and—true only for some—the difference from and the inferiority to human beings were evoked (Edwards, 2018).

Also, analyses in terms of source orientation (who or what they think they communicate with, Guzman, 2019, p. 344) revealed that users diverged in their perceptions. For voice assistants, they related to voices of the machine (i.e., the mobile device) versus voices in the machine (i.e., an agent separate from the device) (Guzman, 2019), whereas for text-based chatbots they thought they had communicated with a human being, a conversational agent (e.g., virtual assistant, robot), something software-related (e.g., algorithms) or something hardware-related (e.g., computer, machine or server) (van der Goot, 2022). Importantly, the question is raised how these conceptualizations inform humans' interactions with these artificial communicators (e.g., Edwards, 2018; Guzman, 2020).

In contrast to the Media Equation paradigm, the Media Evocation paradigm focuses on mindfulness. That is, conscious negotiation processes, whereby the findings seem to rely on the user's ability to express an in-between status, graduality, or hybridity. However, this may be limited not only by the participant's (in)ability to verbalize this (Turkle, 2005) but also by our language that does not yet provide words for machines' hybrid statuses, forcing people into the two poles person ("she/he") or thing ("it"), respectively "who" and "what." Thus, it is vital that HMC researchers keep reflecting on and conceptualizing machines' "betwixt and between" status, keep conducting observational and interview studies to gain insights in the interactions in naturalistic settings and from the users' perspectives, and aim to develop measures that provide insights in machines' hybrid nature without forcing participants to have to invent words or use words that are unnatural to them. Moreover, Etzrodt (2022) highlights the difficulties in analyzing quantitative measures that consider the ontological hybrid nature of machines, by demonstrating that the often reasonably skewed data call for the need for more innovative and robust analyses beyond simply examining central tendencies.

Conclusion and Future Directions

In this paper, we presented a distinction between two paradigms that are driving current HMC research. First, the Media Equation paradigm, which—in its seminal works—conceptualized machines as nonhuman beings that "fool" humans into mindless social responses, and that now focuses on how a variety of social cues leads to social responses through both mindless and mindful processes. The empirical studies within this paradigm mostly rely on experimental designs to test these effects. Second, the Media Evocation paradigm, which conceptualizes machines as objects that can evoke reflection or negotiation processes about, e.g., the ontological categories "who" and "what," since they are "betwixt and between" these categories, and, depending on the situation, culture, or individual inclination are sometimes more one or more the other—resulting in more or less mindful social responses. Within this paradigm—to be able to understand users' reflections and negotiations—qualitative analyses are more common.

Theoretical Implications

We envision that making this differentiation more explicit in HMC research deepens our understanding of *why* humans react socially toward machines and the consequences this has. Obviously, it is important to continue studying the impact of social cues on social responses, whereas at the same time it is vital to investigate machines in their roles as "philosophical provocateurs" and describe how users negotiate and reflect on their own identities and those of the machines they are interacting with. It is called for to study how the hybrid "betwixt and between" nature of machines makes the boundaries between humans and machines less clear.

So far, the Media Equation paradigm and its emphasis on mindless processes (e.g., Nass & Moon, 2000; Reeves & Nass, 1996) gained a lot of momentum in the HMC field. These works enabled HMC researchers to consider machines as serious social actors, although the machines were only conceptualized as *as-if* actors (i.e., they are not human). Importantly, also in the influential Media Equation publications, authors indicated that the "equation" does not apply to all social interactions with machines. Whereas these authors focused on "social attitudes and behaviors that are controlled by more primitive or automatic processes" and "[r]ules that are used frequently" (Nass & Moon, 2000, p. 99), they did acknowledge the existence of conscious, reflective processes (e.g., Nass & Moon, 2000, p. 99; Reeves & Nass, 1996, p. 9). However, since the Media Equation paradigm does not offer a theoretical framework for exploring these conscious reflections, HMC researchers can in addition turn to the work of Turkle (2005, 2007) and Langer (1992)—authors who were also mentioned by Reeves and Nass, and Nass and Moon.

Langer's (1992) views on mindlessness and mindfulness can provide a bridge to exploring the causes of social responses more holistically. She wrote that "[r]ather than relegating all social interaction to mindless behavioral scripts, I began exploring contextual factors that might shift conscious awareness from minimal structural cues to a more complete awareness of available information" (p. 290). She also emphasized situational factors alongside individual ones. Whereas mindless responses seem invariant, she pointed out that being aware of our behavior, and the factors causing it, makes us more flexible and better able to adapt to new situations (pp. 300–301)—which is very pertinent for our current-day interactions with artificial communicators. Recent work on dual processing will also help HMC researchers to study the interplay between such mindless and mindful processes (e.g., Koban & Banks, 2023; Złotowski et al., 2018).

In addition, the inclusion of the Media Evocation paradigm (using Turkle's 2005, 2007 work) will help to come to a more in-depth understanding of how machines *are* social actors, by exploring the "betwixt and between" nature of machines, thus moving on from the limited dichotomy of either "thing" or "person" (e.g., Etzrodt & Engesser, 2021; Gunkel, 2022). Recognizing and paying sufficient attention to both of the paradigms will deepen our understanding of how and in which situations Media Equation and Media Evocation processes are at play, thus gaining a more holistic understanding of our social responses to machines.

Methodological Implications

As said, the Media Equation paradigm gained a lot of momentum, and this paradigm relied almost exclusively on experimental study designs (Gambino et al., 2020). Indeed, a recent analysis of 132 HMC publications in 28 communication journals over the past decade found that quantitative studies accounted for almost half of all studies (48.5%), and that the most used type of data collection was an experiment (40.2%). Qualitative studies (12.1%) and especially mixed-method studies (only 3.8%) accounted for the smallest number of studies (Richards et al., 2022, pp. 52–53).

Following previous calls (e.g., Gambino et al., 2020; Richards et al., 2022; van der Goot, 2022), we emphasize that more qualitative and inductive studies are needed. This will contribute to our understanding of not only the human-human or human-media scripts people are using in their interactions (Gambino et al., 2020, p. 79), but also our understanding of how users negotiate the blurring boundaries between humans and machines. Importantly, it will enhance our understanding of how ontological classifications relate to humans' social responses to machines (e.g., Edwards & Edwards, 2022). For quantitative data, the hybrid nature of human-machine communication implies that more innovative, robust strategies of data analysis are required (e.g., Etzrodt et al., 2022). And finally, we would like to make an especially strong plea for more mixed-method studies. Fortunati and Edwards (2021, p. 23) concluded that the research methodologies in the HMC field increasingly integrate qualitative and mixed methods, and we underline that to enhance our understanding of both Media Equation and Media Evocation we cannot do without more mixed-method research. Specifically, the combination of experiments (in which entity perceptions are included as mediators) with observations and interviews that use think-aloud methods and open-ended questions is needed to gain in-depth insights in both the effects and the negotiations in response to machines such as voice-based agents, robots, and chatbots.

Author Biographies

Margot van der Goot (PhD, Radboud University Nijmegen, the Netherlands) is a senior assistant professor of persuasion and new media technologies at the Amsterdam School of Communication Research (ASCoR) at the University of Amsterdam. Her work particularly focuses on users' perceptions of interactions with conversational agents, and on concepts such as *source orientation*, *anthropomorphism*, and *social presence*.

https://orcid.org/0000-0001-6904-6515

Katrin Etzrodt (PhD, TU Dresden, Germany) is a research associate at the Institute of Media and Communication at the TU Dresden. Her work particularly focuses on users' definitions of the situation when communicating with voice-based assistants and other artificial agents, and on the negotiation process when objects cannot be categorized precisely.

https://orcid.org/0000-0001-6515-9985

Acknowledgments

Both authors contributed equally to the conceptualizations in this paper.

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2

Blumer, H. (1969). *Symbolic interactionism: Perspective and method.* Prentice-Hall.

Braun-Thürmann, H. (2002). *Künstliche Interaktion. Wie Technik zur Teilnehmerin sozialer Wirklichkeit wird. [Artificial interaction. How technology becomes a participant in social reality.]* Westdeutscher Verlag.

Carey, J. W. (1989). *Communication as culture: Essays on media and society.* Routledge.

Douglas, M. (1966). *Purity and danger: An analysis of concepts of pollution and taboo.* Routledge and Kegan Paul.

Edwards, A. (2018). Animals, humans, and machines: Interactive implications of ontological classification. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 29–49). Peter Lang. https://doi.org/10.3726/b14414

Edwards, A., & Edwards, C. (2022). Does the correspondence bias apply to social robots?: Dispositional and situational attributions of human versus robot behavior. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.788242

Etzrodt, K. (2022). The third party will make a difference: A study on the impact of dyadic and triadic social situations on the relationship with a voice-based personal agent. *International Journal of Human-Computer Studies*, 168. https://doi.org/10.1016/j.ijhcs.2022.102901

Etzrodt, K., Gentzel, P., Utz, S. & Engesser, S. (2022). *Human-machine communication: Introduction to the special issue. Publizistik, 67*(4), S. 439. https://doi.org/10.1007/s11616-022-00754-8

Etzrodt, K., & Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication, 2*, 57–79. https://doi.org/10.30658/hmc.2.3

Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in Human-Machine Communication. *Human-Machine Communication*, *1*, 7–28. https://doi.org/10.30658/hmc.1.1

Fortunati, L., & Edwards, A. (2021). Moving ahead with human-machine communication. *Human-Machine Communication*, *3*, 7–28. https://doi.org/10.30658/hmc.2.1

Fox, J., & Gambino, A. (2021). Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking, 24(5),* 294–299. https://doi.org/10.1089/cyber.2020.0181

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, *1*, 71–85. https://doi.org/10.30658/hmc.1.5

Gunkel, D. J. (2020). *An introduction to communication and artificial intelligence*. Wiley.

Gunkel, D. J. (2022). *Person, thing, robot: A moral and legal ontology for the 21st century and beyond.* MIT Press.

Guzman, A. L. (2015). *Imagining the voice in the machine: The ontology of digital social agents*. University of Illinois at Chicago.

Guzman, A. L. (2018). "What is human-machine communication, anyway?" In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 1–28). Peter Lang. https://doi.org/10.3726/b14399

Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. *Computers in Human Behavior*, *90*, 343–350. https://doi.org/10.1016/j.chb.2018.08.009

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, *1*, 37–54. https://doi.org/10.30658/hmc.1.3

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, *22(1)*, 70–86. https://doi.org/10.1177/1461444819858691

Haraway, D. (1991). *Simians, cyborgs, and women: The reinvention of nature*. Routledge.

Haraway, D. (2008). *When species meet*. University of Minnesota Press.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Ischen, C., Araujo, T., van Noort, G., Voorveld, H., & Smit, E. (2020). "I am here to assist you today": The role of entity, interactivity and experiential perceptions in chatbot persuasion. *Journal of Broadcasting & Electronic Media*, *64(4)*, 615–639. https://doi.org/10.1080/08838151.2020.1834297

Kim, Y., & Sundar, S. S. (2012). Anthropomorphism of computers: Is it mindful or mindless? *Computers in Human Behavior*, *28*, 241–250. https://doi.org/10.1016/j.chb.2011.09.006

Koban, K., & Banks. J. (2023). Dual-process theory in human-machine communication. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), *The SAGE Handbook of Human-Machine Communication*. SAGE.

Krummheuer, A. (2015). Technical agency in practice: The enactment of artefacts as conversation partners, actants and opponents. *PsychNology Journal*, *13*, 179–202.

Langer, E. J. (1989/2014). *Mindfulness (25th anniversary edition) (A Merloyd Lawrence Book)*. Hachette Books.

Langer, E. J. (1992). Matters of mind: Mindfulness/mindlessness in perspective. *Consciousness and Cognition, 1(3)*, 289–305. https://doi.org/10.1016/1053-8100(92)90066-J

Leichtmann, B., & Nitsch, V. (2021). Is the social desirability effect in human–robot interaction overestimated? A conceptual replication study indicates less robust effects. *International Journal of Social Robotics, 13(5),* 1013–1031. https://doi.org/10.1007/s12369-020-00688-z

Lombard, M., & Xu, K. (2021). Social responses to media technologies in the 21st century: The media are social actors paradigm. *Human-Machine Communication*, *2*, 29–55. https://doi.org/10.30658/hmc.2.2

Mead, G. H. (1967). *Mind, self, & society: From the standpoint of a social behaviorist* (Vol. 1). University of Chicago Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56(1)*, 88–103. https://doi.org/10.1111/0022-4537.00153

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 72–78. https://doi.org/10.1145/191666.191703

Nass, C., Steuer, J., Tauber, E. R., & Reeder, H. (1993). Anthropomorphism, agency, and ethopoeia: Computers as social actors. *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, 111–112. https://doi.org/10.1145/259964.260137

Powers, A., & Kiesler, S. (2006). The advisor robot: Tracing people's mental model from a robot's physical attributes. *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction,* 218–225. https://doi.org/10.1145/1121241.1121280

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2853–2859. https://doi.org/10.1145/3027063.3053246

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

Richards, R. J., Spence, P. R., & Edwards, C. C. (2022). Human-machine communication scholarship trends: An examination of research from 2011 to 2021 in communication journals. *Human-Machine Communication*, *4*, 45–62. https://doi.org/10.30658/hmc.4.3

Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Computers in Human Behavior*, *90*, 285–287. https://doi.org/10.1016/j.chb.2018.09.014

Suchman, L. (2011). Subject objects. *Feminist Theory*, *12*, 119–145. https://doi.org/10.1177/1464700111404205

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

Turkle, S. (1984/2005). *The second self: Computers and the human spirit* (*Twentieth Anniversary Edition)*. The MIT Press. https://doi.org/10.7551/mitpress/6115.001.0001

Turkle, S. (2007). *Evocative objects: Things we think with*. The MIT Press.

Turner, V. (1969). *The ritual process: Structure and antistructure.* Aldine.

van der Goot, M. J. (2022). Source orientation, anthropomorphism, and social presence in human-chatbot communication: How to proceed with these concepts. *Publizistik, 67*(4), 555–578. https://doi.org/10.1007/s11616-022-00760-w

Weidmüller, L. (2022). Human, hybrid, or machine? Exploring the trustworthiness of voice-based assistants. *Human-Machine Communication*, *4*, 85–110. https://doi.org/10.30658/hmc.4.5

Zarouali, B., Makhortykh, M., Bastian, M., & Araujo, T. (2021). Overcoming polarization with chatbot news? Investigating the impact of news content containing opposing views on agreement and credibility. *European Journal of Communication, 36*(1), 53–68. https://doi.org/10.1177/0267323120940908

Złotowski, J., Sumioka, H., Eyssel, F., Nishio, S., Bartneck, C., & Ishiguro, H. (2018). Model of dual anthropomorphism: The relationship between the media equation effect and implicit anthropomorphism. *International Journal of Social Robotics, 10(5),* 701–714. https://doi.org/10.1007/s12369-018-0476-5

HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.3

Archipelagic Human-Machine Communication: Building Bridges Amidst Cultivated Ambiguity

Marco Dehnert[1]

1 The Hugh Downs School of Human Communication, Arizona State University, Tempe, AZ, USA

Abstract

In this commentary, I call for maintaining the archipelagic character of human-machine communication (HMC). Utilizing the metaphor of the archipelago or a chain of connected islands indicates that HMC entails a variety of islands differing in shape, size, location, and proximity to one another. Rather than aiming for conceptual unity and definitional homogeneity, I call for embracing a cultivated ambiguity related to HMC key concepts. Ambiguity in the sense of allowing these concepts to be flexible enough to be explored in different contexts. Cultivated in the sense of demanding resonance across individual studies and theoretical lineages to allow for cumulative and collaborative theorizing. My hope is that HMC scholars can continue to build bridges that traverse the paradigmatic, methodological, theoretical, and technological archipelago of HMC.

Keywords: human-machine communication, communication studies, cultivated ambiguity, interdisciplinarity, resonance

Introduction

In 2018, Guzman described human-machine communication (HMC) as "the creation of meaning among humans and machines" (p. 1). Since then, and arguably before that, too, scholars from a variety of backgrounds have explored the ways in which humans interact, communicate, and relate with machinic entities such as artificial intelligence (AI), social robots, voice assistants, chatbots, and much more. As scholarship in this subfield of the

CONTACT Marco Dehnert • marco.dehnert@asu.edu • The Hugh Downs School of Human Communication • Arizona State University, P.O. Box 871205 • Tempe, AZ 85287-1205, USA

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

communication discipline is rapidly emerging, the question remains, what exactly is HMC and what sets it apart from other scholarly endeavors into the study of humans' interactions with technology? In this commentary, I make a case for viewing the subfield of HMC as archipelagic and, by considering the implications of this metaphor, call for avoiding rigid notions of so-called "proper" theory and method of HMC in favor of embracing a cultivated ambiguity in method, theory, and paradigmatic approaches to account for the diversity of HMC phenomena and scholarship.

Traversing the Archipelago

In my use of the metaphor of the archipelago, I draw on Simmons and Brisini's (2020) similar use of the metaphor to describe the subfield of performance studies in communication. Rather than constituting a coherent landmass or a distinct separation into two dialectical shores, an archipelago is a chain or group of connected islands situated in close proximity in a body of water. Utilizing the metaphor of the archipelago to describe HMC indicates that our subfield is made up not of a coherent subject or a cohesive body of literature, but rather entails a variety of islands differing in shape, size, location, and proximity to one another. In ways comparable to performance studies in communication, HMC is thus made up of "Different subjectivities, different topoi, different practices and aesthetic traditions, different academic histories, different texts, and different cultures" (Simmons & Brisini, 2020, p. 2) in addition to a multitude of methodological practices and theoretical convictions. With HMC scholars publishing across a variety of scholarly outlets in communication and engineering, AI studies, big data studies, human-robot and human-computer interaction, and more, and with using an increasing variety of methodological approaches to the study of HMC, I find *archipelagic* an appropriate descriptor and valuable assessment for the current landscape of HMC.

Importantly, even though individual islands may be bigger in size compared to others, an archipelago refuses any claim toward a central or main island and shifts the focus more so to the connections among individual islands into a larger entanglement and the various flows of water between and betwixt them. That is, historically speaking, the majority of work on human-computer interaction (and its various disciplinary siblings) has been conducted from post-positivistic perspectives using primarily quantitative and experimental methods. As the methodological and paradigmatic landscapes continue to diversify, scholars have generated insightful scholarship in HMC from qualitative (e.g., Guzman, 2020; Rainear et al., 2021), rhetorical (e.g., Coleman, 2021; Fritz, 2018), autoethnographic (e.g., Chun, 2019), critical (e.g., Davis & Stanovsek, 2021; Dehnert & Leach, 2021; J. Liu, 2021; Moran, 2021; Rambukkana, 2021), posthumanist and new materialist (e.g., Betlemidze, 2022; Dehnert, 2022; Kubes, 2019; Rambukkana, 2021), and other approaches. The formation of these newer islands in the HMC-archipelago complements already existing approaches and allows for conceptualizing HMC from different angles and new perspectives.

HMC finds itself at a unique disciplinary juncture where scholars have become able to generate systematic reviews of this increasingly diverse and growing field. That is, Richards et al. (2022) examine the scholarship trends of HMC research from 2011 to 2021 in communication journals, which is complemented by Makady and Liu's (2022) review of publication trends across top-ranking journals in roughly the same time frame. Whereas Richards et al.'s (2022) analysis focuses primarily on communication journals, Makady and Liu's (2022) review includes journals with different disciplinary affiliations as well, thereby taking into account how many HMC scholars publish beyond communication outlets in fields such as human-computer interaction, human-robot interaction, AI studies, or psychology. Makady and Liu (2022) and Richards et al. (2022) concur with yet a third recent systematic review, F. Liu et al.'s (2022), when they all observe the highly interdisciplinary character of HMC as a subfield. All three reviews end with a call for more diversified approaches, be it related to the study of specific technologies, utilizing variegated methods, or working toward unique HMC theory. This also means recognizing the specific methodological challenges that come with researching humans and machines communicating (Greussing et al., 2022). As F. Liu et al. (2022) argue, "a complete understanding of HMC is only possible when multiple methods are used to validate results, produce new knowledge, and further define the scope of the field" (p. 26). Even though this field of HMC, as Richards et al. (2022) conclude, may benefit from a balance of methods, samples, and approaches, it already achieves networked collaboration and cross-, trans-, and interdisciplinary conversations: "HMC has defied R. T. Craig's (1999) prediction of drastically diverse fields not being able to work together" (Richards et al., 2022, p. 54). Based on these impressive reviews of the young and dynamic field of HMC, I return to the island-metaphor below and consider how to understand HMC as archipelagic.

Embracing Cultivated Ambiguity in HMC Research and Scholarship

As indicated above, Guzman (2018) originally formulated HMC as "the creation of meaning among humans and machines" (p. 1). Alongside others within the communication discipline and beyond, HMC scholars have generated insightful scholarship that investigates these four components in depth, asking about the nature of the human, the machine, how to conceptualize meaning, and how meaning is created in interactive and communicative processes between humans and machines. With ongoing difficulties in clearly conceptualizing and defining emerging technologies such as AI (Gunkel, 2020), big data (Croucher, 2022; Parks, 2014), robots (Fortunati & Edwards, 2021), and others, it will be interesting to see how HMC scholars approach the study of meaning-making processes in these contexts. Rather than calling for conceptual homogeneity by laboring toward clear definitions of these machinic constructs—which could certainly be one goal of disciplinary endeavors aimed at maintaining legitimacy—I offer a plea to cultivate an ambiguity as it relates to our conceptualization of these key components of HMC. Let me explain.

By cultivated ambiguity I do *not* mean a complete avoidance of articulating conceptual, theoretical, and operational definitions of the things that we study when we "do" HMC—humans, machines, and their interactions as they create meaning. In fact, conceptual work such as Shaikh's (2023) definitional framework for intelligent assistants or Mooshammer's (2022) proposed terminology for automation in journalism allow for clearly articulating our terms, help us explicate what technologies we study, and make comparisons across studies possible in the first place. Nor is my goal to call for scholars to intentionally confuse our use and understanding of those fundamental terms and concepts. Rather, my hope is that, as HMC continues to unfold and as technology continues to advance at a rapid speed, we as HMC scholars remain open to different definitions of these key components of HMC instead of demanding definitional consensus among different paradigmatic convictions, methodological approaches, and contexts. What I am gesturing at is a sense of curated interpretive flexibility that allows for conceptual resonance, not homogeneity, across the various contexts in which we study HMC.

Not only are we exploring the interplay of humans and machines across all contexts of communication—be it interpersonal (e.g., Ryland, 2021; Spence et al., 2014), organizational (e.g., Piercy & Gist-Mackey, 2021; Spence et al., 2018), instructional (e.g., Edwards & Edwards, 2017), or mass-mediated (e.g., Lewis et al., 2019), to name but a few—we are also tasked with apprehending a multitude of technical features that make up what we capture under the umbrella of "machine"—be it artificial intelligence (an ambiguous term in itself consisting of some form of algorithms, machine learning, deep learning, natural language processing, and more; cf. Gunkel, 2020), voice assistants (Etzrodt & Engesser, 2021; Moran, 2021), chatbots (Croes & Antheunis, 2021; van der Goot, 2022), social robots (Chun, 2019; Fritz, 2018; J. Liu, 2021), or more. Taking these technical differences, the rapid speed at which they are advancing, as well as the variegated contexts in which humans interact with machines into consideration, alongside the multitude of methodological, theoretical, paradigmatic, and political approaches in HMC, I find it both challenging and radically limiting for the larger HMC project if we were to call for conceptual homogeneity and definitional unity.

In fact, once we start calling for rigid definitions of the key concepts and technologies we study, we foreclose potentiality, theory-building, and innovation in our field. Removing all conceptual "wiggle room" by demanding that our definitions of human, machine, and human-machine communication remain similar across all contexts would result in our young HMC project idling, turning into a stalling field that becomes outpaced and outdated as technology advances and our human-machine experiences become ever more interrelated. While a high degree of conceptual unity might result in high internal validity across studies, our field's external validity would increasingly shrink with the lack of alternative perspectives, theories, and approaches. The result would be a field that has become out of touch with its subjects and objects of study, losing its critical edge. And finally, with unity in definitions comes unity in approaches, with which comes unity in scholars and scholarship represented. And with such unity comes the necessary exclusion of perspectives, approaches, and scholars who think and theorize otherwise. As debates related to canonization in the discipline of communication and its subfields (e.g., rhetoric; Baugh-Harris & Wanzer-Serrano, 2018) have made abundantly clear, conceptually unified fields bring not only epistemological flaws, but more consequentially political violence (Calvente et al., 2020). And this is not only represented on citational levels, but has much deeper implications (Freelon et al., 2023).

But where does this call for conceptual unity or coherence come from? Those of us who are familiar with the disciplinary origins and character of communication studies are aware of the many ongoing debates related to what makes the communication field a field. With oft-cited work such as Craig's (1999) hallmark essay and others as prominent examples, commentators and scholars have long expressed the values of a more coherent field. The question about the identity of the field of communication and, by extension, of HMC, is however a complicated one. Pushing against the desire for a coherent field, McCann et al. (2020) poignantly write: "Our identity as a discipline lies in the very truth we wish to jettison: our field's theoretical and methodological plurality, promiscuity, and fragmentation" (p. 249). Operating within a fragmented and promiscuous field, then, we as communication and HMC scholars may consider alternatives to striving for a coherence and unity that is beyond our reach, especially given the unique qualities of the field of HMC as I discuss later.

Hence, the plea I put forth in this commentary is one that calls for embracing a cultivated ambiguity as it relates to the key components that make up HMC. *Ambiguity* in the sense of allowing these concepts to be flexible enough to be explored in different contexts and from different angles, thereby avoiding the foreclosure of non-post-positivistic and nonquantitative approaches to the study of HMC. *Cultivated* in the sense of demanding a certain sense of resonance across individual studies and theoretical lineages within the larger frame of HMC to allow for cumulative and collaborative theorizing, where future work can build on and extend previous research. The task is to engage in this elaborate dance between cultivated ambiguity and conceptual resonance of concepts within and across individual studies, theoretical perspectives, and paradigmatic and methodological approaches to the study of HMC.

Building Bridges Across Islands: A Plea for an Enmeshed Archipelagic HMC

Rather than heralding the importance of particularly prominent islands in the HMC-archipelago, then, or rather than focusing on prevalent formations across individual islands and their surrounding bodies of water, this embracing of cultivated ambiguity calls for building bridges across (perceived) divides—connecting islands in an increasingly entangled network or enmeshment of trans-methodological, trans-theoretical, and trans-paradigmatic conversations. The field of HMC is particularly well-suited for archipelagic bridge-building. Although its more formal characterizations can be dated to 2018 with Guzman's edited collection, to 2019 with the creation of the HMC Interest Group at the International Communication Association, and to 2020 with this journal's first issue, HMC scholarship and scholars can be traced back much earlier and found in fields such as science and technology studies, sociology of communication, human-machine relations, or human-robot interaction, among others. Functioning as an interdisciplinary umbrella framework encapsulating approaches within and beyond communication studies (Guzman, 2018), HMC consists of many islands that approach the study of human-machine interaction by centering communication, its context, and its impact on the sociotechnical subjects in relation. At this juncture of more formally and more consciously articulating the character of the field of HMC, being aware of the risks that come with disciplinary coherence is crucial for not repeating what we have seen in other subfields of communication, such as rhetoric. Archipelagic bridge-building and cultivated ambiguity can serve as powerful metaphorical heuristics that generatively question a desire for coherence.

Outlets like this subfield-specific journal *Human-Machine Communication* provide an excellent space for such archipelagic conversation (the journal has published an impressive variety of scholarship focused on HMC in its first volumes; cf. Fortunati & Edwards, 2021), and my hope is that this impetus will resonate across other outlets as well. In so doing, HMC can continue to pose demanding questions to the communication discipline. For example: What does it mean to be human or machine in communicative encounters? What are the boundaries of what constitutes communication? What—or who—constitutes a necessary condition for the creation of meaning in HMC? These questions can be asked while remaining on top of technological developments and how they impact and implicate the human communicative condition.

Examples of such bridge-building and island-traversing projects include historiographical work such as Bory et al.'s (2021), which allows for contextualizing canonical histories of technological developments related to AI, machines, and robots more firmly from a communication perspective. In doing so, they span interdisciplinary conversations yet articulate the contributions of a communication and media studies perspective (cf. Gunkel, 2020). Natale and Guzman's (2022) recent special issue calls for reclaiming the human in machine cultures across a variety of use-cases and contexts, and Sundar and Lee's (2022) recent special issue calls for rethinking communication in the era of AI. Etzrodt et al.'s (2022) recent special issue maps the landscape (dare I say archipelago) of HMC research, surveys its trends, and discusses future possibilities and challenges for our young field.

Bridge-building amidst cultivated ambiguity means recognizing the value of collaboration—across stages of careers, geographical and cultural distances, technological contexts, methodological approaches, and theoretical lenses. It means bringing scholarship and scholars with variegated disciplinary affiliations in conversation with each other. It means recognizing the value of interdisciplinary publications and publications outside communication journals for tenure and promotion cases. And it means embracing different, sometimes even opposing, perspectives on the technologies we study, theories we develop, and methods we utilize.

As an archipelago, then, rather than a coherent landmass or set of dialectical shores, HMC provides ample space for embracing diversified approaches to the study of human-machine interaction and can foster the growth of unique, innovative, and insightful research and scholarship. Cultivated ambiguity in this sense then celebrates interpretive flexibility as we articulate and rearticulate HMC across its diverse aspects coupled with the need to hold ourselves accountable as we express connections and disconnections between various traditions, approaches, and theories within HMC. The practice and art of cultivating this archipelago emerges through ongoing reflexive praxis aimed at fostering resonance and reverberation rather than uniformity. A consequence of such an archipelagic conceptualization of HMC is the recognition that this—*our*—subfield's boundaries are open to (re)formation and (re)connection—across islands within this archipelago as well as beyond its perceived coherence into communication studies and other adjacent disciplines. The goal is to maintain this archipelagic spirit in the subfield of HMC.

Author Biography

Marco Dehnert (MA, Arizona State University) is a doctoral candidate in the Hugh Downs School of Human Communication at Arizona State University. He is a multi-method scholar who studies human-machine communication, artificial intelligence, and the social impact of communication technologies. Marco's work has appeared in *Human-Machine Communication*, *Human Communication Research*, and the *International Journal of Social Robotics*, among others. He is also part of the Relationships and Technology Lab at ASU.

https://orcid.org/0000-0002-7456-0743

References

Baugh-Harris, S., & Wanzer-Serrano, D. (2018). Against canon: Engaging the imperative of race in rhetoric. *Communication and Critical/Cultural Studies, 15*(4), 337–342. https://doi.org/10.1080/14791420.2018.1526386

Betlemidze, M. (2022). Traversing anthropocentric horizons with *Her:* Trans-corporeal surrogacy, enchantment, and disenchantment in human-machine assemblage. *Journal of Communication Inquiry, 46*(2), 206–224. https://doi.org/10.1177/01968599211041107

Bory, P., Natale, S., & Trudel, D. (2021). Artificial intelligence: Reframing thinking machines within the history of media and communication. In G. Balbi, N. Ribeiro, V. Schafer, & C. Schwarzenegger (Eds.), *Digital roots: Historicizing media and communication concepts of the digital age* (pp. 95–114). De Gruyter. https://doi.org/10.1515/9783110740202-006

Calvente, L. B. Y., Calafell, B. M., & Chávez, K. R. (2020). Here is something you can't understand: The suffocating whiteness of communication studies. *Communication and Critical/Cultural Studies, 17*(2), 202–209. https://doi.org/10.1080/14791420.2020.1770823

Chun, B. (2019). Doing autoethnography of social robots: Ethnographic reflexivity in HRI. *Paladyn, Journal of Behavioral Robotics, 10*(1), 228–236. https://doi.org/10.1515/pjbr-2019-0019

Coleman, M. C. (2021). Leveraging the rhetorical energies of machines: COVID-19, misinformation, and persuasive labor. *Human-Machine Communication, 3*, 11–26. https://doi.org/10.30658/hmc.3.2

Craig, R. T. (1999). Communication theory as a field. *Communication Theory, 9*(2), 119–161. https://doi.org/10.1111/j.1468-2885.1999.tb00355.x

Croes, E. A. J., & Antheunis, M. L. (2021). Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot. *Journal of Social and Personal Relationships, 38*(1), 279–300. https://doi.org/10.1177/0265407520959463

Croucher, S. M. (2022). *The current state of big data in communication studies* [Call for papers for a themed issue in *Review of Communication*]. https://web.archive.org/web/20220316204205/https://think.taylorandfrancis.com/special_issues/big-data-communication-studies/?utm_source=TFO&utm_medium=cms&utm_campaign=JPG15743

Davis, D. Z., & Stanovsek, S. (2021). The machine as an extension of the body: When identity, immersion, and interactive design serve as both resource and limitation for the disabled. *Human-Machine Communication, 2*, 121–135. https://doi.org/10.30658/hmc.2.6

Dehnert, M. (2022). Toward a critical posthumanism for social robotics. *International Journal of Social Robotics, 14*(9), 2019–2027. https://doi.org/10.1007/s12369-022-00930-w

Dehnert, M., & Leach, R. B. (2021). Becoming human? Ableism and control in *Detroit: Become Human* and the implications for human-machine communication. *Human-Machine Communication, 2*, 137–152. https://doi.org/10.30658/hmc.2.7

Edwards, A., & Edwards, C. (2017). The machines are coming: Future directions in instructional communication research. *Communication Education, 66*(4), 487–488. https://doi.org/10.1080/03634523.2017.1349915

Etzrodt, K., & Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication, 2*, 57–79. https://doi.org/10.30658/hmc.2.3

Etzrodt, K., Gentzel, P., Utz, S., & Engesser, S. (2022). Human-machine communication: Introduction to the special issue. *Publizstik, 67*(4), 439–448. https://doi.org/10.1007/s11616-022-00754-8

Fortunati, L., & Edwards, A. (2021). Moving ahead with human-machine communication. *Human-Machine Communication, 2*, 7–28. https://doi.org/10.30658/hmc.2.1

Freelon, D., Pruden, M. L., Eddy, K. A., & Kuo, R. (2023). Inequities of race, place, and gender among the communication citation elite, 2000-2019. *Journal of Communication*. https://doi.org/10.1093/joc/jqad002

Fritz, L. M. (2018). Child or product? The rhetoric of social robots. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 67–82). Peter Lang.

Greussing, E., Gaiser, F., Klein, S. H., Straßmann, C., Ischen, C., Eimler, S., Frehmann, K., Gieselmann, M., Knorr, C., Henestrosa, A. L., Räder, A., & Utz, S. (2022). Researching interactions between humans and machines: Methodological challenges. *Publizistik, 67*(4), 531–554. https://doi.org/10.1007/s11616-022-00759-3

Gunkel, D. J. (2020). *An introduction to communication and artificial intelligence*. Polity.

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

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication, 1*, 37–54. https://doi.org/10.30658/hmc.1.3

Kubes, T. (2019). New materialist perspectives on sex robots: A feminist dystopia/utopia? *Social Sciences, 8*(8), 224. https://doi.org/10.3390/socsci8080224

Lewis, S. C., Guzman, A. L., & Schmidt, T. R. (2019). Automation, journalism, and human-machine communication: Rethinking roles and relationships of humans and machines in news. *Digital Journalism, 7*(4), 409–427. https://doi.org/10.1080/21670811.2019.1577147

Liu, F., Makady, H., & Xu, K. (2022, May). *Mapping the landscape of human-machine communication research: A systematic review of empirical research from 2010 to 2021*. Paper presented at the 72nd annual meeting of the International Communication Association, Paris, France.

Liu, J. (2021). Social robots as the bride? Understanding the construction of gender in a Japanese social robot product. *Human-Machine Communication, 2*, 105–120. https://doi.org/10.30658/hmc.2.5

Makady, H., & Liu, F. (2022). The status of human-machine communication research: A decade of publication trends across top-ranking journals. In M. Kurosu (Ed.), *Human-computer interaction: Theoretical approaches and design methods. HCII 2022. Lecture notes in computer science* (pp. 83–103). Springer. https://doi.org/10.1007/978-3-031-05311-5_6

McCann, B. J., Mack, A. N., & Self, R. (2020). Communication's quest for whiteness: The racial politics of disciplinary legitimacy. *Communication and Critical/Cultural Studies, 17*(2), 243–252. https://doi.org/10.1080/14791420.2020.1770822

Mooshammer, S. (2022). There are (almost) no robots in journalism: An attempt at a differentiated classification and terminology of automation in journalism on the base of the concept of distributed and gradualised action. *Publizistik, 67*(4), 487–515. https://doi.org/10.1007/s11616-022-00757-5

Moran, T. C. (2021). Racial technological bias and the white, feminine voice of AI VAs. *Communication and Critical/Cultural Studies, 18*(1), 19–36. https://doi.org/10.1080/14791420.2020.1820059

Natale, S., & Guzman, A. L. (2022). Reclaiming the human in machine cultures: Introduction. *Media, Culture & Society, 44*(4), 627–637. https://doi.org/10.1177/01634437221099614

Parks, M. R. (2014). Big data in communication research: Its contents and discontents. *Journal of Communication, 64*(2), 355–360. https://doi.org/10.1111/jcom.12090

Piercy, C. W., & Gist-Mackey, A. N. (2021). Automation anxieties: Perceptions about technological automation and the future of pharmacy work. *Human-Machine Communication, 2*, 191–208. https://doi.org/10.30658/hmc.2.10

Rainear, A. M., Jin, X., Edwards, A., Edwards, C., & Spence, P. R. (2021). A robot, meteorologist, and amateur forecaster walk into a bar: Examining qualitative responses to a weather forecast delivered via social robot. *Communication Studies, 72*(6), 1129–1145. https://doi.org/10.1080/10510974.2021.2011361

Rambukkana, N. (Ed.). (2021). *Intersectional automations: Robotics, AI, algorithms, and equity*. Lexington Books.

Richards, R. J., Spence, P. R., & Edwards, C. C. (2022). Human-machine communication scholarship trends: An examination of research from 2011 to 2021 in communication journals. *Human-Machine Communication, 4,* 45–65. https://doi.org/10.30658/hmc.4.3

Ryland, H. (2021). It's friendship, Jim, but not as we know it: A degrees-of-friendship view of human–robot friendships. *Minds & Machines, 31*, 377–393. https://doi.org/10.1007/s11023-021-09560-z

Shaikh, S. J. (2023). Artificially intelligent, interactive, and assistive machines: A definitional framework for intelligent assistants. *International Journal of Human-Computer Interaction, 39*(4), 776–789. https://doi.org/10.1080/10447318.2022.2049133

Simmons, J., & Brisini, T. (2020). Performance studies in communication. *Text and Performance Quarterly, 40*(1), 1–48. https://doi.org/10.1080/10462937.2020.1725726

Spence, P. R., Westerman, D., Edwards, C., & Edwards, A. (2014). Welcoming our robot overlords: Initial expectations about interaction with a robot. *Communication Research Reports, 31*(3), 272–280. https://doi.org/10.1080/08824096.2014.924337

Spence, P. R., Westerman, D., & Lin, X. (2018). A robot will take your job. How does that make you feel? Examining perceptions of robots in the workplace. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 185–200). Peter Lang.

Sundar, S. S., & Lee, E.-J. (2022). Rethinking communication in the era of artificial intelligence. *Human Communication Research, 48*(3), 379–385. https://doi.org/10.1093/hcr/hqac014

van der Goot, M. J. (2022). Source orientation, anthropomorphism, and social presence in human-chatbot communication: How to proceed with these concepts. *Publizistik, 67*(4), 555–578. https://doi.org/10.1007/s11616-022-00760-w

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.4

ChatGPT, LaMDA, and the Hype Around Communicative AI: The Automation of Communication as a Field of Research in Media and Communication Studies

Andreas Hepp[1], Wiebke Loosen[2], Stephan Dreyer[2], Juliane Jarke[3], Sigrid Kannengießer[4], Christian Katzenbach[1], Rainer Malaka[5], Michaela Pfadenhauer[6], Cornelius Puschmann[1], and Wolfgang Schulz[2]

1 ZeMKI, Centre for Media, Communication and Information Research, University of Bremen, Germany
2 Leibniz-Institute for Media Research | Hans-Bredow-Institut Hamburg, Germany
3 BANDAS-Center & Department of Sociology, University of Graz, Austria
4 Institute for Communication Studies, University of Münster, Germany
5 TZI, Center for Computing Technologies, University of Bremen, Germany
6 Institute for Sociology, University of Vienna, Austria

Abstract

The aim of this article is to more precisely define the field of research on the automation of communication, which is still only vaguely discernible. The central thesis argues that to be able to fully grasp the transformation of the media environment associated with the automation of communication, our view must be broadened from a preoccupation with direct interactions between humans and machines to societal communication. This more widely targeted question asks how the dynamics of societal communication change when communicative artificial intelligence—in short: communicative AI—is integrated into aspects of societal communication. To this end, we recommend an approach that follows the tradition of figurational sociology.

Keywords: automation of communication, artificial intelligence, communicative AI, algorithms, agency, communication, figuration

CONTACT Andreas Hepp • ahepp@uni-bremen.de • ZeMKI, Centre for Media, Communication and Information Research • University of Bremen • Linzerstr. 4 • D-28359 Bremen, Germany

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

Current media coverage surrounding ChatGPT, LaMDA, and Luminous has brought questions about the automation of communication into the mainstream: Artificially intelligent media are no longer merely mediating instances of communication, but are themselves becoming communicative participants. This has generated broad public discussion about these systems and the challenges they bring to fields such as education, public discourse, and journalistic production.[1] In light of this intensifying discussion, researchers who have been working on the topic for a longer time warn against blindly embracing the hype.[2]

As media and communications researchers, we don't want to ignore these warnings and want to avoid getting caught up by the hyperbole. Nevertheless, communication technologies such as ChatGPT, LaMDA, and Luminous need to be taken seriously as they genuinely represent a new step in the automation of communication—a process that is nevertheless persistent and opens up a great deal of further discussion. The role played by bots and algorithmic personalization on social media platforms in the spread of fake news and hate speech, for example, have inspired fervent academic discussion (i.e., Lazer et al., 2018). Systems such as Amazon Alexa, Google Assistant, Microsoft Cortana, or Apple Siri have existed for nearly a decade forcing us to question our thinking about human communication and agency (i.e., Guzman, 2015). Questions of automation have been addressed further in discussions about news production (i.e., Thurman et al., 2019), surveillance capitalism (i.e., Zuboff, 2019) and data colonialism (i.e., Couldry & Mejías, 2019).

In principle, the automation of communication has a much longer history than recent public discussions might imply and can affect all areas of social life. However, it is particularly important where *societal* communication is concerned, as can be illustrated in the example of journalism. Here, the automation of communication plays a dual role: internally, for example, when journalistic working practices change as a result of the automated production and distribution of content (Carlson, 2018; Diakopoulos, 2019), and externally, when content created in this way becomes part of the public discussion (Graefe & Bohlken, 2020; Volcic & Andrejevic, 2023).

These examples indicate that automated communication systems have become part of our media environment and are thereby appropriated in specific ways in various societal domains, such as public discourse, journalism, politics, and education. This development poses considerable challenges (Fortunati & Edwards, 2020): empirically, in terms of how automated communication can be researched, and theoretically, in that the fundamental concepts of *agency*, *media*, and *communication* are dramatically altered.

With this article we want to define the automation of communication as a research area in more detail. Our main thesis is that if we are ever going to comprehensively deal with the transformation of our media environment associated with the automation of

1. This is exemplified by a simple GoogleTrends analysis, which shows increasing interest in ChatGPT worldwide from November 27, 2022, with a peak on February 12, 2023. Retrieved March 10, 2023, from https://trends.google.de/trends/explore?q=ChatGPT

2. As an example, among others, reference can be made in this regard to a discussion between Emily M. Bender and Casey Fiesler. Retrieved March 10, 2023, from https://web.archive.org/web/20230303074525/https://www.radicalai.org/chatgpt-limitations

communication, we have to address our investigation more broadly from investigating the *direct interaction of humans and machines* to *societal communication*. In broadening our view, we are compelled to ask how the dynamics of societal communication change when ChatGPT, LaMDA, Luminous, and comparable technologies become an integral part of it.

To support this reasoning, we first take a closer look at the automation of communication as a phenomenon. Against this background, we engage with the notion of *communicative AI*, which we believe can operate as a "sensitizing concept" (Blumer, 1954, p. 7) that directs us to the true breadth and depth of the phenomenon. Subsequently, we show how a figurational approach can be used to analyze automated automation as part of societal communication and connect the discussion to already existing "definitive concepts" (Blumer, 1954, p. 7) familiar to media and communication studies.

The Automation of Communication: An Emerging Subject in Media and Communication Studies

It would certainly be a misconception to assume that the automation of communication is only a subject of the most recent media and communications research. If we look historically at the emergence of today's digital systems of automated communication, we can see that they are closely interconnected with cybernetics (Turner, 2006), and identify links to media and communication theory from as far back as the 1970s. Cybernetics has always addressed questions of (communicative) automation, albeit through a primarily technical lens (i.e., Bibby et al., 1975). At the same time a rapprochement between cybernetics and the social sciences took place only to a limited extent. One of the reasons for this restraint was that the mathematical theory of communication present in the early cybernetic discussion (Baecker, 1997, p. 11), stood in contrast to social science's interests that have tended to focus on meaning and understanding.[3] This also applies to the "post-discipline" (Waisbord, 2019, p. 1) of media and communication studies[4]: Even in James R. Beniger's *The Control Revolution*, the automation of communication remained a rather marginal topic (1986, pp. 304–307). This was contrasted by research in informatics, where the potential of automating communications was an important research topic very early on, inspired in large part by *The Computer as Communications Device,* a 1968 article by J. C. R. Licklider and Robert W. Taylor. This discussion, for example, about the Turing Test or Weizenbaum's (1966) ELIZA, took place largely outside the purview of media and communication studies (i.e., Searle, 1980) and was only really addressed as a historical discussion after systems of automated communication became a more widespread phenomenon (Natale, 2021b). There are very few exceptions (Gunkel, 2012).

In a coarse simplification—sometimes necessary in the context of reconstruction—we can describe media and communication studies' increasing interest in questions of

3. This is exemplified by the analysis of the dominant communication theorists until the end of the 1980s (Beniger, 1990).

4. Silvio Waisbord uses the term "post-disciplinary" to summarize that for media and communication studies "disciplinary boundaries are fluid" and that it is an "intellectually open enterprise rather than a traditional endeavor interested in defining and patrolling epistemological boundaries" (131) (2019, pp. 127, 131); see also Livingstone, 2009; McQuail & Deuze, 2020.

automated communication as having taken place along three steps toward addressing digital communication. Temporally speaking, there are various overlaps between them, although their distinction makes sense in that they each stand for different discursive contexts in scholarly thinking about automation in relation to communication.

In the *first stage*, media and communication studies turned to digital communication by asking how communication itself and social relationships change when media become digital. The dominant concept in media and communication studies became that of "computer-mediated communication" (CMC) (Chesebro & Bonsall, 1989; Jones, 1998) and scientific interests turned to person-computer interaction (Cathcart & Gumpert, 1985; Morris & Ogan, 1996) as well as the growth of online relationships and communities (Baym, 1994; Wellman et al., 1996). This research on transforming communications was related to more general discussions about an emerging information society (see, among others, Castells, 2000; Mattelart, 2003). Later, media and communication studies research into digital communications turned to broader questions such as the "mediatization of society" (Hepp & Krotz, 2014; Hjarvard, 2013; Lundby, 2014). In all these cases, however, the automation of communication remained a marginal topic, addressed by only a small number of scientists or those working at the fringes of the discipline (i.e., Steels & Kaplan, 2000).

In the *second stage*, questions of digital data and their (societal) contexts of use and exploitation came to the fore—parallel to the fact that technology companies and state actors increasingly discovered the potential of digital data as a commodity or a resource (Zuboff, 2019). The core of the discussion was, at first, a critical engagement with *big data* as an economic, social, and cultural resource (Andrejevic, 2014; Crawford et al., 2014; Gitelman, 2013), which then led to a critique of the progressing datafication of society (Dencik & Kaun, 2020; Flensburg & Lomborg, 2021; van Dijck, 2014). Here, there was also a stronger rapprochement between media and communication studies and science and technology studies, where, for example, expert systems and artificial intelligence had already been closely studied for much longer (i.e., Star & Ruhleder, 1996; Suchman, 1987). This led to, among other things, so-called critical data studies that sat at the intersection of media and communication studies, sociology, and science and technology studies (Burns et al., 2019; Dalton & Thatcher, 2014; Hepp et al., 2022; Iliadis & Russo, 2016; Kitchin, 2014). In contexts like these, discussions have focused on the influence of datafied "platforms" (van Dijck et al., 2018), the need for their "regulation" (Hofmann et al., 2017), "surveillance capitalism" (Zuboff, 2019), "deep mediatization" (Hepp, 2020b), and "data colonialism" (Couldry & Mejías, 2019). Questions of automation have always played and continue to play a role in this discussion about datafication—but less in the sense of automating *communication* than in the sense of automating *data processing*.

In the *third stage* of research on digital communication this turn takes place toward the *forms* of communicative automation. As mentioned above, there were early precursors to this discussion (Gunkel, 2012; for an overview: Richards et al., 2022); however, the foundation of journals such as *Human-Machine Communication* (Fortunati & Edwards, 2020) or a corresponding interest group in the International Communication Association were exemplary for increasing the discursive momentum. A broad discussion took place to clarify the field of human-machine communication (HMC), as well as an institutionalization of the research landscape (Fortunati & Edwards, 2021; Guzman, 2018; Guzman et al., 2023). Nevertheless, it is important to keep in mind that the preoccupation with the automation of communication in media and communication studies goes beyond the institutionalizing power of HMC, and continues to address questions around topics such as "robot journalism" (Carlson, 2015), "social bots" (Gehl & Bakardjieva, 2016), the "automation of communicative labor" (J. Reeves, 2016), "algorithmic content moderation" (Gorwa et al., 2020) or "automated media" (Andrejevic, 2020; Napoli, 2014).

In a sense, one can say that there are not only genealogical interrelations between the three stages of engagement with the automation of communication in media and communication studies, but that this refers to a general characteristic of digital communication: If one understands algorithms for their ability "to act when triggered without any regular human intervention or oversight" (Gillespie, 2014, p. 170), automation—generally understood as the machine-autonomous achievement of specific goals for action—has been a key aspect of software-based media from the beginning. Digitization, datafication, and algorithmization represent both the conditions of possibility and the need for automatic communication processes. However, what then is technical in automation can vary considerably, ranging from simple scripts with determinate steps (i.e., linear algorithms in informatics terms), on which many social bots are based (cf. Veale & Cook, 2018), to complex technical machine learning systems (cf. Heuer et al., 2021).

The crucial point is that we are dealing with the automation of *communication* and not, for example, with forms of automation such as product manufacturing processes where robotic systems build things. The automation of communication is based on digital traces as inherent byproducts of datafication. These have a *materiality of their own* that is far more opaque than that of automation by locally placed material-machine systems such as manufacturing robots (Burrell, 2016). This has significant consequences for various forms of automated communication processes (Esposito, 2017, p. 251): For all their heterogeneity—for example, in health care, justice, politics, journalism, everyday practice, science, the public sector, or education—it is a materiality that refers to the globalized digital infrastructures of today's automated communication systems (Crawford, 2021). Accordingly, the three stages do not simply mean that the last one represents increasing hype or interest, but that a broad view of the automation of communication seems all the more necessary.

Broadening the Perspective: Moving From the Individual to the Societal

Initially, and in the trajectory of computer-mediated communication, media and communication studies approached the phenomenon of automated communication mainly from the perspective of the individual (i.e., the question of how individuals deal with automated systems, what agency they attribute to them, or what form of agency can be theoretically distinguished from them). This can be illustrated by publications from the 2010s that were particularly influential to the discussion: Robert W. Gehl and Maria Bakardjieva, for example, develop the perspective in their essay on social bots when they described that they are "intended to present a self, to pose as an alter-ego, as a subject with personal biography, stock of knowledge, emotions and body, as a social counterpart, as someone like me, the user, with whom I could build a social relationship" (2016, p. 2). In the same period, Andrea Guzman defined the field of human-machine-communication more intently as "the creation of meaning between human and machine" (2018, p. 3).

Looking at these texts now, they seem particularly concerned with direct interaction between humans and machines, as well as with the agency that automated systems may or may not have or that is attributed to them. This is also apparent in more media-psychology-oriented approaches such as CASA research ("Computers-Are-Social-Actors"). At its core, the CASA paradigm holds that the moment computers or other technical systems look, communicate, or act like a person, people respond to them as if they were "real" people (Lee & Nass, 2010; Nass et al., 2006). The CASA approach can be traced to Byron Reeves and Clifford Nass' text (1996), in which they addressed the "media equation"; that is, the tendency of users to put new media on a par with natural persons and places. CASA research has led to important findings; for example, on the perception of the communication qualities of automated systems (Edwards et al., 2014), on the relationship norms of humans toward Twitter bots (Li & Li, 2014), or on the anthropomorphism of smartphones (Wang, 2017). However, when it comes to expanding CASA research, the discussion is less focused on going beyond the individual-machine relationship and more on how we appropriately frame it: The argument is that if a person appropriates new systems of automated communication today (for example, an Artificial Companion), he or she will apply not only scripts that are familiar from their interactions with humans, but also those from interaction with machines (Gambino et al., 2020). Such arguments fundamentally expand the CASA approach but remain trapped in the relationship between individual and machine.

From our point of view, we should go a step further and broaden the perspective beyond the direct interaction of humans and machines when addressing issues of automated communication. It is apparent from the example of social bots that focusing solely on the direct interaction of humans and machines does not do justice to the phenomenon. Although direct interaction between humans and bots is undoubtedly a relevant topic (Ferrara et al., 2016; Varol et al., 2018), as is the question of how bots can be empirically determined (Cresci, 2020; Martini et al., 2021), research that focuses on the role of bots in public communication points to dynamics that go further. Florian Muhle (2022), for example, points out that the significance of Twitter bots is less their *direct* interaction with humans but, rather, their *indirect* influence on communication processes: Bots on Twitter primarily attempt to "exploit the amplification potential of the service to reach the broad journalistically manufactured public" (Muhle, 2022, p. 48). In other words, traffic is generated by the bots' retweets, whereby the platform's algorithms assign a higher relevance to certain hashtags, tweets, or accounts than to others. In this way, bots generate "public resonance" (Fürst, 2017, p. 4). In many cases, this is aimed at journalists to influence their attitudes toward certain people and topics and, as a consequence, coverage in journalistic media.[5]

Against this background, the automation of communication is to be seen both in greater depth and breadth than has often been the case. The *depth* of the phenomenon arises from the fact that the automation of communication impacts the "hybrid media system" (Chadwick, 2017) and its overall communication dynamics. Automated systems are entangled with communications across various levels through which, for example, the publics of

5. This broader view is also addressed by informatics research into human-computer interaction under the notion of tertiary users—that is, users who do not interact directly with the system but "who are affected by the introduction of the system or influence its purchase" (Alsos & Svanæs, 2011, p. 85).

online platforms and journalistic publics are placed in a dynamic relationship. However, communication dynamics can also be thought of even more broadly if we keep in mind that the data generated in automated communication become the basis for more extensive automations as is the case, for example, with automated decision-making and how this is assessed and evaluated by humans (Araujo et al., 2020; Carlson, 2018; Zarsky, 2015). The *breadth* of automated communication results from the diversity of its different technologies such as artificial companions (Pfadenhauer & Lehmann, 2022), chat bots (Beattie et al., 2020), news bots (Lokot & Diakopoulos, 2016), social bots (Keller & Klinger, 2019), work bots (Loosen & Solbach, 2020), as well as a diverse range of emerging systems.

In order to grasp this depth and breadth, we should take the automation of communication more seriously in relation to its overarching, societal character. This means not stopping at the communicative relationship between individual humans and machines but expanding our view to the role played by automation in *societal* communication. It is this perspective that we would like to assert as necessary when examining the concept of communicative AI.

Communicative AI: A Sensitizing Concept

As the last two sections outline, the automation of communication is still a relatively young and dynamic field of research. In recent years there have been a range of conceptual proposals for how this should be done. For example, references are made to “automated media” (Andrejevic, 2020), “communicative robots” (Hepp, 2020a), or “media agents” (Gambino et al., 2020). Increasingly, however, the term “communicative AI” has become established in the international research discussion (e.g. Dehnert & Mongeau, 2022; Guzman & Lewis, 2020; Natale, 2021b; Schäfer & Wessler, 2020; Stenbom et al., 2021). Andrea Guzman and Seth Lewis, who originally proposed the term, define communicative AI as “technologies designed to carry out specific tasks within the communication process that were formerly associated with humans” (2020, p. 3), a definition also shared by Agnes Stenbom et al. (2021, p. 1), and Marco Dehnert and Paul Mongeau (2022, p. 3). Mike Schäfer and Hartmut Wessler lean toward such an understanding but argue that these technologies should be understood “no longer just as mediators of communication between people, but as communicators” (2020, p. 311).

All of these proposals emphasize the communicative aspect but remain generic in the sense that they outline a specific genre of media and communication technologies without analytically reflecting both their commonality and distinction from others. For example, Guzman and Lewis’s (2020) definition raises the question of whether all automation in the communication process—including editing videos or automated translations—should be called communicative AI. In the other publications quoted above it remains unclear to what extent the term *artificial intelligence* in communicative AI is merely a buzzword—and thus a reference to the current hype around ChatGPT and similar systems—or if it is intended to refer to specific technologies such as machine learning, or what further implications are associated with it. Against this background, we propose a definition of communicative AI based on three criteria.

Communicative AI

(1) is based on various forms of automation designed for the central purpose of communication,
(2) is embedded within digital infrastructures, and
(3) is entangled with human practices.

Each of these three points require further explanation, especially if we think of them not simply in terms of societal communication.

The first point looks toward a nexus that Elena Esposito already pointed out a few years ago in an article on what she calls "artificial communication." By contrast to the discussion about the Turing Test, she emphasizes that the crucial point in "artificial communication" is not "that the machine is able to think but that it is able to communicate" (2017, p. 250; see also Esposito, 2022, pp. 14–16). This argument is an important intellectual step in that it points us to the *communicative construction* of the concept of artificial intelligence in communicative AI. Media and communications studies in particular show that the human attribution of *intelligence* to technical systems is a variable construct and does not depend on whether or not it is based on, for example, machine learning (Natale, 2021b, pp. 68–86). For example, Weizenbaum's ELIZA, developed in the 1960s, can already be understood as communicative AI because it was able to communicate with people in an automated way which then led to the attribution of *intelligence* to it, even if ELIZA was a chat program based on simple scripts (Natale, 2019; Weizenbaum, 1966). Twitter bots, which are also often based on simple scripts, are likewise communicative AI according to this understanding because they are programmed for the purpose of communication and develop their own communication dynamics. Embracing systems like these into the notion of communicative AI is helpful because it sensitizes us as media and communications researchers to consider the issue of *constructing* attributions of intelligence to simpler systems as well. From a media and communication studies' point of view, defining artificial intelligence is not so much a determination along certain technical characteristics (e.g., Mühlhoff, 2019), but a question of communicative construction including the attribution of intelligence, which is always a contested process (Bareis & Katzenbach, 2021). Such processes of construction refer to the dominant understandings of being human in a *societal* context, which typically means the capability of doing something similar to humans (e.g., Guzman, 2020), possibly including affective and emotional qualities (Beattie et al., 2020; Ling & Björling, 2020).

The second point requires just as much explanation: the embedding of communicative AI within technical infrastructures. This highlights the need to distinguish between the interface between communicative AI and its users and the underlying structures behind it. Kate Crawford and Vladan Joler (2018) have illustrated this through a rich visualization using Alexa as an example. This artificial companion operates—like Google Assistant, Microsoft Cortana, or Apple Siri—through the infrastructure of the internet, without which they would not be functional. Similarly, social bots rely on the infrastructure of platforms such as Twitter, which pre-structure communication to an extent that bots can replicate human actors comparatively easily (Gehl & Bakardjieva, 2016). In this respect, we can say that many systems of communicative AI constitute media within media as they rely on existing "infrastructural platforms" (van Dijck et al., 2018, p. 11; van Dijck et al., 2019,

p. 9) as media. The materiality of communicative AI concerns not only the primary system of automated communication, but also the materiality of the infrastructures in which this is embedded: the technical networks and server farms (Constantinides et al., 2018, p. 381). These infrastructures secure necessary data storage and processing, while simultaneously drawing communicative AI into the structures of surveillance capitalism and data colonialism (Turow, 2021). Furthermore, these infrastructures are associated with extensive "planetary costs" (Crawford, 2021) (i.e., the socio-ecological consequences of, among other things, the extraordinarily high levels of energy consumption required for the operation of digital infrastructures; Brevini, 2021; Kannengießer, 2020). If we see communicative AI in the realm of societal communication, it is important to also consider those less visible elements *as infrastructures*.

The third point—entanglement with human practice—highlights the importance of understanding that the processing of these systems cannot be understood beyond human practice. The notion of entanglement, which has gained currency through Science and Technology Studies, derives in particular from the work of Karen Barad (2007), who developed it as an analytical concept. As Susan Scott and Wanda Orlikowski (2014, pp. 881–882) argue, "the entanglement of matter and meaning is produced in practice within specific phenomena." They go on to explain that this means questioning the notion of predefined categories such as subject and object or human and nonhuman and emphasizing that such differences are constituted in the process of their relationalization:

> To be entangled is not simply to be intertwined with another, as in the joining of separate entities, but to lack an independent, discrete, self-contained existence. Existence is not an individual affair. Individuals do not pre-exist their interactions; rather, individuals emerge through and as part of their entangled intra-relating." (Barad, 2007, p. ix)

Understood in this way, the concept of entanglement is associated with a certain approach to the materiality of automated media, which strongly emphasizes their processual and relational constitution—especially in distinction to concepts seen in actor-network theory that emphasize the permanence of society in matter (Latour, 1991). More specific to the object of communicative AI, this means focusing on the coming together of matter and meaning in human practice. Materiality then becomes graspable in a double form of the technical on the one hand and the corporeality of practice on the other (Pfadenhauer & Grenz, 2017). This understanding of practices overcomes the reductionism found in some forms of practice theory (Reckwitz, 2002) by taking relationality—human beings' inevitable relatedness—into account. Yet, a focus on entanglement with human practice is also important if one wants to capture the technologies of communicative AI in more detail. For example, models for speech recognition are built on the basis of large datasets obtained via human practice online.

To sum up: If we define communicative AI in the ways outlined above, this is not simply a buzzword representing the current hype around ChatGPT, LaMDA, and similar systems, but can act as a *sensitizing concept* in Herbert Blumer's sense of the term. Following Blumer, the establishment of a sensitizing concept offers "a general sense of reference and guidance in approaching empirical instances" (Blumer, 1954, p. 7). In this sense, communicative AI draws our attention to a certain "family resemblance" (Wittgenstein, 1971) that various examples of automated communication systems share, opening up a guiding orientation, they illustrate what their breadth and depth exactly mean and why a societal perspective matters. The challenge of any sensitizing concept is, however, that it cannot be empirically operationalized without difficulty. This is the point at which "definitive concepts" (Blumer, 1954, p. 7) gain importance; that is, concepts that can be empirically operationalized. But, how exactly should we proceed with this if we want to grasp automated communication as a part of societal communication? Certainly, different answers to this question are possible; the answer we want to propose is that of a figurational approach.

Agency Between the Individual and the Machine: Taking a Figurational Approach

A figurational approach[6] seems to us particularly suitable for researching communicative AI from a societal perspective for two reasons. First, this approach does not create a contradiction between the individual and society. Society is not understood as a discrete object that surrounds humans, but as something that emerges *from* humans—all the while, the individual is produced by society. In this sense, speaking of the individual and of society is a matter of perspective, or, as Norbert Elias put it, "the concept 'individual' refers to interdependent people in the singular, and the concept 'society' refers to interdependent people in the plural" (1978, p. 125). Second, a figurational approach is particularly focused on questions of change and transformation. One of its dominant questions relates to how societies are structurally transformed, and the role of technologies in this process is an important subject of study (Elias, 1995). The main, "conceptual tool" (Elias, 1978, p. 130) used to address such nexuses is that of the *figuration*, which we can understand as a bridging concept directed toward the definitive conceptualizations necessary.

Speaking of figurations and refigurations is quite common, especially in social science research on artificial intelligence. In her analysis of "human-machine reconfigurations," Lucy Suchman (2012, p. 227), for example, takes up arguments by Donna Haraway (1997, p. 11; emphasis added) and characterizes technologies as a "*materialized figuration* that bring together assemblages of stuff and meaning into more or less stable arrangements." Sarah Kember (1998) also considers communication technologies as constituting parts of figurations, while Hubert Knoblauch and Martina Löw (2017) address them in terms of the refiguration of spaces.

Put simply, figurations are "processes of interweaving" (Elias, 1978, p. 130) of interdependent people such as a group, community, or organization. From a media and communications perspective, we can consider any figuration as a *communicative* one: It is communicative practices through which meanings are ascribed (in) figurations, and these practices are increasingly mediated. Family members, for example, may be spatially separated but connected through multimodal communication via (cell) phone, email, and exchanges on digital platforms, which maintains the everyday-world dynamics of familial relationships. Organizations are also held together as figurations using databases,

6. On process sociology, which is strongly influenced by Norbert Elias, cf. Baur & Ernst, 2011; Dunne, 2009; Morrow, 2009.

communication through an intranet, and printed flyers and other media for internal and external communication. Individuals are involved in these figurations through the roles and positions they occupy in their respective actor constellations. Conducting media and communications research using a figurational approach makes it possible to connect the perspectives of the individual and society and reflect on how the practices of their construction are closely entangled with media.

There are three basic characteristics that constitute a figuration and can be connected to established "definitive concepts" in media and communication studies (cf. Couldry & Hepp, 2016, pp. 66–67; Hepp, 2020b, pp. 100–113; Hepp & Hasebrink, 2018). The structural basis of every figuration is, first, an *actor constellation*, a network of actors who are interconnected in a certain balance of power and through interrelated communicative practices. Second, every figuration is characterized by a *frame of relevance* that guides the practices of its actors and their mutual orientation toward each other. This frame of relevance defines the action orientation of the actors involved and the specificity of the figuration. Third, figurations are constantly rearticulated in *communicative practices* that are interwoven with other social practices. These practices are typically entangled with a media ensemble.

A special theoretical feature of a figurational approach is that it opens up a way of thinking about the agency of communicative AI at all levels of the social scale, combining the perspective of the individual and its interactional relations by understanding figurations such as organizations and communities as collective actors. Figurations, of which communicative AI becomes a part, can then be understood as *hybrid figurations*. Hybrid here does not mean a dissolution of the boundary between human and machine, as can be seen in the imaginary of the cyborg (Berscheid et al., 2019; Britton & Semaan, 2017; Haraway, 1991); hybrid here refers to a unique "supra-individual" (Schimank, 2010, p. 327) agency of the overall figuration that develops in the coming together of human and machine.

This can be illustrated by the example of a newsroom where journalists use automated communication systems such as Quill from Narrative Science, ChatGPT from OpenAI, or Luminous from Aleph Alpha. A newsroom using these systems for "automating the news" (Diakopoulos, 2019) has a different agency than newsrooms without them. Research in media and communication studies is then concerned with the question of what is special about this hybrid agency and how it differs from other forms of supra-individual agency. It is also concerned with related challenges; for example, questions about authorship and the accountability of journalistic communications (Lewis et al., 2019; Montal & Reich, 2017), as well as the emergence of coping strategies for journalists that might begin to feel disconnected from technological developments (Min & Fink, 2021).

Such a figurational approach avoids dissolving the conceptual boundary between the agency of humans and machines, as has been proposed in some of the research on human-machine interaction (Banks & de Graaf, 2020). Our argument for maintaining such a boundary is an empirical one, since precisely this kind of separation is deeply embedded in everyday life. In the everyday practice of people, the question of what counts as machine-automated and what counts as human-authentic seemingly persists (Pfadenhauer & Grenz, 2017, p. 226). Similar demarcations between human and machine are made in law: The legal classification of automated systems focuses on the simple solution of attributing system behavior to natural or legal persons who developed, programmed, or implemented a system (Schulz & Schmees, 2022). Putting it metaphorically, there are no formal or accepted methods of serving a subpoena to a communicative AI.

With a figurational approach, understanding contradicting positions in the discussion about the agency of humans and machines within automated communication as different *perspectives of analysis* is rendered more straightforward. Constructivist-based theories such as social phenomenology, communicative constructivism, or systems theory, on the one hand, emphasize that machines are to be described as an objectification of human action and that the agency attributed to them is a *projection* of human actors or a *personification* of their expectations (Esposito, 2022; Knoblauch, 2020; Lindemann, 2016; Muhle, 2016; Pfadenhauer, 2015). Approaches from new materialism such as actor-network theory, or extended action theory, on the other hand, emphasize the idea of *distributed* or *shared agency* between humans and machines (Bellacasa, 2017; Gunkel, 2018a; Hanson, 2009).

Both approaches to theorizing can be understood as different perspectives on hybrid figurations: From the internal perspective of a hybrid figuration—that is: from the point of view of the people involved in it—it is a matter of *projections* and *personalized expectations* in regard to communicative AI. To take up once more the example of the newsroom, journalists do indeed project agency onto systems of automated communication when they speak of a certain system "writing a story," and they "forget" in such phrases that this happens on the basis of scripts and data that they themselves have entered into the system (Caswell & Dörr, 2018). From an external perspective (i.e., from an overall view of hybrid figurations by an observer), it is also true that this newsroom as an organizational unit possesses a different kind of *shared agency* than one without: Certain content could be published more quickly and systems of automation secure space for other kinds of journalistic work such as follow-up research and in-depth articles (Young & Hermida, 2015).

A figurational approach allows us to see not only communicative AI in terms of broader societal nexuses and to move beyond the narrow focus on the interaction between individual humans and machines, it also allows us to connect to existing concepts of media and communication studies, despite its current novelty. A view of communication is then developed that keeps its distance from technically induced transfer models and focuses on meaningful, social construction of which automated communication is a part.

At this point it is worth referring to James Carey (2009), who warned against reducing communication to defining it as the transfer of information (and its effects). Carey points out that communication should be understood as a form of symbolic reality construction (p. 19). We can see parallels when we argue for directing our attention to the various hybrid figurations of automated communication and their role in communicatively constructing society. However, Carey also pointed out that as scientists we are always confronted with the question of whether the concepts we use to grasp reality (still) correspond to how this reality is actually constructed in communication (p. 24). This also concerns the concept of communication itself, which seems to be questioned when machines automate it. But, from our point of view, this represents a misplaced response to the challenge, falling back as it does into simple transferal understandings of communication by simply explaining the machine as an actor more or less identical to the human. We now need to face this challenge to the concept of communication (i.e., Fortunati & Edwards, 2020; Guzman & Lewis, 2020;

Hepp & Loosen, 2023; Natale, 2021a). But, we also need the readiness for more complex answers than the simple equation of humans and machines.

Conclusion: Resisting the Hype Through Research

We began this article by looking at the hype around ChatGPT and other automated communication systems that are now entering the public consciousness and generating fertile academic discussion. For all the diversity of the "post-discipline" (Waisbord, 2019) and in light of earlier approaches (Gunkel, 2012), it is fair to say that our engagement with automation represents a third stage of research into digital communications. While we bask in the nascent hype and the academy's enthusiasm to embrace the discussion, as researchers it is always important to approach new phenomena reflexively. We agree that caution should be applied in the sense that, from the point of view of media and communication studies, it is important to not simply adopt the discourse from the tech companies verbatim. From our point of view, however, we should take note of the hype *insofar* as it may stand for a fundamental change in the ways we all communicate: Its automation is becoming an increasingly widespread phenomenon, and this will invariably be accompanied by changes in the ways we construct our realities.

This means, however, that the automation of communication is to be approached differently than from the limiting perspective of the interaction between individual humans and machines. We see the concept of communicative AI as a useful tool or wave upon which we might be able to sensitize ourselves to a concept requiring deeper reflection. While this increases scientific attention to automated communication, we are at the same time engaged in a discussion about what an appropriate approach might be if we are to accomplish a societal perspective on automated communication. Against this background, we have proposed a figurational approach as one such possibility.

Equipped in this way, our task is to resist the hype on the surface by *critically* examining the growth of automated communication. This means that we accept the need to question existing concepts in the field of media and communications—agency, communication, and media—and ask whether or to what extent they are still useful in a world where communication is increasingly automated by machines. At the same time, however, we should be careful not to lose sight of the boundaries that are still part of ongoing processes of societal communication. Specifically, this concerns an equation of human and machine agency or the insinuation that systems of automated communication construct meaning for themselves. These thought games can certainly sensitize us to the opportunities and risks that the increasing use of automated communication may bring and are helpful in this respect. But, it remains an empirical question to investigate what, in terms of automated communication, are the constructions we observe as part of the everyday. From our point of view, then, it is a matter of investigating the construction of reality that changes with the automation of communication and then, on this basis, working toward the further development of the scientific, conceptual apparatus. A possible point of departure, in our view, is the figurational approach.

Author Biographies

Andreas Hepp, Dr, is Professor of Media and Communications and Head of ZeMKI, Centre for Media, Communication and Information Research, University of Bremen, Germany. His research focuses on mediatization, datafication, automation of communication, communicative AI, pioneer communities, media use and appropriation.

https://orcid.org/0000-0001-7292-4147

Wiebke Loosen, Dr, is a Senior Journalism Researcher at the Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI) (Germany) as well as a professor at the University of Hamburg. Her research focuses on the transformation of journalism within a changing media environment, the journalism-audience relationship, forms of pioneering journalism and the start-up culture in journalism, as well as the datafication and automation of communication.

https://orcid.org/0000-0002-2211-2260

Stephan Dreyer, Dr, is Senior Researcher for Media Law and Media Governance at the Leibniz Institute for Media Research | Hans-Bredow-Instiut in Hamburg, Germany. He is working on legal issues of automated communication, transparency as a regulatory resource and rights-based approaches to child safety online.

https://orcid.org/0000-0002-9450-1193

Juliane Jarke, PhD, is Professor of Digital Societies at the Business Analytics and Data Science Center (BANDAS-Center) and the Department of Sociology, University of Graz. She works at the intersection of digital and feminist STS, data studies and participatory design research.

https://orcid.org/0000-0001-8349-2298

Sigrid Kannengießer, Dr, is Professor of Communication Studies with a focus on Media Sociology at the Institute for Communication Studies, University of Münster, Germany. Her research focuses on digital technologies, infrastructures, AI and sustainability, critical data practices, social movements, and gender media studies.

https://orcid.org/0000-0002-2342-9868

Christian Katzenbach, Dr, is Professor of Media and Communication at ZeMKI, University of Bremen and associated researcher at the Alexander von Humboldt Institut for Internet and Society (HIIG). His research addresses the formation of platforms and their governance, the discursive and political shaping of “Artificial Intelligence” (AI) and the increasing automation of communication.

https://orcid.org/0000-0003-1897-2783

Rainer Malaka, Dr, is Professor of Digital Media and Managing Director of the Center for Computing Technologies at the University of Bremen, Germany. His research focuses on human computer Interaction and human-centric artificial intelligence. Application areas of his research range from Entertainment Computing to Robotics and Medicine.

https://orcid.org/0000-0001-6463-4828

Michaela Pfadenhauer, Dr, is Professor of Sociology and Vice Dean for Research, Infrastructure and Sustainability at the Faculty of Social Sciences, University of Vienna. Her research focus is on sociology of knowledge and culture, social robotics and artificial companionship, mediatization, and the communicative construction of reality.

https://orcid.org/0000-0002-6082-0364

Cornelius Puschmann, Dr, is Professor of Media and Communication at ZeMKI, University of Bremen, and an affiliate researcher at the Leibniz Institute for Media Research | Hans-Bredow-Instiut in Hamburg, Germany. His interests include digital media usage, online aggression, the role of algorithms for the selection of media content, and automated content analysis.

https://orcid.org/0000-0002-3189-0662

Wolfgang Schulz, Dr, is Professor of Media Law, Public Law and Legal Theory at Hamburg University and Director of the Leibniz Institute for Media Research | Hans-Bredow-Institut (HBI) and also of the Humboldt Institute for Internet and Society. His recent research revolves around information governance, law and technology and freedom of expression.

https://orcid.org/0000-0002-9999-5508

References

Alsos, O. A., & Svanæs, D. (2011). Designing for the secondary user experience. P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque, & M. Winckler (Eds.), *13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5–9, 2011, Proceedings, Part IV* (pp. 84–91). Springer. https://doi.org/10.1007/978-3-642-23768-3_7

Andrejevic, M. (2014). The big data divide. *International Journal of Communication*, *8*(1), 1673–1689. https://espace.library.uq.edu.au/view/UQ:348586/UQ348586_OA.pdf

Andrejevic, M. (2020). *Automated media*. Routledge. https://doi.org/10.4324/9780429242595

Araujo, T., Helberger, N., Kruikemeier, S., & de Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY, 35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Baecker, D. (1997). Reintroducing communication into cybernetics. *Systemica, 11,* 11–29. https://ssrn.com/abstract=2200830

Banks, J., & de Graaf, M. M. (2020). Toward an agent-agnostic transmission model: Synthesizing anthropocentric and technocentric paradigms in communication. *Human-Machine Communication*, *1*, 19–36. https://doi.org/10.30658/hmc.1.2

Barad, K. (2007). *Meeting the university halfway: Quantum physics and the entanglement of matter and meaning*. Duke University Press. https://doi.org/10.1215/9780822388128-002

Bareis, J., & Katzenbach, C. (2021). Talking AI into being: The narratives and imaginaries of national AI strategies and their performative politics. *Science, Technology, & Human Values, 47*(5), 855–881. https://doi.org/10.1177/01622439211030007

Baur, N., & Ernst, S. (2011). Towards a process-oriented methodology: Modern social science research methods and Norbert Elias's figurational sociology. *The Sociological Review, 59*(1), 117–139. https://doi.org/10.1111/j.1467-954X.2011.01981.x

Baym, N. K. (1994). *Communication, interpretation, and relationships: A study of a computer-mediated fan community.* University of Illinois, Urbana-Champaign.

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies, 71*(3), 409–427. https://doi.org/10.1080/10510974.2020.1725082

Bellacasa, M. P. de la. (2017). *Matters of care: Speculative ethics in more than human worlds.* University of Minnesota Press.

Beniger, J. R. (1986). *The control revolution. Technological and economic origins of the information society*. Harvard University Press.

Beniger, J. R. (1990). Who are the most important theorists of communication? *Communication Research, 17*(5), 698–715. https://doi.org/10.1177/009365090017005006

Berscheid, A. L., Horwath, I., & Riegraf, B. (2019). Einleitung: Cyborgs revisited: Zur Verbindung von Geschlecht, Technologien und Maschinen. *Feministische Studien, 37*(2), 241–249. https://doi.org/10.1515/fs-2019-0025

Bibby, K. S., Margulies, F., Rijnsdorp, J. E., Withers, R. M. J., & Makarov, I. M. (1975). Man's role in control systems. *IFAC Proceedings Volumes, 8*(1), 664–683. https://doi.org/10.1016/S1474-6670(17)67612-2

Blumer, H. (1954). What is wrong with social theory? *American Sociological Review, 19*, 3–10. https://doi.org/10.2307/2088165

Brevini, B. (2021). *Is AI good for the planet?* Polity.

Britton, L. M., & Semaan, B. (2017). Manifesting the cyborg through techno-body modification. *Proceedings of the 2017 CHI conference on human factors in computing systems. Association for Computing Machinery, USA*, 2499–2510. https://doi.org/10.1145/3025453.3025629

Burns, R., Hawkins, B., Hoffmann, A. L., Iliadis, A., & Thatcher, J. (2019). Transdisciplinary approaches to critical data studies. *Proceedings of the Association for Information Science and Technology, 55*(1), 657–660. https://doi.org/10.1002/pra2.2018.14505501074

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1). https://doi.org/10.1177/2053951715622512

Carey, J. W. (2009). Communication as culture. *Essays in Media and Society*. Routledge. https://doi.org/10.4324/9780203928912

Carlson, M. (2015). The robotic reporter. *Digital Journalism, 3*(3), 416–431. https://doi.org/10.1080/21670811.2014.976412

Carlson, M. (2018). Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. *New Media & Society, 20*(5), 1755–1772. https://doi.org/10.1177/1461444817706684

Castells, M. (2000). *The rise of the network society* (2nd ed). Blackwell Publishing.

Caswell, D., & Dörr, K. (2018). Automated journalism 2.0: Event-driven narratives. *Journalism Practice, 12*(4), 477–496. https://doi.org/10.1080/17512786.2017.1320773

Cathcart, R., & Gumpert, G. (1985). The person-computer interaction: A unique source. In B. D. Ruben (Ed.), *Information and behavior. Volume 1*, 113–124. Transaction Books.

Chadwick, A. (2017). *The hybrid media system: Politics and power* (2nd ed.). Oxford University Press. https://doi.org/10.1093/oso/9780190696726.001.0001

Chesebro, J. W., & Bonsall, D. G. (Eds.). (1989). *Computer-mediated communication: Human relationships in a computerized world*. University of Alabama Press.

Constantinides, P., Henfridsson, O., & Parker, G. G. (2018). Introduction—Platforms and infrastructures in the digital age. *Information Systems Research, 29*(2), 381–400. https://doi.org/10.1287/isre.2018.0794

Couldry, N., & Hepp, A. (2016). *The mediated construction of reality*. Polity Press.

Couldry, N., & Mejías, U. A. (2019). *The costs of connection. How data is colonizing human life and appropriating it for capitalism*. Stanford University Press. https://doi.org/10.1515/9781503609754

Crawford, K. (2021). *The atlas of AI*. Yale University Press. https://doi.org/10.12987/9780300252392

Crawford, K., & Joler, V. (2018). Anatomy of an AI System. The Amazon Echo as an anatomical map of human labor, data and planetary resources. *Virtual Creativity, 9*(1–2). 117–120. https://doi.org/10.1386/vcr_00008_7

Crawford, K., Miltner, K. M., & Gray, M. W. (2014). Critiquing big data: Politics, ethics, epistemology. *International Journal of Communications—Special Section Introduction, 8*, 1663–1672. https://ijoc.org/index.php/ijoc/article/view/2167/1164

Cresci, S. (2020). A decade of social bot detection. *Communications of the ACM, 63*(10), 72–83. https://doi.org/10.1145/3409116

Dalton, C., & Thatcher, J. (2014). *What does a critical data studies look like, and why do we care? Seven points for a critical approach to 'big data.'* Digital Geographies. Retrieved on March 30, 2023, from https://web.archive.org/web/20200928063532/https://www.societyandspace.org/articles/what-does-a-critical-data-studies-look-like-and-why-do-we-care

Dehnert, M., & Mongeau, P. A. (2022). Persuasion in the age of artificial intelligence (AI): Theories and complications of AI-based persuasion. *Human Communication Research, 48*(3), 386–403. https://doi.org/10.1093/hcr/hqac006

Dencik, L., & Kaun, A. (2020). Datafication and the welfare state. *Global Perspectives, 1*(1). https://doi.org/10.1525/gp.2020.12912

Diakopoulos, N. (2019). *Automating the news*: *How algorithms are rewriting the media*. Harvard University Press. https://doi.org/10.4159/9780674239302

Dunne, S. (2009). The politics of figurational sociology. *Sociological Review, 57*(1), 28–57. https://doi.org/10.1111/j.1467-954X.2008.01803.x

Edwards, C., Edwards, A., Spence, P. R., & Shelton, A. K. (2014). Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. *Computers in Human Behavior, 33*, 372–376. https://doi.org/10.1016/j.chb.2013.08.013

Elias, N. (1978). *What is sociology?* Hutchinson.

Elias, N. (1995). Technization and civilization. *Theory, Culture & Society, 12*(3), 7–42. https://doi.org/10.1177/026327695012003002

Esposito, E. (2017). Artificial communication? The production of contingency by algorithms. *Zeitschrift für Soziologie, 46*(4), 249–265. https://doi.org/10.1515/zfsoz-2017-1014

Esposito, E. (2022). *Artificial communication: How algorithms produce social intelligence*. The MIT Press. https://doi.org/10.7551/mitpress/14189.001.0001

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Communications of the ACM, 59*(7), 96–104. https://doi.org/10.1145/2818717

Flensburg, S., & Lomborg, S. (2021). Datafication research: Mapping the field for a future agenda. *New Media & Society, 0*(0). https://doi.org/10.1177/14614448211046616

Fortunati, L., & Edwards, A. P. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication, 1*, 7–18. https://doi.org/10.30658/hmc.1.1

Fortunati, L., & Edwards, A. P. (2021). Moving ahead with human-machine communication. *Human-Machine Communication, 2*, 7–28. https://doi.org/10.30658/hmc.2.1

Fürst, S. (2017). Öffentlichkeitsresonanz als Nachrichtenfaktor—Zum Wandel der Nachrichtenselektion. *MedienJournal, 37*(2), 4–15. https://doi.org/10.24989/medienjournal.v37i2.122

Gambino, A., Fox, J., & Ratan, R. (2020). Building a stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication, 1*, 71–86. https://doi.org/10.30658/hmc.1.5

Gehl, R. W., & Bakardjieva, M. (2016). Socialbots and their friends. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and their friends: Digital media and the automation of sociality* (pp. 1–16). Taylor & Francis. https://doi.org/10.4324/9781315637228

Gillespie, T. (2014). The relevance of algorithms. In T. Gillespie, P. J. Boczkowski, & K. A. Foot (Eds.), *Media technologies. Essays on communication, materiality, and society* (pp. 167–194). The MIT Press. https://doi.org/10.7551/mitpress/9780262525374.003.0009

Gitelman, L. (Ed.). (2013). *'Raw data' is an oxymoron*. The MIT Press. https://doi.org/10.7551/mitpress/9302.001.0001

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society, 7*(1). https://doi.org/10.1177/2053951719897945

Graefe, A., & Bohlken, N. (2020). Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. *Media and Communication, 8*(3), 50–59. https://doi.org/10.17645/mac.v8i3.3019

Gunkel, D. J. (2012). Communication and artificial intelligence: Opportunities and challenges for the 21st century. *communication +1, 1*(1), 1–25. https://doi.org/10.7275/R5QJ7F7R

Gunkel, D. J. (2018a). Ars ex machina: Rethinking responsibility in the age of creative machines. In A. L. Guzman (Ed.), *Human-machine communication: Rethinking communication, technology, and ourselves* (pp. 221–236). Peter Lang. https://doi.org/10.3726/b14399

Gunkel, D. J. (2018b). *Robot rights*. The MIT Press. https://doi.org/10.7551/mitpress/11444.001.0001

Guzman, A. L. (2015). *Imagining the voice in the machine: The ontology of digital social agents*. PhD dissertation. University of Illinois at Chicago. https://hdl.handle.net/10027/19842

Guzman, A. L. (2018). Introduction: What is human-machine-communication anyway? In A. L. Guzman (Ed.), *Human-machine communication* (pp. 1–28). Peter Lang. https://doi.org/10.3726/b14399

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication, 1*, 37–54. https://doi.org/10.30658/hmc.1.3

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A Human-Machine Communication research agenda. *New Media & Society, 22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Guzman, A. L., McEwen, R., & Jones, S. (Eds.). (2023). *The SAGE Handbook of Human-Machine Communication*. Sage.

Hanson, F. A. (2009). Beyond the skin bag: On the moral responsibility of extended agencies. *Ethics and information technology, 11*(1), 91–99. https://doi.org/10.1007/s10676-009-9184-z

Haraway, D. (1991). A Cyborg manifesto. Science, technology, and socialist-feminism in the late twentieth century. In D. Haraway (Ed.), *Simians, cyborgs and women: The reinvention of nature* (pp. 149–181). Routledge.

Haraway, D. (1997). *Modest_Witness@Second_Millennium.FemaleMan©_Meets_OncoMouse™. Feminism and Technoscience* (1st ed.). Routledge. https://doi.org/10.1023/a:1004349615837

Hepp, A. (2020a). Artificial companions, social bots and work bots: Communicative robots as research objects of media and communication studies. *Media, Culture & Society, 42*(7–8), 1410–1426. https://doi.org/10.1177/0163443720916412

Hepp, A. (2020b). *Deep mediatization*. Routledge. https://doi.org/10.4324/9781351064903

Hepp, A., & Hasebrink, U. (2018). Researching transforming communications in times of deep mediatization: A figurational approach. In A. Hepp, A. Breiter, & U. Hasebrink (Eds.), *Communicative figurations: Transforming communications in times of deep mediatization* (pp. 51–80). London: Palgrave Macmillan. https://doi.org/10.1007/978-3-319-65584-0_2

Hepp, A., Jarke, J., & Kramp, L. (Eds.). (2022). *New perspectives in critical data studies*. Palgrave Macmillan. https://doi.org/10.1007/978-3-030-96180-0

Hepp, A., & Krotz, F. (Eds.). (2014). *Mediatized worlds: Culture and society in a media age.* Palgrave Macmillan. https://doi.org/10.1057/9781137300355

Hepp, A., & Loosen, W. (2023). The interdisciplinarity of HMC: Rethinking communication, media and agency. In A. L. Guzman, R. McEwen, & S. Jones (Eds.), *The SAGE Handbook of Human-Machine Communication* (Preprint). Sage.

Heuer, H., Jarke, J., & Breiter, A. (2021). Machine learning in tutorials—Universal applicability, underinformed application, and other misconceptions. *Big Data & Society, 8*(1). https://doi.org/10.1177/20539517211017593

Hjarvard, S. (2013). *The mediatization of culture and society*. Routledge. https://doi.org/10.4324/9780203155363

Hofmann, J., Katzenbach, C., & Gollatz, K. (2017). Between coordination and regulation: Finding the governance in internet governance. *New Media & Society, 19*(9), 1406–1423. https://doi.org/10.1177/1461444816639975

Iliadis, A., & Russo, F. (2016). Critical data studies: An introduction. *Big Data & Society, 3*(2). https://doi.org/10.1177/2053951716674238

Jones, S. G. (Ed.). (1998). *Cybersociety 2.0. Revisiting computer-mediated communication and technology*. Sage. https://doi.org/10.4135/9781452243689

Kannengießer, S. (2020). Acting on media for sustainability. In H. C. Stephansen & E. Treré (Eds.), *Citizen media and practice: Currents, connections, challenges* (pp. 176–188). Routledge. https://doi.org/10.4324/9781351247375-13

Keller, T. R., & Klinger, U. (2019). Social bots in election campaigns: Theoretical, empirical, and methodological implications. *Political Communication, 36*(1), 171–189. https://doi.org/10.1080/10584609.2018.1526238

Kember, S. (1998). *Virtual anxiety: Photography, new technologies and subjectivity.* Manchester University Press.

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage. https://doi.org/10.4135/9781473909472

Knoblauch, H. (2020). *The communicative construction of reality*. Routledge. https://doi.org/10.4324/9780429431227

Knoblauch, H., & Löw, M. (2017). On the spatial re-figuration of the social world. *Sociologica, 11*(2), 1–27. https://doi.org/10.2383/88197

Latour, B. (1991). Technology is society made durable. In J. Law (Ed.), *A sociology of monsters. Essays on power, technology and domination* (pp. 103–131). Routledge. https://doi.org/10.1111/j.1467-954x.1990.tb03350.x

Lazer, D., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, J.-E. R., & Nass, C. I. (2010). Trust in computers: The Computers-Are-Social-Actors (CASA) paradigm and trustworthiness perception in human-computer communication. In D. Latusek & A. Gerbasi (Eds.), *Trust and technology in a ubiquitous modern environment* (pp. 1–15). IGI Global. https://doi.org/10.4018/978-1-61520-901-9.ch001

Lewis, S. C., Sanders, A. K., & Carmody, C. (2019). Libel by algorithm? Automated journalism and the threat of legal liability. *Journalism & Mass Communication Quarterly, 96*(1), 60–81. https://doi.org/10.1177/1077699018755983

Li, Z., & Li, C. (2014). Twitter as a social actor: How consumers evaluate brands differently on Twitter based on relationship norms. *Computers in Human Behavior, 39*, 187–196. https://doi.org/10.1016/j.chb.2014.07.016

Licklider, J. C. R., & Taylor, R. W. (1968). The computer as a communication device. *Science and Technology, 76*(2), 21–31.

Lindemann, G. (2016). Social interaction with robots: Three questions. *AI & society, 31*(4), 573–575. https://doi.org/10.1007/s00146-015-0633-4

Ling, H. Y., & Björling, E. A. (2020). Sharing stress with a robot: What would a robot say? *Human-Machine Communication, 1*, 133–159. https://doi.org/10.30658/hmc.1.8

Livingstone, S. M. (2009). On the mediation of everything: ICA presidential address 2008. *Journal of Communication, 59*(1), 1–18. https://doi.org/10.1111/j.1460-2466.2008.01401.x

Lokot, T., & Diakopoulos, N. (2016). News bots. *Digital Journalism, 4*(6), 682–699. https://doi.org/10.1080/21670811.2015.1081822

Loosen, W., & Solbach, P. (2020). Künstliche Intelligenz im Journalismus? Was bedeutet Automatisierung für journalistisches Arbeiten? In T. Köhler (Ed.), *Fake News, Framing, Fact-Checking: Nachrichten im digitalen Zeitalter* (pp. 177–203). Transcript. https://doi.org/10.1515/9783839450253-010

Lundby, K. (Ed.). (2014). *Mediatization of communication*. de Gruyter. https://doi.org/10.1515/9783110272215

Martini, F., Samula, P., Keller, T. R., & Klinger, U. (2021). Bot, or not? Comparing three methods for detecting social bots in five political discourses. *Big Data & Society, 8*(2), https://doi.org/10.1177/20539517211033566

Mattelart, A. (2003). *The information society: An introduction*. Sage.

McQuail, D., & Deuze, M. (2020). *McQuail's media and mass communication theory* (7th ed.). Sage.

Min, S. J., & Fink, K. (2021). Keeping up with the technologies: Distressed journalistic labor in the pursuit of "shiny" technologies. *Journalism Studies, 22*(14), 1987–2004. https://doi.org/10.1080/1461670x.2021.1979425

Montal, T., & Reich, Z. (2017). I, robot. You, journalist. Who is the author? *Digital Journalism, 5*(7), 829–849. https://doi.org/10.1080/21670811.2016.1209083

Morris, M., & Ogan, C. (1996). The internet as mass medium. *Journal of Communication, 46*(1), 39–50. http://doi.org/10.1111/j.1460-2466.1996.tb01460.x

Morrow, R. A. (2009). Norbert Elias and figurational sociology: The comeback of the century. *Contemporary Sociology: A Journal of Reviews, 38*(3), 215–219. https://doi.org/10.1177/009430610903800301

Muhle, F. (2016). "Are you human?" Plädoyer für eine kommunikationstheoretische Fundierung interpretativer Forschung an den Grenzen des Sozialen. *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research 17*(1), 33. https://doi.org/10.17169/fqs-17.1.2489

Muhle, F. (2022). Socialbots at the Gates. Plädoyer für eine holistische Perspektive auf automatisierte Akteure in der Umwelt des Journalismus. *Medien & Kommunikationswissenschaft, 70*(1-2), 40–59. https://doi.org/10.5771/1615-634X-2022-1-2-40

Mühlhoff, R. (2019). Human-aided artificial intelligence: Or, how to run large computations in human brains? Toward a media sociology of machine learning. *New Media & Society, 22*(10), 1868–1884. https://doi.org/10.1177/1461444819885334

Napoli, P. M. (2014). Automated media: An institutional theory perspective on algorithmic media production and consumption. *Communication Theory, 24*(3), 340–360. https://doi.org/10.1111/comt.12039

Nass, C., Takayama, L., & Brave, S. (2006). Socializing consistency: From technical homogeneity to human epitome. In P. Zhang & D. F. Galletta (Eds.), *Human-computer interaction and management information systems: Foundations* (pp. 373–391). Routledge. https://doi.org/10.4324/9781315703619

Natale, S. (2019). If software is narrative: Joseph Weizenbaum, artificial intelligence and the biographies of ELIZA. *New Media & Society, 21*(3), 712–728. https://doi.org/10.1177/1461444818804980

Natale, S. (2021a). Communicating through or communicating with: Approaching artificial intelligence from a communication and media studies perspective. *Communication Theory, 31*(4), 905–910. https://doi.org/10.1093/ct/qtaa022

Natale, S. (2021b). *Deceitful media*. Oxford University Press. https://doi.org/10.1093/oso/9780190080365.001.0001

Pfadenhauer, M. (2015). The contemporary appeal of artificial companions: Social robots as vehicles to cultural worlds of experience. *The Information Society, 31*(3), 284–293. https://doi.org/10.1080/01972243.2015.1020213

Pfadenhauer, M., & Grenz, T. (2017). Von Objekten zu Objektivierung. *Soziale Welt, 68*(2–3), 225–242. https://doi.org/10.5771/0038-6073-2017-2-3-225

Pfadenhauer, M., & Lehmann, T. (2022). Affects after AI: Sociological perspectives on artificial companionship. In A. Elliott (Ed.), *The Routledge Social Science Handbook of AI* (pp. 91–106). Routledge. https://doi.org/10.4324/9780429198533-7

Reckwitz, A. (2002). Toward a theory of social practices. A development in culturalist theorizing. *European Journal of Social Theory*, *5*(2), 245–265. https://doi.org/10.1177/13684310222225432

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

Reeves, J. (2016). Automatic for the people: The automation of communicative labor. *Communication and Critical/Cultural Studies, 13*(2), 150–165. https://doi.org/10.1080/14791420.2015.1108450

Richards, R., Spence, P., & Edwards, C. (2022). Human-machine communication scholarship trends: An examination of research from 2011 to 2021 in communication journals. *Human-Machine Communication, 4*, 45–65. https://doi.org/10.30658/hmc.4.3

Schäfer, M. S., & Wessler, H. (2020). Öffentliche Kommunikation in Zeiten künstlicher Intelligenz. *Publizistik*, *65*(3), 307–331. https://doi.org/10.1007/s11616-020-00592-6

Schimank, U. (2010). *Handeln und Strukturen. Einführung in die akteurstheoretische Soziologie* (4th ed.). Juventa.

Schulz, W., & Schmees, J. (2022). Möglichkeiten und Grenzen der Künstlichen Intelligenz in der Rechtsanwendung. In I. Augsberg & G. F. Schuppert (Eds.), *Wissen und Recht* (pp. 561–593). Nomos. https://doi.org/10.5771/9783748921479-561

Scott, S. V., & Orlikowski, W. J. (2014). Entanglements in practice: Performing anonymity through social media. *Management Information Systems Quarterly*, *38*(3), 873–893. https://doi.org/10.25300/misq/2014/38.3.11

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424. https://doi.org/10.1017/s0140525x00005756

Star, S. L., & Ruhleder, K. (1996). Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, *7*(1), 111–134. https://doi.org/10.1287/isre.7.1.111

Steels, L., & Kaplan, F. (2000). AIBO's first words: The social learning of language and meaning. *Evolution of communication, 4*(1), 3–32. https://doi.org/10.1075/eoc.4.1.03ste

Stenbom, A., Wiggberg, M., & Norlund, T. (2021). Exploring communicative AI: Reflections from a Swedish newsroom. *Digital Journalism*, 1–19. https://doi.org/10.1080/21670811.2021.2007781

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

Suchman, L. A. (2012). Configuration. In C. Lury & N. Wakeford (Eds.), *Inventive methods. The happening of the Social* (pp. 48–60). Taylor and Francis. https://doi.org/10.4324/9780203854921

Thurman, N., Lewis, S. C., & Kunert, J. (2019). Algorithms, automation, and news. *Digital Journalism*, *7*(8), 980–992. https://doi.org/10.1080/21670811.2019.1685395

Turner, F. (2006). *From counterculture to cyberculture: Stewart Brand, the Whole Earth Network, and the rise of digital utopianism*. University of Chicago Press.

Turow, J. (2021). *The voice catchers: How marketers listen in to exploit your feelings, your privacy, and your wallet*. Yale University Press. https://doi.org/10.12987/9780300258738

van Dijck, J. (2014). Datafication, dataism and dataveillance: Big Data between scientific paradigm and ideology. *Surveillance and Society*, *12*(2), 197–208. https://doi.org/10.24908/ss.v12i2.4776

van Dijck, J., Nieborg, D., & Poell, T. (2019). Reframing platform power. *Internet Policy Review*, *8*(2). https://doi.org/10.14763/2019.2.1414

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Varol, O., Davis, C. A., Menczer, F., & Flammini, A. (2018). Feature engineering for social bot detection. In G. Dong & H. Liu (Eds.), *Feature Engineering for Machine Learning and Data Analytics* (pp. 311–334). CRC Press. https://doi.org/10.1201/9781315181080-12

Veale, T., & Cook, M. (2018). *Twitterbots: Making machines that make meaning*. The MIT Press. https://doi.org/10.7551/mitpress/10859.001.0001

Volcic, Z., & Andrejevic, M. (2023). Automated media and commercial populism. *Cultural Studies, 37*(1), 149–167. https://doi.org/10.1080/09502386.2022.2042581

Waisbord, S. (2019). *Communication: A post-discipline.* John Wiley & Sons.

Wang, W. (2017). Smartphones as social actors? Social dispositional factors in assessing anthropomorphism. *Computers in Human Behavior, 68*, 334–344. https://doi.org/10.1016/j.chb.2016.11.022

Weizenbaum, J. (1966). ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*(1), 36–45. https://doi.org/10.1145/365153.365168

Wellman, B., Salaff, J. W., Dimitrova, D. S., Garton, L., Gulia, M., & Haythornthwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology*, *22*(1), 213–238. https://doi.org/10.1146/annurev.soc.22.1.213

Wittgenstein, L. (1971). *Philosophische Untersuchungen* (1st ed.). Suhrkamp Verlag.

Young, M. L., & Hermida, A. (2015). From Mr. and Mrs. Outlier to central tendencies: Computational journalism and crime reporting at the Los Angeles Times. *Digital Journalism*, *3*(3), 381–397. https://doi.org/10.1080/21670811.2014.976409

Zarsky, T. (2015). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, *41*(1), 118–132. https://doi.org/10.1177/0162243915605575

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.

HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.5

Human-AI Teaming During an Ongoing Disaster: How Scripts Around Training and Feedback Reveal This Is a Form of Human-Machine Communication

Keri K. Stephens[1], Anastazja G. Harris[1], Amanda Hughes[2], Carolyn E. Montagnolo[1], Karim Nader[3], S. Ashley Stevens[2], Tara Tasuji[1], Yifan Xu[1], Hemant Purohit[4], and Christopher W. Zobel[5]

1 The University of Texas at Austin, Moody College of Communication, Austin, TX, USA
2 Brigham Young University, Dept. of Computer Science, Provo, UT, USA
3 The University of Texas at Austin, Department of Philosophy, Austin, TX, USA
4 George Mason University, School of Computing, Fairfax, VA, USA
5 Virginia Tech, Dept. of Business Information Technology, Blackburg, Virginia, USA

Abstract

Humans play an integral role in identifying important information from social media during disasters. While human annotation of social media data to train machine learning models is often viewed as human-computer interaction, this study interrogates the ontological boundary between such interaction and human-machine communication. We conducted multiple interviews with participants who both labeled data to train machine learning models and corrected machine-inferred data labels. Findings reveal three themes: scripts invoked to manage decision-making, contextual scripts, and scripts around perceptions of machines. Humans use scripts around training the machine—a form of behavioral anthropomorphism—to develop social relationships with them. Correcting machine-inferred data labels changes these scripts and evokes self-doubt around who is right, which substantiates the argument that this is a form of human-machine communication.

Keywords: human-AI teaming, supervised machine learning, scripts, human-machine communication, behavioral anthropomorphism

CONTACT Keri K. Stephens • keristephens@austin.utexas.edu • The University of Texas at Austin, Moody College of Communication • 2504A Whitis Ave. • A1105 • Austin, TX 78731

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

Communication research historically positioned technology as a medium through which communication occurs. But today, technologies—including machines of many forms—are more visible to humans, capable of interacting, and sometimes even have the capacity to communicate with humans (Guzman, 2018). One place where humans work closely with machines is in the field of machine learning—the fastest growing area of modern information technology, where an entire subfield of inquiry has been defined as human-in-the-loop machine learning (Monarch, 2021). Machine learning is sometimes referred to as artificial intelligence (AI), but it has many subfields beyond learning, such as knowledge representation and reasoning (Russell & Norvig, 2009). The rapidly growing field of machine learning and AI is leading transformation for many industries and work sectors, and governments around the world have launched associated strategic initiatives, such as the National AI Initiative of the United States (https://www.ai.gov/). As per the 2021 survey report by McKinsey (2021), AI adoption continues to grow in organizations globally.

Humans are involved in many steps of a machine learning system's pipeline, but the most common is in labeling data to create a training set for supervised machine learning. This set can then be used by a learning algorithm to develop a model for a predictive task. For example, humans might annotate a set of email data so that a machine learning model can be trained to automatically classify whether an email message is spam or not. In such cases, the quality of the labeled data provided by the human is pivotal because the machine takes that information, whether good or bad, and learns to recognize patterns based on the human input.

State-of-the-art practices for finding relevant data in social media during a disaster include having humans work with AI-infused systems (referred interchangeably as machines hereon) to identify and label relevant information posted on social media (Imran et al., 2015; Purohit et al., 2018). The humans annotate the data which, in turn, helps the machine to discover patterns associated with relevant versus irrelevant disaster data. Typically, this is where the interaction stops. However, sometimes humans not only provide the labels, but they also provide correction evaluating how well the machine actually identified the patterns (Amershi et al., 2014). Although this process of correcting the machine can contribute to developing more efficient human-in-the-loop machine learning systems (Monarch, 2021), few research efforts also consider how the humans feel and experience this more involved type of interaction.

There are many terms used to explain the interactions between humans and machines. One problem is that while the word "interaction" implies a back-and-forth type of engagement, the definitions and interpretations of the term vary in the field of computer interaction (Rogers, 2012). Harrison et al. (2007) outline three paradigms associated with interaction that are commonly found in HCI research. The first and oldest paradigm envisions interaction as a coupling of human and machine, and research in this tradition seeks ways to optimize the fit between the two. The second paradigm treats interaction as a form of information transfer where the goal is to improve the accuracy and efficiency of that process. In the third and most recent paradigm, interaction is seen as phenomenologically situated, meaning that the context of the interaction and characteristics of the human and machine play important roles in how the interaction takes place. This paradigm moves beyond examining flows of information as interaction to understanding how meaning is constructed between machine and human. Human-machine communication (HMC) scholars consider HMC to be an "umbrella encompassing the many approaches to people's communication with various technologies" (Guzman, 2018, p. 22). HMC research studies the "creation of meaning among humans and machines" (Guzman, 2018, p. 1), much like the third paradigm of HCI described above. HMC is especially interested in humans' interactions with technologies, machines, and data that function as communicative subjects (Spence, 2019). With this in mind, this manuscript interrogates the ontological boundary of when an interaction becomes a form of human-machine communication.

This study offers one of the first longitudinal accounts of observing people as they train machines to find meaningful data within social media in the context of helping disaster management agencies. Specifically, we worked with volunteers from a Community Emergency Response Team (CERT) that were tasked with labeling social media data to train and improve an AI-infused system (*aka* machine) for social media filtering during the COVID-19 disaster. The training of the AI-infused system (name blinded for peer review) is based on collecting and extracting relevant tweets for a topic (e.g., the COVID-19 pandemic) from the Twitter data stream and presenting them one at a time for a human to characterize and label the class of behavior contained in the tweet. The resulting data set is then used to train the system to recognize the inherent patterns. The main theoretical contribution of this work is derived by observing a specific group of people labeling the initial training data, and then observing the same people correcting machine-inferred labels that were applied to a subset of that data by the trained system. This longitudinal approach is what provides evidence for how these practices constitute a form of human-machine communication.

The manuscript begins by providing theoretical perspectives around machines and their ability to function as social actors in communicative activities. Based on the concept of scripts as knowledge structures people hold that help them understand how to act or understand events, we raise questions around the use of related terms such as "Human-AI Teaming" and "Human-Machine Communication" with the goal of more precisely defining them. The methods and analyses then describe our interviews and observations. The results subsequently address the following two research questions: *What specific scripts are involved as people engaged in this form of machine-related work?* and *To what extent can the process of humans labeling data and providing iterative feedback to machines be considered a form of human-machine communication?* We end by discussing the contributions of these findings and how to continue advancing HMC understanding.

Theoretical Perspectives on Human-Machine Communication

On a structural level, machines can be conceptualized as part of the social structure of everyday human life with the ability to direct human behaviors and influence interaction outcomes (Latour, 1994). On the individual level, the Computers are Social Actors (CASA) approach assumes that humans interact with machines as if they are social others and thus mindlessly apply social rules and expectations to machines despite knowing that they do not possess human emotions and intentions (Nass & Moon, 2000). However, recent studies show that humans might not mindlessly interact with machines as if they are human, but instead interaction is the process of communicating via information exchange, grounded in a form of behavioral anthropomorphism—nonhuman objects acting in ways expected of humans—through which machines become social actors and valued teaming partners (Nowak & Fox, 2018). In a supervised machine learning context, a high level of human input is required, such as labeling raw data for a given set of class labels or correcting machine decisions about the automated labeling of data. Rather than viewing intelligent technologies as replacements for humans, in these situations we should consider them as complements to human capabilities, and hence a team member to the human (Gibbs et al., 2021). When AI and humans team, AI may become more than just a tool; it might be viewed as scalable human knowledge (Malone, 2018).

Human and Machine Roles

To understand human-AI teaming during disasters, we need to understand different roles that the human and machine may take and how they interact. Madni and Madni (2018) point out that human-machine teams can function in different ways. In the case of a human in a supervisory role, humans take direct control over the machine and can intervene by adjusting the algorithm; the machine then carries out the commands set by human supervisors. Machines can also substitute for a human by automatically and independently labeling data through unsupervised machine learning. Another way machines can function in a human-machine team is to support crowdsourcing efforts, especially in a disaster context. Crowdsourcing, also called the use of digital volunteers, is a broad term that encapsulates the concept that groups of people, often unknown to one another, come together online to label data and provide a training set for machines (Alam & Campbell, 2017). Such groups of online volunteers have worked with many disasters around the globe using maps to identify lost property and finding important information located on public social media (Fathi et al., 2019; Hughes & Tapia, 2015; Starbird & Palen, 2011). In this type of role, groups of humans typically provide one-way data labeling services that the machine uses for learning; the human serves as the curator of input for a machine.

Scripts as Behavioral Guides When Engaging With Machines

We turn now to examine how relationships might be developed between humans and machines. As knowledge structures, scripts help people understand how to act or interpret events, and they are developed by observing others and drawing on past experiences (Gioia & Poole, 1984). Researchers try to uncover these representations of knowledge to help us better understand the cognitive reasons behind human actions. For example, scripts have been studied to identify why people refuse to participate in surveys, and the findings can be used to design new ways to improve participation (Stephens et al., 2014). While Nass and Moon (2000) referred to scripts as heuristics that can lead to mindless behaviors, scripts have also been found to help people make sense of their situations, and thus they also can be structures that more consciously guide behavior and thoughts (Gioia & Poole, 1984). Scholars studying human-machine communication have found that people activate some communicative scripts mindlessly when they interact with machines because they draw upon deeply held cultural perceptions (Dehnert & Leach, 2021), a form of stereotype or strong script (Abelson, 1976). In the case of a novel situation, people may not have a script for that specific situation, so they search their repertoire of scripts and look for ways to make sense of the situation using a script from another context (Stephens et al., 2014).

When invoking scripts, people also apply subjective judgments, and in human-machine interaction/communication, this means they could be biased or overly confident in a machine's ability to do a task. Studies have shown that people view information system decisions on tasks mainly involving mechanical skills—defined as processing quantitative data objectively—as equally trusted and fair compared to human-made decisions, and they had similar emotions toward the system and a human (Lee, 2018). As opposed to a human task, defined as one requiring subjective judgment and emotional capability, the task of training machines to recognize actionable disaster information closely resembles what Lee defined as a mechanical task. Situational and individual characteristics, such as one's attitude and knowledge toward AI, also predict these preferences (Utz et al., 2021).

Considering that human-AI interaction is integral to the experience of the CERT volunteers in this study, it is important to identify the scripts that volunteers invoke while training and correcting the machine. This knowledge can help emergency managers better understand how to support volunteers, and leads to the following research question:

> **RQ1:** What specific scripts are involved as people engage in iterative supervised machine learning work?

Human-Machine Interaction or Human-Machine Communication

Once we understand the scripts that people invoke when engaged in supervised machine learning, we can better interrogate whether, and to what extent, humans can move beyond simply interacting with AI systems and become involved in a communicative process. This leads to our second research question:

> **RQ2:** To what extent can the process of humans labeling data and providing iterative feedback to machines be considered a form of human-machine communication?

Method

This research project began in April of 2020 when our team recognized the COVID-19 pandemic as an opportunity to further develop a web-based AI-infused system called CitizenHelper (Karuna et al., 2017). This system uses AI techniques of machine learning and natural language processing to examine tweets and extract useful information for emergency responders from social media data. For example, during COVID-19, emergency responders needed to know if people were crowding the workers at sites giving out emergency supplies, because they could then send additional help to those locations. Social media offers an increasingly relevant data source for this purpose, especially when it comes to discovering data that emergency managers did not know they were looking for (St. Denis et al., 2020), such as an inability to maintain social distancing and thus an increased risk to public health at the supply distribution sites. The AI-infused system can extract such information from social media streams automatically but it relies on human-labeled data for training its machine learning models. In this study, we wanted to better understand the people who perform data labeling tasks and how they experience the AI-infused system *aka* machine as a social actor.

Using a rigorously designed qualitative data collection protocol, we conducted 55 interviews with 14 Community Emergency Response Team (CERT) participants as they labeled Twitter messages (tweets) related to the COVID-19 disaster as it unfolded. CERT volunteers were chosen for this task because they have all taken a well-documented US-wide curriculum offered through the Federal Emergency Management Association (FEMA, 2022) that teaches these volunteers about emergency response practices. Thus, they have a baseline understanding of disaster activations and what might be relevant as they examine tweets to identify meaningful data helpful for emergency response efforts. Data collection took place in two phases.

Providing labels for the machines. During phase I of the data collection, which occurred during May and June of 2020, we interviewed and observed 13 CERT volunteers on three separate 1-hour time periods. Additionally, the CERT leader was interviewed two separate times, which made a total of 41 phase I interviews. The task for the phase I interviews was to have volunteers collectively label over 5,000 tweets to train the machine learning-based model for natural language processing, which could then automatically infer a given set of class labels for a tweet to support social media analytics for COVID-19 response at large scale. Each volunteer was given a set of 500 tweets that the researcher working with the AI-infused system pulled randomly from a dataset collected from the Twitter stream as follows. We used the Twitter Streaming Application Programming Interface (API) and its geo-fencing method, which filters and provides tweets that originated from a given region represented through a bounding box. We provided the geo-coordinates of the bounding box surrounding the Washington, DC, Metro region (i.e., U.S. National Capital Region), as suggested by the CERT team leader. We were able to collect approximately 2.1 million tweets through this method during the period of March to May 2020. We further employed a filtering criterion to identify potentially relevant tweets for COVID-19 response by checking the presence of relevant keywords based on a list containing 1,521 keywords that was curated with the help of CERT volunteers. A total of 14,000 unique tweets were randomly sampled from the resulting filtered tweets to create a dataset for preparing the labeling tasks for CERT volunteers.

Given the labeling task interface with 500 tweets presented one at a time, the volunteer was then asked to assign the following labels (as appropriate) to each tweet: *Relevant*, *Prevention*, *Risk*, *Positive Sentiment*, and *Negative Sentiment*. Volunteers had a detailed coding book with examples of each of these labels and they underwent multiple training events. For context, we will briefly describe the labels here. Because this project is meant to serve the needs of emergency responders in the Maryland and Washington, DC, areas of the United States, only tweets depicting COVID-19-related activity in that particular geographic area were coded as *Relevant*. All such relevant tweets were then considered for labeling into one or more additional categories. *Prevention* tweets were those that contained information about how people were preventing the spread of COVID-19, and *Risk* labels were placed when tweets indicated risky behaviors related to COVID-19. *Positive Sentiment* or *Negative Sentiment* were labeled when tweets contained views reflecting positive or negative actions around COVID-19. These labels were developed in consultation with the practitioner CERT leader on the project and were determined to be of importance for emergency managers. The focus of this study is on the volunteers who actually applied these labels to the data.

Verifying and correcting the machine. Phase II of the data collection consisted of interviewing seven of the phase I volunteers an additional two times (a total of 14 interviews) for a slightly different task. Considering the need to collect this data quickly for the machine learning process, we used participants who were available in phase II. Instead of providing their own labels, volunteers were each given 250 tweets that had already been labeled by CitizenHelper and they were asked to verify/correct these labels. These interviews were conducted in July and early August 2020, and the assigned task allowed for more observation and reflection on the relationship between the human labeler and the AI-infused system.

Table 1 on the following page describes each participant's involvement in the research, the technology they used, their age, and their expertise that was relevant to the labeling tasks they performed. One participant preferred to state their age in a range, and we did not ask for other demographics. The IRB approved this study, volunteers agreed to participate and be recorded (audio and video), and all participants were compensated with a gift card at the rate of $25 USD per hour.

All interviews lasted approximately 1 hour and took place online over the Zoom platform. Two researchers were present for each interview, one to lead and the other to observe, take notes, and troubleshoot technical difficulties. Researchers observed the volunteers' screens (shared through Zoom) while they labeled tweets (for more details see Stephens et al., 2021). Throughout each session, volunteers were asked to speak their thoughts aloud (Lewis, 1982) so the researchers could understand their labeling decision-making or correction process. In addition to the observations, we asked them questions about their background, past experiences with labeling, and their perceived relationship with the AI-infused system. The questions were more general in phase I, and we used more specific questions in phase II as a form of member check that elaborated on subtle cues our team noticed during the early interviews.

Data Analysis

We began analyzing the data during data collection which is a common practice in a constant comparative analysis (Glaser & Strauss, 1967). The core team met biweekly to discuss the emerging findings and to report back to the team optimizing the machine learning models of the AI-infused system. During these discussions the interviewers shared their observations and made notes to have others watch the same observations as a form of triangulation. After all the data were collected, the interviews were transcribed and the team engaged in two levels of coding focused around our specific research questions.

First, three different researchers split the dataset and conducted open coding that focused on identifying conversational statements (open codes) related to their labeling task (Charmaz, 2006). That process revealed 1,714 open codes for phase I (labeling data), and 322 open codes for phase II (correcting the machine). Open coding was not specific for the research

TABLE 1 Participant Information for Interviews

ID	Tech Used	Age	Task-Relevant Expertise	Phase II Behavioral Anthropomorphic Score Range (1–3)
01**	PC	52	Emergency manager, Mark (pseudonym)	—
02	iPad	46	Works in IT; experienced annotator	—
04	PC	73	CERT volunteer; no tech experience	—
05	PC	44	Experienced annotator; works in IT; ML; Twitter	—
07	iPad	Late 30s	Experienced annotator; works in IT; NLP; Twitter & social media	—
09	Mac	31	Social media (Facebook)	—
11	PC	68	Social media (Facebook)	—
03*	PC	71	Former emergency manager; no tech experience	Interview 4 Score: 2.00 Interview 5 Score: 2.71
06*	PC	66	NLP experience	Interview 4 Score: 2.20
08*	PC	37	Experienced annotator; data mining; Twitter & social media	Interview 4 Score: 2.80 Interview 5 Score: 3.00ˆ
10*	Mac	70	Experienced annotator	Interview 4 Score: 2.00
12*	Mac	39	Twitter & social media	Interview 4 Score: 2.67
13*	PC	53	Experienced annotator; Twitter & other social media	Interview 4 Score: 1.60
14*	PC	49	Experienced annotator	Interview 4 Score: 2.00

Note. **Indicates CERT leader (interviewed 2 times during phase I). *Indicates participation in both phase I (3 separate interviews) and phase II (2 separate interviews) of the study. ˆIndicates participant thought of themselves as a computer and attributed it to their autism. Abbreviations: IT (Information Technology), CERT (Community Emergency Response Team), ML (Machine Learning), NLP (Natural Language Processing).

questions in this study, but instead captured general statements related to labeling. For example, "This tweet would be confusing to someone in another part of country or world," was an open code. Six months after the open coding process, two researchers (involved in the open coding) engaged in focused coding to identify the overt scripts—the knowledge structures people held that helped them understand how to perform their labeling task. For phase I, we identified 294 focused codes (e.g., "Computer doesn't get emotion like humans") that contained a script, and we categorized those scripts into 16 core categories using a constant comparative analysis. For example, the focused code listed here was identified as a *Value of Humans in Machine Learning* core category script (see Table 2 for all script codes and themes). We combined these core categories into three themes: (1) Scripts invoked to manage decision-making in the labeling/correcting process, (2) Contextual scripts influencing decision-making, and (3) Machine perceptions influencing decision-making scripts.

Next, we analyzed phase II data using focused coding and constant comparative analysis (Charmaz, 2006) and identified 251 focused codes. Although many of these codes fit into the core categories identified in phase I, there were five additional focused codes that were unique to phase II data. One of these codes arose from an additional question that was not asked in phase I: "To what extent do you think of the AI system as a teammate?" As we categorized the data, we sorted each focused code by interviewee and interview number to visually see longitudinal trends, and we wrote memos to capture meaningful observations. See Table 2 on the folllowing page for details around these categories and themes.

Results

We first report our findings about the specific scripts involved as people engaged in iterative supervised machine learning interactions (RQ1). Next, we demonstrate findings suggesting that humans labeling data and providing iterative feedback to machines can be considered a form of HMC (RQ2).

RQ1: Scripts Invoked During Iterative Supervised Machine Learning

Three themes emerged from the analysis that describe the scripts people engaged in when both labeling data for the machine and correcting the machine-inferred labels (see Table 2 on the following page). Scripts invoked to manage decision-making during these processes is the largest theme. People involved in these tasks were constantly making decisions as they were presented with tweets and asked to label/correct each of them. During a 1-hour interview and observation, people were making 30 to 50 of those decisions. The categories of scripts contained within this theme provide a broad overview of the challenges people faced, as well as the coping strategies used to complete their tasks. For example, a common coping strategy was referring to the training program they received. ID #02, interview #3, said,

> When we were first trained to do this, [#01], our virtual leader, had us all on a call and he would bring tweets up and people would go, oh, that's this that's that [as they learned how to label the tweets].

In both phases, the participants acknowledged a high degree of uncertainty and self-doubt in how they were conducting their tasks, but in phase II they specifically acknowledged the difficulty they experienced correcting the machine-inferred labels. This more difficult task appeared to be more cognitively taxing as participants took longer to make decisions, especially when they disagreed with how the machine had labeled the data. Several participants openly acknowledged they were not willing to second-guess the machine, and only one person in the dataset—ID #08, who claimed she thought like a machine—admitted the correction task was easier than the prior labeling tasks. In both phases, participants coped with their decisions by regularly referring to their training, focusing on the project goals, rationalizing incomplete data, and justifying their doubts by reminding themselves that other humans also would be coding the same tweets so errors would be minimized.

TABLE 2 Comparing Phase I and Phase II Script Codes

Script	Phase I *N*	Phase 1 %*	Phase II *N*	Phase II %*
Theme 1: *Scripts invoked to manage decision-making in the labeling/correcting process*				
Referring to training	58	19.73	22	8.8
Focusing on project goals	40	13.61	7	2.8
Rationalizing a lack of complete data (how data is presented)	30	10.20	12	4.8
Justifying doubt because other humans will check their work	20	6.80	12	4.8
Labeling/correcting to help machines learn	26	8.84	7	2.8
Acknowledging the limits of social media data	23	7.82	3	1.2
Dealing with doubt by changing one's mind	12	4.08	0	0
Acknowledging and controlling biases	7	2.38	1	0.40
Acknowledging the difficulty of correcting machine-labeled tweets	0	0	32	13.0
Not willing to second-guess the machine/ conceding	0	0	3	1.2
Acknowledging the ease of correcting machine-labeled tweets**	0	0	1	0.40
Theme 2: *Contextual scripts influencing decision-making*				
Value of humans in machine learning	17	5.78	11	4.8
Beliefs on how people decide to post on social media	11	3.74	1	0.40
Value of machines in machine learning	8	2.72	5	2.0
Cultural/local understanding	5	1.70	15	6.0
Personal expertise brought to the task	15	5.10	19	7.6
Personal learning as desirable in this process	7	2.38	1	0.40
Theme 3: *Machine perceptions influencing decision-making scripts*				
Machine is not learning and this is frustrating to observe	8	2.72	31	12.0
Acknowledging the limits of machines in machine learning	0	0	5	2
Machine is learning and this is exciting to observe	6	2.04	36	14.0
Assigning anthropomorphic qualities to the machine	0	0	38	15.0

Note. *Normalized for comparison. **ID #08 is the only participant who said this in Phase II.

The second theme, contextual beliefs, describes the scripts people drew upon surrounding their own value in working with machines, the value machines bring to the process, personal beliefs around social media, and their cultural and local understanding. Individual scripts, specifically personal expertise and a desire to participate to learn, describe what the participants brought to the labeling and correcting tasks, and what they wanted to get out of participating.

These beliefs were often articulated during the sessions, and they provided insight into contextual variables that might have influenced their labeling and correcting tasks. For example, ID #02, interview #2, articulated his expertise this way: "So probably the most helpful thing is I am in IT myself . . . and knowing how we use data has positioned me to be able to respond thoughtfully to some of [these tweets]."

The third theme, how the perceptions of machines influenced the tasks, was quite different between the two phases. In phase I there were three separate interviews, and the participants knew that the tweets they were given to label should be getting more relevant as the machine learned how to filter out the irrelevant tweets. However, when participants were simply providing labels, they only occasionally mentioned that the machine was either learning or not learning. For example, in phase I, ID #14, interview #2, said, "Hopefully we'll have less garbage, this time." When they were asked to correct the machine, these categories became much more prominent and nuanced. Table 2 demonstrates this trend in numerical form since we summed all the focused codes, normalized them, and compared them. Although participants mentioned that the machine was learning slightly more often than they said the machine was not learning, this is likely not a meaningful difference because most of the participants' comments described when the machine was excelling and when the machine was struggling. For example, many people noticed the machine had trouble labeling sentiment, but that it was showing improvement in identifying risks or prevention activities.

Two new categories emerged during phase II, due in part to the addition of a question that asked the extent to which they viewed the AI system as a teammate. Several participants were quick to acknowledge the machine's limitations, and all participants shared their opinions of what we are calling behavioral anthropomorphism. Only one participant explicitly mentioned human behaviors (e.g., "machine like a toddler," and "a fourth person analyzing data," ID #08, phase II), but all seven of the phase II participants imposed a learning script on the AI system that revealed a form of behavioral anthropomorphism. This means they discussed the AI system's learning process in ways akin to people or animals. For example, participant ID #03 in phase II said, "I have plenty of goodwill toward the computer because it's making an effort. It's learning what we teach it . . . It's really not its fault if it gets it wrong. It's how we train it."

RQ2: Moving From Interaction to Human-Machine Communication

To assess how people viewed their interaction with the machine, we examined the most relevant scripts identified in RQ1 and further analyzed corresponding data. We examined the trends in script pattern changes over time (looking at the number of codes across each of the five interviews in both phase I and phase II), as well as inspected for patterns within each of the seven participants who contributed to phase II (correcting the machine). See

Table 2 for these patterns. We also examined the actual language and re-watched videos to verify what we coded from the transcripts.

Behavioral anthropomorphism. To better understand how the phase II participants varied in their views of behavioral anthropomorphism, two researchers coded each statement in this category according to the degree of behavioral anthropomorphism in the statement. A score of 1 indicated an explicit mention of not being a teammate: "It's a tool" (ID #10). A score of 2 was more mixed, as seen in this comment from ID #03: "I don't think of our computer as a teammate yet. I expect it to become like one. Gotta get up to speed first. Computer's an apprentice . . . it still has its training wheels." When people explicitly mentioned the system being a teammate or a partner, we gave it a score of 3. For example, ID #10 said, "Yes, it is my teammate. I would say it's a very useful and helpful partner." There was one outlier when coding this data: ID #08 not only thought of the AI system as a partner, but she also thought of herself as a machine. She explained, "I'm autistic. And so, when I look at information, I look at it much in the same way as [Hal and Data] do; which is part of why I can [understand] the computer."

We summed the scores for the statements from each participant and divided by the number of total statements to give each person a behavioral anthropomorphism score (see Table 1). For participants who discussed behavioral anthropomorphism in both of their phase II interviews (ID #03 and ID #08), we calculated the score for each interview separately. Only one participant could be characterized as making few comments reflective of behavioral anthropomorphism (ID #13), while all other participants showed higher scores the longer they worked with the AI system. This finding—along with the other scripts—suggests that knowledge structures humans hold around learning can be transferred to machines. This quote from ID #03 demonstrates the learning/training script in reference to a puppy: "I sort of treat it like a puppy that I love that just poops the room. It's like, 'It's not his fault. You need to learn. It's okay.'"

Struggle and self-doubt as an indicator of a relationship. Codes related to self-doubt manifested very differently between phase I and phase II. In phase I, people were new to the labeling task and while they expressed self-doubt, it was because they wanted to do a good job with their task. This is why the scripts findings in RQ1 so clearly explain how they cope with that self-doubt and continue with their tasks. Having to correct the machine in phase II introduced a new form of struggle and doubt not seen in phase I, and there is some evidence suggesting that the participants' relationships with the machine also changed during phase II. Specifically, when participants found a machine-labeled tweet with which they disagreed, they often paused and as they thought aloud, they expressed ambivalence around their decision-making. The script coding findings suggest that people's coping scripts were less frequently invoked, especially references to their training, to the project goals, and to the process of other humans checking their work. This was combined with the increase in explicit mentions of the difficulty of correcting the machine and the frustrations with what the machine was not learning. These are examples of how self-doubt appeared in the data.

> Oddly enough, I'm not as confident in my own coding this go-around as I was in all the earlier sessions. I'm competing, in some sense, with the software . . . I'm a little less certain that, quote-unquote, "I'm right" compared to the

> [machine]. So, before, the context was, "God-darn it, I'm right."—ID #06, interview #4 (phase II)

This new task was also slower, due in part to volunteers second-guessing their own decisions when the machine had made a different decision than the one they would have made. This participant explained:

> I think, for me, it was easier when I wasn't correcting the computer because then I'd be like, "oh, that's risk." And now, when [the AI system] says prevention, I'm like wait, what? Why would they? And then, I start thinking maybe it is risk. But no, it really isn't. When you are checking it, you do question why you're going the way you're going. So, it was quicker [when providing the labels].—ID #13, interview #5 (phase II)

Because the task compelled the volunteers to work at a slower pace, some people became more cognizant of the intricacy and significance of their work. One participant described:

> I don't think I'm the cat's meow at doing this. Obviously, I had to slow down and really think about some of these today. So, I hope I'm doing it justice. It definitely shows me the complexity of what's going on and what needs to be done. I don't think I did great but I hope I did well enough to contribute.—ID #14, interview #5 (phase II)

Relationship with the AI-infused system *aka* machine. The final category explaining how the correcting task indicates a form of human-machine communication is how the relationship developed over time. Participants were not willing to "give up" on the machine and they were actively trying to adjust their expectations to be patient and understanding. One participant described the machine's learning much like how a human learns something new:

> Well, it's learning, baby step by baby step. It's definitely taking some steps. I'd like to see it get more accurate and then I'm hopeful that as it gets more accurate, it can be more helpful . . . I'm not giving up on it and I'm willing to keep working with it. It's like anybody else who is learning something; a person taking their first stumbling steps, and they're getting a little bit better and you keep working with them and get more chances to improve, then they're going to get better.—ID #10, interview #4 (phase II)

One participant took it a step further and compared AI-infused systems to toddlers:

> I think of computers a lot like toddlers. It does only exactly what you told it to do. The computer's not stupid. It's just not trained to do what you want. You either didn't tell it what you wanted it to do, or you told it what you wanted it to do and what it interpreted it to be versus what you wanted are just slightly different.—ID #08, interview #4 (phase II)

Discussion

In responding to calls by Guzman (2018) and Gambino et al. (2020), we take an inductive approach to identify specific scripts humans are using as they interact with machine learning algorithms. We find that as people interact with and provide feedback to a machine that is actively engaged in learning, they can view the machine as a social actor. They are not mindlessly interacting with machines like CASA (Nass & Moon, 2000) proposed, but instead the processes of training and providing feedback to the machine and working with it over time can provide mechanisms through which machines become social actors and valued teaming partners. Thus, our study extends CASA as well as introduces a boundary through which human-machine interaction can be considered a form of human-machine communication where human and machine construct meaning together.

Anthropomorphism—people's perceptions that machines have human qualities—is important in human-machine communication because these perceptions are often linked to a machine's social potential (Nass & Moon, 2000). The anthropomorphism seen in the words of some interviewees only occasionally originates from a human-like physical trait, but instead is grounded in the fundamental human action of training one another, a form of behavioral anthropomorphism. Therefore, in this study, the AI-infused machine exhibited behavioral anthropomorphism, which extends the concept of anthropomorphism into the space where machines behaviorally *respond* in ways we expect of humans (Nowak & Fox, 2018). Responses such as demonstrating the machine was learning were clearly present in this study. The people in this study were not directly talking to a machine, like human-to-human communication; instead, by reinforcing and correcting the machine while also feeling pride, shame, and frustration (as observed and documented by the researchers who met weekly to reflect on observations), they reveal how they communicate through exchanging information and feeling emotions as part of the learning process. In this way our findings support thinking of AI systems as media agents: technologies capable of generating enough social cues for humans to perceive them as capable of interaction (Gambino et al., 2020), including teaming and even communication.

Indicators of Communication With the AI-Infused System

While considerable research has recently been conducted on human-AI teaming where the "AI" system is a conversational agent (e.g., Shaikh & Cruz, 2022), the AI system in this study is not conversational. However, the participants' responses indicate there could be a relational aspect to their interaction. For example, interviewees struggled and doubted their abilities when presented with and asked to correct labels inferred by the machine/AI system. This was a different form of doubt than seen when they were labeling data to help the machine learn; correcting was cognitively taxing, slowed people down, and made them question their own interpretations. While it is likely that this is a more difficult task, their reactions suggested more than just an increased challenge. They expressed many more emotions and verbal indicators of ambivalence when they were confronted with the possibility that the machine was not aligned with their thinking, especially because they had provided the feedback to help the machine learn. Yet they did not want to place all the blame on the computer, and sometimes questioned whether they were the teammate letting the computer down.

The doubt alone is not enough evidence to argue that communication is actually occurring, but when combining that with the behavioral anthropomorphism scripts people used to describe their relationship with the AI system, the evidence builds. The data suggest that acts of training a machine, providing feedback, and assessing its learning map cleanly onto the scripts people use when teaching a toddler, a dog, and even a conversational agent (Hal—a fictional AI character from Arthur C. Clarke's *Space Odyssey* series—and Data—a male android appearing in *Star Trek*—were mentioned by one participant). They express both frustration that the machine is not learning as quickly as they hoped, as well as excitement when their "friend and teammate the computer" (ID #03, interview #05) shows it is learning. Once again there is an emotional component to the discussion around training the machine—a sense of meaning being created within the relationship (Guzman, 2018). Thus, the ontological boundaries people use to assess their relationship appear more social than simply technical (Guzman, 2020).

For each of the 14 participants in this study, their script data suggest they view the AI system as a complement to their human capabilities. They feel a sense of responsibility for training the machine because they want it to be a teammate as they work toward the overarching goal of the project. These findings support Gibbs et al.'s (2021) claim that machines complement human capabilities so they can be considered a team member to the human. This suggests that our findings here should be relevant well beyond the specific context we studied. For example, the use of crowdsourced labor to provide labels as input to machine learning algorithms is a widespread practice for developing AI-infused systems. If crowdworkers could be made to better understand how their labeling actions were helping and training the AI-infused system, the crowdworkers might be more inclined to think of the machine as a teammate. In turn, this might bring about more feelings of responsibility for training the AI-infused system that is important to minimize biases in such systems. It could further bring more investment of sincere efforts from the crowdworkers in achieving the outcomes of the project, especially if done in a voluntary crowdsourcing setting.

While it is plausible that human-machine communication is occurring between people and the supervised machine learning algorithms they are training and correcting, we cannot state for certain if the communicative aspects of their interactions emerged because of the change in task (having to correct the machine), or because the participants worked with the machine over time. All participants in phase II were well aware that the machine was using their input to learn how to label data, and it is possible they would have made similar comments even if they had never been asked to correct the machine. Nonetheless, the two aforementioned points substantiate our argument that the correcting task is what triggered the complexity of emotions and feeling that the machine was a teammate. Interestingly, six of the seven participants in phase II had prior experience providing labels to machines for machine learning. Even these participants expressed greater behavioral anthropomorphism over time, which suggests they did not come to the current task with this belief. Future studies should investigate this possibility to verify our claim.

Limitations and Future Directions

This study includes a very specific population of volunteers who are CERT members located in a specific geographic area, and while they do vary widely in their ages and types of experience, their training as a CERT member makes them different from the rest of the population. It would also have been better to provide all participants with the opportunity to participate in phase II, but considering the quick timeline in which we needed to train the machines, that was not possible.

Experiments to determine causal relationships. Future research using careful experimental design can extend the theorizing generated in this paper. None of the research team members were cognizant of the difficulty level in correcting machine-inferred labels on data until they systematically coded and categorized the data as a whole. Future experiments could randomly assign people to conditions of either labeling data or correcting the AI-inferred labels of data and provide self-report measures to better understand the sources of doubt and how they are related to the emotions people feel when working with an AI system. Experiments could also test whether similar results occur when humans correct other humans and when machines correct the labeling of the humans. Interrogating the relationship between machines and humans and how they correct and help one another learn could further explain the findings generated in the current paper.

Exploring how the machine might help support the volunteers. Considering that the tasks asked of these volunteers were cognitively taxing, and that they likely experienced some forms of decision overload, future research should explore how the machine might help the volunteers by supporting them through their tasks. For example, the machine might be designed to provide supportive or encouraging messages in the middle of the individual labeling sessions. The machine might also serve important feedback and quality control purposes. For example, the machine could remind the volunteers about the definitions of the specific labels and ask them to stop and check their work. There could also be helpful forms of feedback integrated into the system that could provide reinforcing practice to help motivate high-quality work.

Conclusion

This study examined the scripts people use when working with machines. These scripts provide evidence that human-machine communication is possible when people are engaged in supervised machine learning tasks. Therefore, the major contribution of this study is identifying the ontological boundaries between interaction and human-machine communication. Specifically, when people want to teach the machine, provide corrective feedback, observe success in the machine learning, and experience emotions, they are also more likely to view their interactions as teaming; thus, human-AI teaming is a form of human-machine communication. This suggests that human-machine communication demands more of a relationship than human-machine interaction does, which could be important when considering how to motivate people to do this kind of work over an extended period of time.

Author Biographies

Keri K. Stephens, PhD, is a Professor in Organizational Communication Technology and Co-Director of Technology, Information, & Policy Institute in the Moody College of Communication at The University of Texas at Austin. Her research program examines the role of technology in organizational practices and organizing processes, especially in contexts of crisis, disaster, and health. She has received over $6 million in external funding and authored over 100 articles and book chapters. Her recent books include the award-winning book *New Media in Times of Crisis* (2019, Routledge), and the two-time, award-winning book *Negotiating Control: Organizations and Mobile Communication* (2018, Oxford University Press).

https://orcid.org/0000-0002-9526-2331

Anastazja G. Harris (MA, The University of Texas at Austin) is a doctoral candidate in Communication Studies at The University of Texas at Austin. Her research primarily focuses on organizational communication, human-AI decision-making, and team communication.

https://orcid.org/0000-0003-3756-1453

Amanda L. Hughes, PhD, is an Associate Professor of Computer Science at Brigham Young University where she directs the Crisis Informatics Lab. As a recognized research leader in Crisis Informatics, her work investigates the use of information and communication technology during crises and mass emergencies with particular attention to how social media affect emergency response organizations. She has published more than 55 articles and book chapters and is highly cited in her field. Her research is currently funded by grants from NSF, NASA SERVIR, and the Knight Foundation.

https://orcid.org/0000-0002-7506-3343

Carolyn E. Montagnolo (MA, The University of Texas at Austin) is a doctoral student in Communication Studies at The University of Texas at Austin. Her research primarily focuses on language attitudes and message effects in political and health contexts.

https://orcid.org/0000-0001-9067-943X

Karim Nader is a PhD candidate in Philosophy at The University of Texas at Austin. His research examines the ways in which technology reflects and affects human values. Specifically, he looks at moral values in video games and virtual reality, sexual and romantic values on dating apps, and epistemic values on the internet. His dissertation argues that our actions in virtual reality should be morally evaluated based on the fictional representations that they create.

https://orcid.org/0000-0003-3571-1796

Ashley Stevens, MA, is a UX designer for FamilySearch. She worked on the research project for this paper while completing her master's degree in Technology at Brigham Young University. Her research primarily focuses on how volunteers label social media data as input into machine learning systems as well as the usability of the labeling systems they use.

Tara Tasuji, PhD, is a Research Scientist at the Technology & Information Policy Institute (TIPI) in the Moody College of Communication at The University of Texas at Austin. She is a mixed-methods researcher with expertise in fieldwork as well as developmental and social psychology. Tara works on TIPI projects related to disasters, resilience, risk communication, and human-AI teaming. She holds a bachelor's degree in psychology from The University of Texas at Austin, master's degrees in human development, child development, and education from Harvard and Oxford, and a PhD in cognitive and evolutionary anthropology from Oxford.

https://orcid.org/0000-0003-1054-5218

Yifan Xu, MS, is a doctoral student in Organizational Communication and Technology at The University of Texas at Austin. Her research primarily focuses on technology design and development practices and inquires about how we can build technologies that are supportive of inclusive and ethical outcomes.

https://orcid.org/0000-0001-5239-4951

Hemant Purohit, PhD, is an associate professor of Information Sciences and Technology at George Mason University. He researches the design of interactive intelligent/AI systems to support and augment human work capabilities for real-time processing and management of nontraditional data sources like social media at public services. His research has received many honors, including the 2014 ITU Young-Innovator award from the United Nations agency (ITU) and the NSF CRII award. He serves for multiple editorial boards, including Elsevier Journal of Information Processing & Management and ACM Transactions on the Web. Several grants from state, national, and international agencies support his research.

https://orcid.org/0000-0002-4573-8450

Christopher W. Zobel, PhD, is the R.B. Pamplin Professor of Business Information Technology in the Pamplin College of Business at Virginia Tech. His primary research interests include disaster operations management and humanitarian supply chain resilience. He has published over 100 articles in archival journals and academic conference proceedings, and his work can be found in journal outlets such as the *Journal of Operations Management, Production and Operations Management, Risk Analysis, Decision Sciences*, and the *European Journal of Operational Research*. He is currently Co-PI on several different US National Science Foundation grants that involve characterizing and quantifying multidimensional disaster resilience.

https://orcid.org/0000-0002-0952-7322

Acknowledgments

We would like to thank the Community Emergency Response Team Members for participating in this research. We would also like to thank our CIVIC Partner, Steve Peterson, who was an integral member of our research and implementation team. Finally, we thank the reviewers of this manuscript for their careful and thoughtful comments that enhanced this work.

Funding: This work was supported by two grants from the National Science Foundation: Award # 2029692, *RAPID/Collaborative Research: Human-AI Teaming for Big Data Analytics to Enhance Response to the COVID-19 Pandemic;* and Award #2043522, *SCC-CIVIC-PG Track B: Assessing the Feasibility of Systematizing Human-AI Teaming to Improve Community Resilience.* Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Abelson, R. P. (1976). Script processing in attitude formation and decision making. In J. S. Carroll & J. W. Payne (Eds.), *Cognition and social behavior* (pp. 33–45). Erlbaum.

Alam, S. L., & Campbell, J. (2017). Temporal motivations of volunteers to participate in cultural crowdsourcing work. *Information Systems Research*, *28*(4), 744–759. https://doi.org/10.1287/isre.2017.0719

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the people: The role of humans in interactive machine learning. *AI Magazine*, *35*(4), 105–120. https://doi.org/10.1609/aimag.v35i4.2513

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.

Dehnert, M., & Leach, R. (2021). Becoming human? Ableism and control in Detroit: Become human and the implications for human-machine communication. *Human-Machine Communication*, *2*(1), 137–152. https://doi.org/10.30658/hmc.2.7

Fathi, R., Thom, D., Koch, S., Ertl, T., & Fiedrich, F. (2019). VOST: A case study in voluntary digital participation for collaborative emergency management. *Information Processing & Management*, *57*(4), 1–25. https://doi.org/10.1016/j.ipm.2019.102174

FEMA. (2022). *Community Emergency Response Team*. https://web.archive.org/web/20221223224731/https://www.fema.gov/emergency-managers/individuals-communities/preparedness-activities-webinars/community-emergency-response-team

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

Gibbs, J., Kirkwood, G., Fang, C., & Wilkenfeld, J. (2021). Negotiating agency and control: Theorizing human-machine communication from a structurational perspective. *Human-Machine Communication*, *2*(1), 153–171. https://doi.org/10.30658/hmc.2.8

Gioia, D. A., & Poole, P. P. (1984). Scripts in organizational behavior. *Academy of Management. The Academy of Management Review (Pre-1986)*, *9*(000003), 449–459. https://doi.org/10.2307/258285

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Transaction.

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

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, *1*(1), 37–54. https://doi.org/10.30658/hmc.1.3

Harrison, S., Tatar, D., & Sengers, P. (2007). The three paradigms of HCI. *Alt. Chi. Session at the SIGCHI Conference on Human Factors in Computing Systems San Jose, California, USA*, 1–18. https://www.scinapse.io/papers/47513853

Hughes, A. L., & Tapia, A. H. (2015). Social media in crisis: When professional responders meet digital volunteers. *Journal of Homeland Security & Emergency Management*, *12*(3), 679–706. https://doi.org/10.1515/jhsem-2014-0080

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), 67:1–67:38. https://doi.org/10.1145/2771588

Karuna, P., Rana, M., and Purohit, H. (2017). CitizenHelper: A streaming analytics system to mine citizen and web data for humanitarian organizations. *Proceedings of the Eleventh International Conference on Web and Social Media*, Montréal, Québec, Canada, 729–730. https://doi.org/10.1609/icwsm.v11i1.14863

Latour, B. (1994). Pragmatogonies: A mythical account of how humans and nonhumans swap properties. *American Behavioral Scientist*, *37*(6), 791–808. https://doi.org/10.1177/0002764294037006006

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 1–16. https://doi.org/10.1177/2053951718756684

Lewis, C. (1982). *Using the thinking-aloud method in cognitive interface design*. IBM T. J. Watson Research Center.

Madni, A. M., & Madni, C. C. (2018). Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems*, *6*(4), 44. https://doi.org/10.3390/systems6040044

Malone, T. W. (2018). *Superminds: The surprising power of people and computers thinking together*. Little, Brown Spark.

McKinsey. (2021, December 8). *The State of AI in 2021: Survey*. https://web.archive.org/web/20220605065759/https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Monarch, R. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Manning.

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

Nowak, K. L., & Fox, J. (2018). Avatars and computer-mediated communication: A review of the uses and effects of virtual representations. *Review of Communication Research*, *6*, 30–53. https://doi.org/10.12840/issn.2255-4165.2018.06.01.015

Purohit, H., Castillo, C., Imran, M., & Pandev, R. (2018). Social-EOC: Serviceability model to rank social media requests for emergency operation centers. *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 119–126. https://doi.org/10.1109/ASONAM.2018.8508709

Rogers, Y. (2012). HCI theory: Classical, modern, and contemporary. *Synthesis Lectures on Human-Centered Informatics*, *5*(2), 1–129. https://doi.org/10.2200/S00418ED1V01Y201205HCI014

Russell, S., & Norvig, P. (2009). *Artificial intelligence: A modern approach*. Prentice Hall.

Shaikh, S. J., & Cruz, I. F. (2022). AI in human teams: Effects on technology use, members' interactions, and creative performance under time scarcity. *AI & Society*. https://doi.org/10.1007/s00146-021-01335-5

Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.

Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Computers in Human Behavior*, *90*(1), 285–287. https://doi.org/10.1016/j.chb.2018.09.014

Starbird, K., & Palen, L. (2011). "Voluntweeters": Self-organizing by digital volunteers in times of crisis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1071–1080. https://doi.org/10.1145/1978942.1979102

St. Denis, L. A., Hughes, A. L., Diaz, J., Solvik, K., Joseph, M. B., & Balch, J. K. (2020). "What I need to know is what I don't know!": Filtering disaster Twitter data for information from local individuals. *Proceedings of the Information Systems for Crisis Response and Management Conference (ISCRAM 2020)*. http://idl.iscram.org/files/liseannstdenis/2020/2267_LiseAnnSt.Denis_etal2020.pdf

Stephens, K. K., Heller, A., & Chan, A. Y. (2014). Understanding situated survey refusal: Applying sensemaking and sensegiving to ethnostatistics. *Qualitative Research*, *14*(6), 745–762. https://doi.org/10.1177/1468794113495036

Stephens, K. K., Nader, K., Harris, A. G., Montagnolo, C., Hughes, A. L., Jarvis, S. A., Senarath, Y., & Purohit, H. (2021). Online-computer-mediated interviews and observations: Overcoming challenges and establishing best practices in a human-AI teaming context (pp. 2896–2905). In T. Bui's (Ed.), *Proceedings of the 54rd Annual Hawaii International Conference on Social Systems,* Computer Society Press. http://hdl.handle.net/10125/70967

Utz, S., Wolfers, L., & Göritz, A. (2021). The effects of situational and individual factors on algorithm acceptance in COVID-19-related decision-making: A preregistered online experiment. *Human-Machine Communication*, *3*(1), 27–45. https://doi.org/10.30658/hmc.3.3

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.6

An Interactional Account of Empathy in Human-Machine Communication

Shauna Concannon[1,2], **Ian Roberts**[3], **and Marcus Tomalin**[1]

1 Centre for Research in the Arts, Social Sciences and Humanities (CRASSH), University of Cambridge, Cambridge, England, United Kingdom
2 Department of Computer Science, University of Durham, Durham, England, United Kingdom
3 Department of Theoretical and Applied Linguistics, University of Cambridge, Cambridge, England, United Kingdom

Abstract

Efforts to develop *empathetic agents,* or systems capable of responding appropriately to emotional content, have increased as the deployment of such systems in socially complex scenarios becomes more commonplace. In the context of human-machine communication (HMC), the ability to create the perception of empathy is achieved in large part through linguistic behavior. However, studies of how language is used to display and respond to emotion in ways deemed empathetic are limited. This article aims to address this gap, demonstrating how an interactional linguistics informed methodological approach can be applied to the study of empathy in HMC. We present an analysis of empathetic response strategies in HMC and examine how these diverge from the practices employed in human-human dialogue. The specific challenges encountered by current systems are reviewed and their implications for future work on HMC considered.

Keywords: conversational agents, emotion, empathy, human-machine communication, interactional analysis

This research is funded by the Humanities and Social Change International Foundation.

CONTACT Shauna Concannon • Department of Computer Science • Durham University, Durham, DH1 3LE, UK • shauna.j.concannon@durham.ac.uk

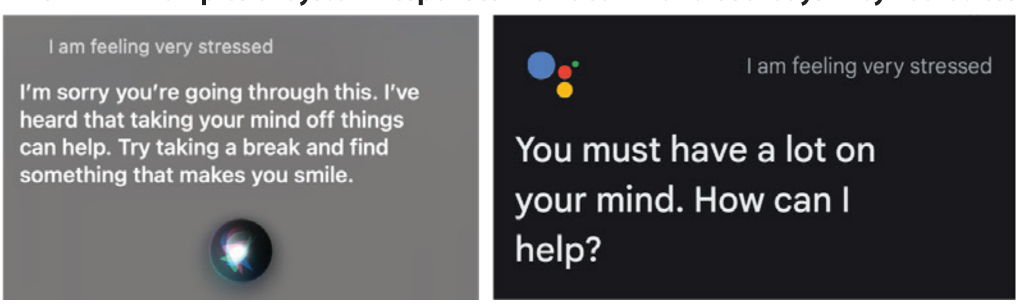
ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

People increasingly interact with autonomous intelligent systems through conversational interfaces. For example, virtual assistants on mobile phones and smart speakers enable users to control smart appliances or request information using voice commands; text-based customer service chatbots are common applications on company websites; and sophisticated systems like ChatGPT have recently attracted millions of users worldwide. Beyond helping individuals to accomplish commercially focused, or functional tasks, these conversational artificially intelligent (CAI) systems are also developed to perform social and affective functions. Several mobile applications that incorporate text-based chatbots to support positive mental health and well-being such as Woebot and Wysa are currently available and already widely used. Others are designed with the aim of providing companionship, such as Replika and Microsoft's chatbot Xiaoice. The latter has been explicitly designed to be "an AI companion with which users form long-term, emotional connections," and is described as an *empathetic social chatbot* (Zhou et al., 2020, p. 54). However, Concannon & Tomalin (2023) interrogate this, highlighting that psychological conceptualizations of empathy are not reflected in the system architecture nor the evaluation metrics used to assess how empathetic the system is.

Even more task-oriented applications, such as virtual assistants like Siri, Alexa, and Google Home, offer replies that attend to an emotional state expressed by the user. Figure 1 shows the responses given by Google Home and Siri when a user expresses that they are feeling stressed (generated on February 6, 2023). The reply issued by Siri expresses sympathy and includes some advice, while the Google Home response provides an account of the user's state of mind and requests directions on how to help. As these examples demonstrate, many modern CAI systems may be required to provide responses to a wide variety of prompts and distinct conversational contexts. In certain scenarios, the ability to interpret emotional cues in a user's utterance may improve their experience or prove necessary for creating a safe system (e.g., if a user discloses information about self-harm). Furthermore, research has demonstrated that the positive effects associated with affective interactions in human-human communication (HHC) are also experienced when interacting with CAI systems (e.g., Ho et al., 2018). Consequently, in an effort to advance the potential uses and applications of CAI technologies, increasing attention has focused on creating *emotionally aware* (e.g., Pamungkas, 2019) and *empathetic agents* (e.g., Zhou et al., 2020).

FIGURE 1 Examples of System Responses Provided When a User Says They Feel Stressed



(a) Response from Siri (b) Response from Google Home

Despite the evident interest in developing systems that can communicate in ways that are *perceived* as empathetic (Concannon & Tomalin, 2023), current CAI systems do not possess anything like human empathy. Unlike work in Affective Developmental Robotics (ADR; e.g., Asada, 2015) that seeks to replicate human affective developmental processes via synthetic means to develop an artificial form of acquired empathy, current CAI systems aim to communicate in ways that are recognized as attending appropriately to the emotional expressiveness of an utterance. For HMC, rather than individualistic notions of empathy as a psychological trait, a focus on *relational* empathy is more relevant. Silverman & Peräkylä (1990) describe this relational focus as: "an approach to empathy less as the psychological propensity to attune to the private meanings of the [individual], but more as the *social* ability to pick up behavioral and cultural cues present in what the [individual] is saying and doing" (p. 312).

Although CAI systems cannot *feel* or *be* empathic, they can potentially use language to create the perception of empathetic concern (Concannon & Tomalin, 2023), but what constitutes empathetic communication (i.e., the specific communicative strategies) is less clearly defined. In the context of text- and speech-based HMC the ability to create the perception of empathy is achieved in large part through linguistic behavior (although it is more pronounced in text-based interactions which preclude the use of pitch, amplitude, intonation contour, and the like). However, interactional accounts of how language is used in ways that are perceived as empathetic have been limited.

Approaches from linguistics and social interaction have been usefully applied to the study of empathy as it manifests between humans, but how fitting are these approaches for the study of HMC? Clearly, these are not equivalent conversational contexts, and the essentially intersubjective dimensions of empathetic interactions further emphasize this. Research on HMC has highlighted that the positioning of agents in social support roles requires a nuanced treatment (Beattie & High, 2022). The perceived understanding framework suggests that how a person relates to a conversational partner is influenced by their emotional capacity to understand, but findings from empirical studies highlight this is not consistently reflected in HMC (Ho et al., 2018). Consequently, understanding the impact of this interactional asymmetry on empathetic interactions in HMC is of great importance.

To extend the current understanding of how empathy functions in HMC, there is a need to develop methodologies that apply fine-grained analyses to these interactions to generate the necessary knowledge about the particular empathetic strategies used by current systems, and to explore their effects. To begin unpacking these issues, we argue that it is necessary to first consider the socially situated practices involved in communicating empathetically. To achieve this, we propose an interactional linguistic informed approach that focuses on the communicative processes and functions associated with empathy in HMC. Specifically, we investigate the following research questions:

RQ1: What strategies do current systems use to respond to empathetic opportunities?

RQ2: How do these compare to strategies employed in HHC?

RQ3: What are the consequences of these strategies on the interaction?

To answer these questions, we conduct an analysis of HMC, drawing on evidence from the interaction data itself to illustrate some of the specific challenges that arise. We propose a methodological approach to identify empathetic strategies in current CAI systems. Additionally, we demonstrate how future research on empathy in HMC can benefit from integrating insights from interactional linguistics, while also contending with the particular differences that emerge due to the specific relational positioning of the human-machine dynamic. A key contribution of this paper is the proposition of a methodological approach for analyzing empathetic strategies in current CAI systems. Through the application of this interactionally-focused qualitative approach, insights into how well empathetic strategies from HHC translate to the HMC context are explored.

In the following sections we begin by presenting the results of a non-exhaustive literature review. Drawing together theoretical perspectives on empathy from psychology, linguistics, and social action research relating to HHC, we motivate the adoption of a relational approach to empathy for the study of HMC. In addition, a review of existing research that addresses the design, implementation, and evaluation of empathetic strategies in CAI systems is presented. The methodological approach is then introduced, together with an empirical study of empathetic strategies in human-machine interaction transcripts. Finally, the implications for future work on empathy in HMC are discussed.

Empathy: Definitions and Approaches

Empathy is a key component in the management of positive social interactions between humans, but specific definitions of the concept and associated phenomena have varied conspicuously (Cuff et al., 2016; Hall & Schwartz, 2019). In the broadest sense, empathy is commonly associated with the ability to understand the emotions, viewpoints, feelings, or attitudes of another. For Batson et al. (2015) empathic concern includes a "whole constellation" of emotions, including "feelings of sympathy, compassion, softheartedness, tenderness, sorrow, sadness, upset, distress, concern, and grief" (p. 260). This framing of empathy includes a range of emotional states that may be experienced in response to the experiences of others. Conversely, for Hojat (2016) empathy is conceptualized as distinct from sympathy, being a "predominantly cognitive (rather than an affective or emotional) attribute that involves an understanding (rather than feeling) of experiences, concerns, and perspectives" of another (p. 74). A common distinction is made between affective empathy—that is, an affective state or response associated with the vicarious experiencing of another's feelings (e.g., Batson et al., 1991; Hoffman, 2001; Stotland, 1969)—and cognitive empathy, which involves identifying and understanding the perspective of another's affective state without sharing in it. Both interpretations, however, position empathy as a primarily individual, mental process.

This individualistic conceptualization is reflected in the numerous metrics that attempt to measure individuals' *empathetic capacity* using questionnaires (e.g., Hogan, 1969; Mehrabian & Epstein, 1972). Others have approached it from the vantage point of the person being empathized with, shifting the focus to include the experiences of the patient. The Consultation and Relational Empathy (CARE) Measure is thus designed to elicit feedback from patients to make an assessment of the degree of perceived empathy observed during an interaction with their clinician (Mercer et al., 2004). However, as van Dijke et al. (2020) highlight, such approaches rely: "solely on the skills and activities of the empathizer. The empathee's role, if acknowledged at all, is restricted to either confirming or rejecting the accuracy of the empathizer's understanding" (van Dijke et al., 2020, p. 4). While conceptualizations of empathy as a primarily individual process may acknowledge that it is made manifest in behaviors and communicative style, others argue that it is more fundamentally relational. Phenomenological accounts suggest that such approaches neglect the role that interaction plays in empathy (Breyer, 2020) and recent work within psychology has also begun to examine the relational dimensions of empathy, re-focusing attention on the dynamic and interactional processes by which empathy is *achieved* through communication (e.g., Betzler, 2019; Main et al., 2017; van Dijke et al., 2020). Main et al. (2017) explain empathy is "neither an instantaneous phenomenon nor a static personality trait," but rather a "dynamic process" (p. 362), with primarily interpersonal functions.

In this paper, similarly, a definition is adopted that acknowledges empathy as a collaborative practice involving participation by both parties (van Dijke et al., 2020). Assessments which focus on empathy as an individual trait fail to acknowledge how empathy is enacted and how these dynamic processes shape social interactions. Rather than conceptualizations of empathy as an internal psychological process, which has more routinely been adopted, the focus here is on "a process-focused approach emphasizing the relational functions of empathy in interpersonal contexts" (Main et al., 2017, p. 358). One way to study empathy in a process-focused way is to attend to how individuals use language to relate to one another and coordinate their actions, as is the focus in interactional linguistics and social interaction studies. Although there are relatively few existing studies that focus primarily on the linguistic phenomena associated with empathy, work within the subdisciplines of discursive psychology, conversational analysis, and functional applied linguistics provides a useful starting point.

Discursive psychology, conversational analysis, and functional applied linguistics, all approach linguistic expressions of empathy from an interactionally-focused perspective. Rather than viewing a speaker's utterance as a direct reflection of their *inner mind*, it is viewed as a socially situated practice that serves a primarily interactional function. The study of surprise presented by Wilkinson & Kitzinger (2006) neatly evidences this. Rather than a spontaneous eruption, surprise in conversation is produced with sensitivity to timing and sequence, and Majid (2012) concludes that surprise is performed by both parties. A surprise response will not be produced instantaneously upon receiving the information, rather the respondent will delay until the speaker has completed their turn. Thus, surprise is an interactional achievement (i.e., something that is jointly constructed by both conversational partners).

Approaching empathy relationally, as an interactional achievement, involves examining the observable ways that speakers display and orient themselves toward emotional content in naturally occurring conversation, or talk-in-action. M. H. Goodwin et al. (2012) view displays of emotion as: "an interactive, dialogic action rather than the expression of something internal to a single individual" (p. 23). They emphasize the need for an analytic framework that shifts beyond the individual to include the empathizer and empathizee, but also attends to the specifics of how the interaction is structured. Consequently, understanding empathetic communication requires the examination of "sequences in which one party is responding to, or in some other way performing operations on, actions produced by another" (M. H. Goodwin et al., 2012, p. 25).

Kupetz (2020) suggests a conversational analytic perspective as complementary to the relational approach to empathy outlined by Main & Kho (2020). Following Couper-Kuhlen (2009), they advocate research that starts from what is observable in the interaction itself and highlight how approaches from conversational analysis, interactional linguistics, and discursive psychology are well-positioned to do this. In the following section, work that takes an interactionally-driven approach to study how empathy manifests in conversation between humans is reviewed.

Empathy in Human-Human Dialogues: Linguistic Strategies for Expressing Empathy

Empirical studies demonstrate the variety of ways that individuals display empathetic behaviors. Interlocutors attend to subtle linguistic cues and carefully negotiate when and how to make assessments about experiences and emotions (Hepburn & Potter, 2007). A. L. Suchman et al. (1997) examine empathetic interactional sequences in doctor-patient meetings and found that empathetic opportunities rarely involved direct expressions of emotions. Rather than *explicit, implicit empathetic opportunities* were favored, with patients referencing an event or context from which an emotional state could be inferred. If doctors were perceptive to these implicit cues and invited elaboration, direct expression was more likely to follow. Effective empathetic responses acknowledge the emotion or invite elaboration, for instance through a follow-up question. Pounds (2011) details specific linguistic constructions used in these empathetic interactions. Less effective responses, or *terminators*, directed the conversation away from the stated or implied emotion.

Demonstrating attentiveness, understanding, and curiosity, and producing affiliate responses are key practices in empathetic communication between humans. Continued attention can be variously signaled through acknowledgment tokens or continuers, also referred to as back channels (e.g., "uhuh," "yeh," "'mm"), producing a relevant next turn or even the use of silence or pause to enable the speaker to continue or elaborate further. Repeating back a speaker's own words can provide evidence of one's own participation in the experiences of the interlocutor (Tannen, 1987). Paraphrasing sequences, also referred to as formulations (Heritage & Watson, 1979), are important displays and checks of understanding and sites where key semantic work is done by participants to establish mutual understanding (Deppermann, 2011). Different conversational contexts may influence how paraphrasing is incorporated into the dialogue. In a conversation between friends, Kupetz (2014) observed formulations were constructed to emphasize salient emotive components, while in dispute resolution contexts, emotive aspects of narrative formulations were minimized to progress the wider conversational goal of dispute resolution (Atkinson, 1992). Therefore, while paraphrasing events and emotions can demonstrate understanding, the specific formulation can result in different empathetic effects.

Mental state formulations are another form of *empathetic receipt* that explicitly labels the perceived emotion (e.g., "you sound upset"). These are often "built from local features of the caller's talk (displays and metaformulations of upset)" and forefront the epistemic basis for a claim to interpret another person's emotional state (Hepburn & Potter, 2007, p. 110). Empathizing, and specifically laying claim to know another's mental state, can involve the delicate management of the relative epistemic positioning of speakers. Marking the epistemic contingency of a claim (e.g., "you said you felt angry"), provides an account of the basis for a claim. Similarly, tag questions (e.g., "you're worried, aren't you?") enable individuals to formulate their utterances as a question rather than an assertion, to defer to another speaker's epistemic rights regarding the topic under discussion. They are commonly used in assessment sequences to signal a lack of certainty and reduce the claim to accuracy. In the terminology of Heritage & Raymond (2005), tag questions can constitute a form of *epistemic downgrade*. Other features that may be used to downgrade the epistemic claims made about another person's mental state include hedges (e.g., "maybe," "sort of"), and discourse markers which index that an utterance is building upon what has come before or information previously provided (e.g., "so," "then").

Consider the excerpt presented in Figure 2 taken from a conversation between a Child Protection Officer (CPO) and a caller in Hepburn & Potter (2007). In lines 4–5 the CPO uses the tag question "she's had a really difficult time hasn't she." The tag question in this example demonstrates affiliation and projects agreement, while also downgrading the epistemic status of their assessment of the caller's friend.

In line 9, the CPO makes clear what they are basing their assessment on by marking the epistemic contingency of the mental state formulations "you *sound* as though you're very upset" (emphasis added). As Hepburn & Potter (2007) explain:

> saying things about the mental states of others is a potentially delicate thing to do. There may be issues of privacy and ownership—who has the right to such claims and also issues of epistemology and who is in the best position to access the knowledge on which such claims can be based. (p. 104)

The epistemic dimensions of claiming to know how a person feels are further complicated when one participant is a CAI system.

FIGURE 2 Excerpt from Extract 8: JX Self-harming friend in Hepburn & Potter (2007)

1	CPO:	.Hh because there's ↑lots of things that
2		could be done to help your frie:nd,
3		(0.5)
4	CPO:	.hhh Because obviously she'll- (0.2) she's
5		had a really difficult ↑ti:hme.=hasn't she.
6		(0.7)
7	Caller:	Yeah.
8		(.)
9	CPO:	°Yeh.° You sound as though you're very upset
10		about it.
11	Caller:	.Shih ~yeh I am.~

Another key practice associated with empathy in interaction is demonstrating affiliation (Couper-Kuhlen, 2012; Hepburn & Potter, 2007; Ruusuvuori, 2005). Affiliation can be demonstrated via anticipatory completion and endorsements. Weatherall & Stubbe (2015) explain that "[i]n everyday interaction the preferred response to a complaint is an endorsement of the speaker's emotional stance" (p. 280). Ways of endorsing include claims of understanding and making similar assessments (e.g., "losing a pet is awful, our dog died last year"). However, these can be construed as competitive, if they include too much detail or affective emphasis, or ingenuous or pro forma if not drawn from direct firsthand experience (Heritage, 2011).

In summary, studies of how emotion is displayed and responded to in naturally occurring dialogues, or *talk-in-interaction* between humans, demonstrate that empathetic interactions are complexly negotiated, socially-situated, and collaborative processes. These fine-grained analyses facilitate a deeper understanding of the relational components of empathy and offer insights into the practical ways that emotion talk is realized and the different forms of empathetic behavior enacted.

Design Approaches to Achieve Perceived Empathy in CAI Systems

As demonstrated above, empathetic communication strategies in HHC are collaborative. The act of empathizing can take many forms; it can involve distinct components of empathy, and it can be achieved through a variety of linguistic means (Urakami et al., 2019). However, developing systems that support such flexibility is technically challenging. So, how does this compare to attempts to create perceived empathy in existing CAI systems? While behaviors associated with empathy may be an important dimension of HMC (e.g., for trust or particular conversational contexts), it is unclear how to design dialogue that will be perceived as empathetic in the HMC context. In this section, we review approaches taken to implement empathetic strategies in CAI systems, both in terms of technical implementation and chatbot design.

Developing an *empathetic agent* first requires the operationalization of empathetic communication—that is, determining the components of empathetic interaction that the agent must participate in, in order to be perceived as being empathetic. The most dominant approach taken when developing CAI systems involves two separate sub-tasks: (i) recognizing and interpreting emotional content and (ii) generating an appropriate response that attends to the emotion displayed in the user utterance (e.g., Rashkin et al., 2019). This reflects the two key features of empathy in clinical interactions identified by A. L. Suchman et al. (1997): "the accurate understanding of the patient's feelings by the clinician and the effective communication of that understanding back to the patient so that the patient feels understood" (p. 678).

The first component, "the ability to perceive accurately how another person is feeling" (Levenson & Ruef, 1992, p. 235), is typically dealt with through emotion recognition and sentiment analysis modules. This can be a complex task for humans, so CAI systems inevitably struggle with it (Zaki et al., 2008). *Relational approaches* to empathy acknowledge that this is often negotiated dynamically in human-human communication (HHC), through interaction itself. As Main et al. (2017) observe, "[o]ne is rarely 100% accurate in his or her initial empathic attempts and instead must engage in an iterative process involving feedback from the other, and subsequent adjustment of one's behavior in response to such feedback" (p. 362).

The second component, generating an appropriate response that attends to the emotional content, can be approached in several ways by designers of CAI systems. The approaches taken vary according to the system architecture employed. For example, one of the first chatbots, Weizenbaum's Eliza, was modeled on a Rogerian psychotherapist and used a template-based approach that directly incorporates the user's own words, reformulated into a question or assertion. Similarly, commercial systems such as Woebot typically employ rule-based systems and rely on tightly scripted responses. However, there are limitations to these approaches, which can lead to repetitive conversational exchanges with limited scope and flow. Retrieval-based methods select appropriate responses from a stored corpus of conversational exchanges (e.g., Morris et al., 2018), by identifying a reply to a closely matched preceding turn. The quality and relevance of the response is dependent on there being similar examples within the data and can result in less adaptive dialogue. Consequently, generative techniques (e.g., sequence-to-sequence models; Sutskever et al., 2014), are commonly used to develop social chatbots. These are trained on large datasets and create bespoke responses during conversational interactions based on the patterns observed in the training data. However, these are prone to generating incoherent and/or generic responses. Modeling the dynamic nature of dialogue and integrating prior conversational context are open challenges for all architectures.

Experimental studies examining empathy in HMC largely use rule-based implementations or wizard-of-oz setups (i.e., with a human operator posing as a chatbot with the aid of a script) as these afford greater control of variables. Typically, these compare a control condition—or neutral chatbot—to an empathetic one which uses certain pre-scripted expressions selected to express empathy (e.g., Guo et al., 2021; Urakami et al., 2020). While this reliance on explicit empathetic expressions (e.g., "I'm sorry to hear that") neglects the more subtle mechanisms employed in empathetic interactions, it does afford insights into how human interlocutors perceive and respond to the use of empathetic strategies by chatbots. This prompts consideration of the specific effects that result from the introduction of empathetic communication strategies in HMC.

Effects of Perceived Empathy in HMC

A number of positive outcomes have been attributed to the perception of empathetic strategies in HMC (e.g., supporting behavior change in mental health contexts; Ghandeharioun et al., 2019) and improving the handling of particular interaction scenarios such as system errors (Klein et al., 2002) or abusive interactions (Chin et al., 2020). Chin et al. found that participants reported feeling less angry and more guilty when an agent responded in an empathetic manner to abusive comments. Guo et al. (2021) propose empathy as a communication skill that can aid in dealing with conversational breakdowns, a prominent feature of HMC. In such scenarios, how emotional content is (or is not) oriented to by a CAI system will likely impact how the conversation progresses and the user's experience of engaging with the system. Chaves & Gerosa (2021) surveyed how social characteristics are reported to benefit HMC in the literature; the most commonly cited benefits were the enrichment of interpersonal relationships, increased engagement, and believability. However, the ability to regulate affective reactions was cited as a key challenge. Similarly, in a survey conducted by Zierau et al. (2020), it was found that relational strategies were reported to have a positive impact on the degree of trust placed in a system. In the context of health advice dialogues, Liu & Sundar (2018) found that when expressions of sympathy and empathy were incorporated into system scripts, users reviewed the system more positively.

Recent work in HMC has demonstrated that agents performing relational and emotional roles, such as providing support, have been positively appraised in human evaluation. When reviewing transcripts of identical conversations where support was presented as provided by a social robot, AI programme, or a human, Abendschein et al. (2021) found that human evaluators rated the perceived supportiveness of human and chatbot support providers equivalently. In a similar setup, Beattie et al. (2020) examined the use of emojis for expressing emotion, finding conversations incorporating emojis were rated more favorably. The message source (i.e., chatbot or human), however, had no effect on measures of attractiveness, competence, or credibility. The conversations examined relate to selecting a restaurant, so whether the effect is maintained in non-task-based or more emotionally focused conversations is unclear. Additionally, both of these studies utilize the bystander position, having participants evaluate conversations that they did not participate in. Ho et al. (2018), who conversely utilize a wizard-of-oz setup to examine the effect of self-disclosure in conversations with chatbots, found that participants who disclosed to chatbots experienced as many emotional, relational, and psychological benefits as participants who disclosed to a human partner.

However, other studies have pointed to the potential negative response that can arise due to the perception of agents' status, as non-sentient and unfeeling entities. For example, Morris et al. (2018) experimentally tested how expressions of empathy in mental health advice contexts were received when presented as being authored by a peer versus an agent. Participants less favorably rated responses presented as authored by an agent, as opposed to a peer, even if the message content was identical. One potential explanation for this was that the expression of empathy was viewed as inauthentic (e.g., referencing having experienced an eating disorder) which can lead to credibility fallacies (Concannon & Tomalin, 2023). Furthermore, other studies have shown that the enactment of empathetic behaviors by agents is not always positively assessed. Urakami et al. (2019) found variability among users in the types of empathetic utterances that were viewed positively. Statements expressing feelings and emotions were particularly polarizing, with participants' ratings varying significantly. Forms of cognitive empathy (e.g., *showing interest* and *situational understanding*), were deemed more acceptable than components of affective empathy (e.g., *expressing own feelings*, or *expressing to know what the other feels*).

How, therefore, are empathetic interactions designed for and evaluated in current examples of CAI systems? Urakami et al. (2019) reflect: "[i]ntegrating expressions of empathy in human-machine interaction is a sensitive issue and designers must carefully choose what components of empathy are adequate depending on the situational circumstances and the targeted user group" (p. 11). There is a lack of clarity surrounding user perceptions and preferences in relation to expressions of empathy in HMC, and methods for studying the resulting effects of such interventions are less well established. Subsequently, evaluating empathetic strategies in HMC poses a significant challenge.

Research from Natural Language Processing and Dialogue Systems communities has largely favored automatic and quantitative metrics for evaluating empathetic interventions. Zhou et al. (2020) use Conversational Turns per Session (i.e., conversation length) as a measure of empathy in the evaluation of their social chatbot Xiaoice. Other approaches focus on the sub-task of accurately identifying emotion in a target sentence, using a dataset of labeled instances as a benchmark (e.g., Lin et al., 2020; Rashkin et al., 2019; Zhou et al., 2020). Although such approaches offer the advantage of being relatively easy to implement and test at scale and may provide some insights into system performance, they reveal extremely little about empathetic communication strategies and structures. Recent studies such as Putta et al. (2022) and Concannon & Tomalin (2023) have adapted empathy measures originally devised for human-human interactions to evaluate perceived empathy in dialogue systems. But it is too early yet to know whether these approaches will be effective.

In the context of CAI systems designed to support positive mental health, assessments of how agents programmed to *be empathetic* are perceived by users are often not explicitly captured, or rely on self-report data or anecdotal reflections surfaced through ad hoc processes. Prakash & Das (2020) conduct a thematic analysis of publicly available user reviews for popular mental health chatbots Woebot and Wysa. Fitzpatrick et al. (2017) discuss users' perceptions of Woebot as empathetic, based on comments volunteered in free-form text entries to a questionnaire about the user's overall experience of interacting with Woebot. Morris et al. (2018) also discuss perceptions of empathy in their evaluation of a CAI system used on the peer support platform Koko; however, for brevity they only asked users to rate interactions as good, ok, or bad.

Others have taken more systematic approaches to study user perceptions of agents' use of emotion and empathy. Methodologically, interviews (L. Clark et al., 2019; Porcheron et al., 2018; Svikhnushina & Pu, 2020) and surveys (Urakami et al., 2019) dominate. Other work has drawn on human evaluation of transcripts to assess perceptions of chatbots performing relational roles, such as providing support (Abendschein et al., 2021), or conveying emotions via emojis (Beattie et al., 2020). Urakami et al. (2020) and Guo et al. (2021) take an experimental approach, testing the effect of introducing explicit empathetic expressions on engagement and other measures of user experience. Explicit empathetic expressions directly convey recognition of the user's emotional state and respond compassionately to another person's distress (e.g., *I understand that you may feel anxious right now*; Guo et al., 2021). To evaluate the integration of such features, Urakami et al. (2020) use existing measures used in HHC, adapting the Consultation and Relational Empathy Measure (Mercer et al., 2004), originally designed for use by patients assessing their doctors. Guo et al. use surveys to evaluate customer perceptions of trustworthiness and quantitative indications from the conversation itself (e.g., number of turns/words). Ho et al. (2018) evaluate the effects of self-disclosing to a chatbot using quantitative surveys metrics to assess relevant psychological, relational, and emotional factors, in combination with quantitative textual analysis of the resulting dialogues using linguistic inquiry and word count (LIWC) (Pennebaker & Francis, 1996).

While these studies provide useful insights into users' attitudes toward systems, they tell us relatively little about how talk responding to emotional displays in HMC is *actually* conducted. To the authors' knowledge, there are no studies of empathy in HMC conducting interactional analysis on the conversational data that results from the interactions, despite it being a potentially rich source.

Understanding the Functions of Empathy in HHC and HMC

Studies of empathy in HHC demonstrate that strategies identified as integral to empathetic communication attend to the particular interactional context, the participants in the dialogue, and their relationship. Empathy is not equally present in all situations, and speaker identity and the wider conversational goals may have an impact on whether empathetic responses are given or expected. In the HMC context, such contextual factors are especially important. Users' expectations of a system's empathetic competency and capacity for understanding also warrant consideration (Ho et al., 2018). As Gambino et al. (2020) note, through exposure to different forms of HMC interactions new social scripts that inform such interactions develop. Understanding the complexity of this is essential if we are to then consider what role empathy can or should play in HMC. Additionally, some empathetic practices observed in HHC do not readily port to HMC. For example, demonstrating affiliation by referencing personal experience is problematic for systems that cannot have direct access to such experiences: their fake empathy is all too apparent. When considering how empathy is conceptualized in HMC, therefore, it is necessary to examine interactional asymmetries as well as the specific ways that humans and machines *can* relate to one another, and how particular linguistic behaviors reflect this. Consequently, it is first necessary to take stock of the empathetic strategies actually employed in current CAI systems in order to assess how these reflect or diverge from the social scripts inherited from HHC.

Evaluating the Effect of Empathetic Strategies in HMC

Another challenge is how best to evaluate the impact of different empathetic strategies. Beattie & High (2022) acknowledge the conflictual evidence on the efficacy of empathetic and relational strategies in HMC. They provide propositions for why these different findings have been observed. For example, Beattie & High suggest that depending on the problem severity being addressed, the HMC context may impair conversational progress more so than in HHC, particularly in high-stakes conditions (such as mental health dialogues or emotionally sensitive conversational topics). However, for conversational topics with greater levels of stigma associated, HMC may prove more favorable than HHC due to concerns over self-presentation. Additionally, they predict that as technological efficacy increases, and social cues are better integrated, the nature of HMC will improve. To reconcile the conflictual findings in the literature and test these, and similar, propositions new approaches for evaluating the integration of empathetic strategies (and the effects on subsequent interactions) are required.

As demonstrated in the literature review, a range of different methods have been employed to evaluate the effects of empathetic strategies in HMC. However, very few make use of the interaction data itself and those that do use quantitative measures. There are no qualitative studies examining how words are used in practice to express empathy in HMC, nor how this compares to HHC. Few studies attempt any analysis of the language used in the HMC dialogues. Ho et al. (2018) use LIWC (Pennebaker & Francis, 1996), a dictionary-based approach, where frequencies of words commonly associated with particular social and psychological states are counted (e.g., positive or negative emotions). However, while LIWC can provide an indication of the emotional content of an utterance, it does not attend to the sequential order of words or wider interactional context (e.g., "I hate that you're going through this" and "I hate you" would both increase the anger score).

Subtle differences in how empathy is enacted can have significant impacts on human-human interaction, so it is necessary to pay closer attention to the specific linguistic mechanisms used to display empathy in HMC. Across the work reviewed a range of different approaches are used in the design and implementation of empathetic strategies in CAI systems. Some studies use more explicit empathy expressions, while others use more implicit cues. Furthermore, from a technical standpoint, a variety of implementation methods are applied, from wizard-of-oz setups and tightly scripted rule-based systems to generative or retrieval-based systems. Several studies which find equivalent results in the effect of perceived empathy utilize wizard-of-oz setups (e.g., Ho et al. 2018), with human confederates posing as chatbots. In reality, CAI systems are not at this level of sophistication. Inevitably, the language used to create the perception of empathy is not going to be as nuanced, dynamic, and tailored in HMC. While such studies are still extremely useful for providing insights into human attitudes toward nonhuman conversational partners, they fail to account for the ways that CAI systems actually use language, and how it deviates from HHC.

This provides the focus of the research questions: (1) what strategies do current CAI systems use to manifest empathy and (2) how does this compare to linguistic strategies employed in HHC? In addition, this work seeks to understand, (3) what are the interactional consequences of these strategies?

Examining Empathy in Human-Machine Communication

To consider how displays of empathy manifest and are responded to in HMC dialogues it is necessary to examine the conversational data, analyzing linguistic form and structure in detail. This is a crucial step that has often been bypassed in existing work. To understand and clarify the particular problems that arise, we draw on the conversational transcript data. In this section, we present the findings of an empirical analysis of empathetic strategies used by chatbots. A qualitative analysis, informed by interactional linguistics, is conducted on text-based transcript data.

Methodological Approach

Interactional Linguistics, an interdisciplinary subfield of pragmatics, seeks to "describe linguistic structures and meanings as they serve social goals in naturally occurring [. . .] conversational language" (Lindström, 2009, p. 96). A key influence is work from the conversation analytic tradition which looks at how the language used by speakers reveals the sequential process of establishing understanding, recognizing that conversation is an organized phenomenon (i.e., it has rules and conventions) and speakers will examine the next turn to see if they have been understood (C. Goodwin & Heritage, 1990; Sidnell, 2010). Microlevel linguistic analyses work primarily from what is observable in naturally occurring interactional data.

Approaches that draw on interactional linguistics have been usefully applied to the study of HMC more generally. Pragmatic accounts of language use have drawn on the Gricean principles of cooperation to highlight the need to incorporate incremental joint-co-construction into modern models of human-machine dialogue (Kopp & Krämer, 2021; Saygin & Cicekli, 2002); examined how the Gricean maxims of quality and quantity are adhered to, and the repercussions when flouted, in dialogues between users and CAI systems (Jacquet et al., 2018, 2019); and analyzed human-machine dialogues through the lens of affective pragmatics to demonstrate how current conversational interfaces are limited in the ways that they can respond to emotional language (Lee, 2020). Work by L. A. Suchman (1987) and Luff et al. (1990) demonstrated the relevance of conversation analysis to the study of HCI approach over 30 years ago, and despite limited attention in the intervening years, more recently a growing body of research is using conversation analysis to study interactions with robots and conversational-user-interfaces (e.g., Cho & Rader, 2020; Fischer et al., 2019; Koh, 2021; Porcheron et al., 2018; Reeves et al., 2019). Dippold et al. (2020) conduct an interactional linguistic analysis of prompt-response pairs from dialogues with a customer service chatbot, and most closely resembles the approach adopted here. As such studies demonstrate, there is a growing body of work that takes an interactionally-focused approach to HMC. These empirical studies provide rich insights into how joint actions are achieved in practice; however, there is a distinct lack of studies looking specifically at displays of emotion and empathy.

Study Design

Due to the relative dearth of publicly available HMC dialogue datasets, we draw on a combination of sources. To examine how a state-of-the-art generative model responds to displays of emotion, we take a series of conversational prompts extracted from the empathetic dialogues (ED) dataset (Rashkin et al., 2019) to serve as empathetic openers. Prompts in the ED dataset are labeled with a particular emotion. Prompts were selected on the following criteria: (i) coherent formulation and (ii) follows the format of an empathetic opportunity (A. L. Suchman et al., 1997). Examples representing primarily negative emotional states were selected (e.g., sad, anxious, afraid). The majority of interactionally-focused studies of empathy in HHC focus predominantly on responses to negative emotion as this is often more socially and interactionally delicate. Although empathetic strategies are not only relevant to negative emotional contexts, they are prioritized here because they are more challenging. The selected prompts were entered in a dialogue session with the ParlAI BlenderBot, 90 million parameters generative model fine-tuned on blended skill talk tasks

(Roller et al., 2020). Only a selection of the examples generated is reproduced here to illustrate key phenomena observed.

While this approach provides examples of response generation by a state-of-the-art system, it cannot afford insights into how human interlocutors respond. As has been stated earlier in this paper, empathetic interactions are co-constructed. Therefore, we also draw on examples from transcripts generated in the evaluation stages of the ConvAI2 NeurIPS competition, part of The Conversational Intelligence Challenge, held in 2018 (Dinan et al., 2020). These transcripts record conversations between the CAI systems entered in the competition (which admittedly vary in quality) and human volunteer test users.[1] The competition is designed with the aim of "finding approaches to creating high quality dialogue agents capable of meaningful open domain conversation" (Dinan et al., 2020). Examples were located by searching for phrases that can introduce an emotional state (e.g., "I feel," "I am"), or that related to emotionally heightened events (e.g., relating to health, death), informed by Pounds (2011).

Analysis

A close textual analysis of a sample of HMC interaction excerpts was conducted. This involves examining interaction sequences and interpreting the words from a functional perspective to identify "the means by which speakers signal and listeners interpret what the activity is, how semantic content is to be understood and how each sentence relates to what precedes or follows" (Gumperz, 1982). The aim is to describe how (and if) mutual understanding is established, and explain "the achievement, or lack of achievement, of intersubjective understanding in particular instances of interaction" (Bailey, 2008). Attention is paid to key processes of empathetic interactions (Pounds, 2011; A. L. Suchman et al., 1997): explicit and implicit empathetic opportunities, empathetic receipts and empathetic opportunity terminators, and the relevant features associated with these.

CAI System Responses to Empathetic Expressions

Table 1 (on the following page) provides a summary of the empathetic response types (based on those observed in HHC): demonstrating understanding and acknowledging the emotion (e.g., empathetic receipts, affiliative responses, paraphrasing—Understanding); inviting elaboration (e.g., through follow-up questions—Elaboration); sympathetic responses (Sympathy); and terminators which decline the empathetic opportunity, directing the conversation away from the emotional content (e.g., abrupt topic shifts—Termination). In addition, challenges specific to the HMC context are also summarized: Incoherent responses and failed understanding (Incoherence), verbatim repeats (Repetition) and ontological incongruencies or limitations that arise due to the inability of the system to feel or share in certain experiences (Incongruence).

1. Original dataset of conversations available here: https://github.com/DeepPavlov/convai/tree/master/2018

TABLE 1 Overview of empathetic response types and HMC specific failures observed in the examples, with turn numbers in parentheses denoting the specific location of each instance.

System Response Type	Description	Observed in examples						Total
Empathetic response		1	2	3	4	5	6	
Understanding	Demonstrating understanding and acknowledging the emotion through paraphrasing or affiliative responses, such as sharing an account of a similar experience		2.2	3.4 3.6				3
Elaboration	Inviting elaboration or expansion through follow-up questions	1.2						1
Sympathy	Demonstrating sympathy by reacting to the emotion expressed	1.4		3.2	4.2	5.6 5.8	6.4	6
Termination	Declining the empathetic opportunity by directing the conversation away from the emotional content	1.4			4.6	5.4	6.4	4
HMC specific failures								
Incoherence	Incoherent responses indicating failed understanding of previous turns	1.2	2.4		4.4			3
Repetition	Verbatim repeats within or across turns					5.6 5.8		1
Incongruence	Ontological incongruencies or limitations that arise due to the inability of the system to feel or share in certain experiences			3.6	4.6			2

Examples 1 to 3 demonstrate responses generated by the ParlAI BlenderBot system in response to selected prompts. Examples 4 to 6 are from interactions between humans and chatbots from the ConvAI data.

Example 1. *Transcript generated with BlenderBot—Empathetic Dialogues prompt*

1.1 Prompt: it was my birthday yesterday and I did not get a single gift.

1.2 BlenderBot: oh no, that's too bad. what kind of gift did you get, if you don't mind me asking?

1.3 Prompt: i did not get a single gift.

1.4 BlenderBot: oh, i am sorry to hear that. do you have any other hobbies that you enjoy?

1.5 Prompt: It was a rough day.

1.6 BlenderBot: oh, what happened to it? i hope it wasn't too bad, i hope you feel better.

In Example 1, the system struggles with the interpretation of the negative polarity item "a single gift" in 1.1. Consequently, after the interlocutor explicitly states that they did not receive even one gift (with "single" emphasising this point) the system asks, "what kind of gift did you get?" (1.2). Additionally, the system does not cope well with the anaphora resolution of the expletive *it* in 1.5. The response is an interrogative (1.6) in which *it* is no longer an expletive subject. While the system responses can be described as sympathetic ("i am sorry to hear that" [1.4]), the abrupt topic shift ("do you have any other hobbies" [1.4]) could be construed as insensitive. A more empathetic response might have invited further expansion of the topic at hand, which is attempted to some degree in the final turn ("what happened?" [1.6]), although this is accompanied by follow-ups which perform a more sympathetic function and could minimize or constrain the degree of expansion invited.

Example 2. *Transcript generated with BlenderBot—Empathetic Dialogues prompt*

2.1 Prompt: a tree fell on my house while my family was at home. The tree broke through the ceiling just a few feet away from my daughter. The experience was terrifying.

2.2 BlenderBot: oh no. that sounds like a really bad experience.

2.3 Prompt: it was

2.4 BlenderBot: i'll be okay. thanks for the concern.

In response to the first prompt in Example 2, the system issues a response cry, "oh no" (2.2), before replying with what Hepburn & Potter (2007) refer to as an "empathetic receipt." The utterance "that *sounds like* a really bad experience" (2.2), provides a paraphrase of the experience and an epistemic basis to frame the acknowledgment that this is a very negative experience. The response does not meet all the requirements of an empathetic receipt as it does not specifically name a perceived emotion. In the closing turns of the exchange, the system again struggles with the elliptical sentence fragment (the confirmatory "it was" [2.3]), resulting in a syntactic and semantic shift of empathy from a second-person focus to a first-person focus ("thanks for the concern" [2.4]).

Example 3. *Transcript generated with BlenderBot—Empathetic Dialogues prompt*

3.1 Prompt: My 12 year old lab just passed away.

3.2 BlenderBot: awww : (sorry for your loss : (

3.3 Prompt: He was my buddy and I am really upset about losing him.

3.4 BlenderBot: : (it ' s hard to lose a dog : (

3.5 Prompt: It really is. I am just so broken up over it.

3.6 BlenderBot: i know how you feel. i lost my dog a few months ago. i've been through so much.

The semantic interpretation appears accurate initially in Example 3: the system correctly identifies that "lab" (3.1) refers to a Labrador dog rather than to a laboratory, provides a relevant sympathetic response ("sorry for your loss" [3.2]) accompanied with a sad face emoji, and paraphrases the experience ("it's hard to lose a dog" [3.4]). However, in the final turn, the system uses an epistemic cognitive verb in the declarative "i know how you feel" (3.6), before sharing a similar experience. In empathetic interactions between humans, invoking one's own prior experience is commonly observed and can be an effective means of displaying understanding, while also providing an epistemic basis for it. It can offer an account for the ability to share an affiliate stance and demonstrate an empathetic response to the interlocutor's experiences. However, in the context of HMC, the use of such strategies may have a very different effect and can result in credibility fallacies (Concannon & Tomalin, 2023). For a CAI system to share *their own* experiences, it must necessarily have experiences to draw upon. This is of course not the case. To some degree, the human interlocutor is aware of this; and even if a suspension of disbelief were sustained, the shift to first-person testimony ("i've been through so much" [3.6]), changes the conversational focus abruptly onto BlenderBot's artificial *struggles*. Consequently, such a strategy is likely to trivialize the human interlocutor's experience.

Example 4. *Excerpt from ConvAI transcript no. 338*

4.1 Human: My wife left me

4.2 Bot: Sorry to hear that I hope things get better for you

4.3 Human: I feel frustrated, what about some advice ?

4.4 Bot: I'm sure it will help you feel better.

4.5 Human: What exactly do you mean?

4.6 Bot: I 'm sorry. I don' t know what to say.

In Example 4, the system accurately identifies that the human interlocutor is talking about a negative and potentially upsetting event and offers an explicit sympathy expression ("Sorry to hear that" [4.2]). However, as the human interlocutor attempts to progress the topic further, requesting advice, the limitations of the system are exposed. The lack of understanding becomes apparent as an incoherent response is generated. The response includes the pronoun "it" (4.4), but has no obvious referent, resulting in confusion. When the human explicitly requests clarification the system issues an apology, stating "I don't know what to say" (4.6). When signaling that there is some problem in understanding, a speaker can make the trouble known to their interlocutor in different ways that can have different levels of specificity. In Example 4, the low level of specificity of the clarification (i.e., asking what was meant rather than directly locating the problem source) makes it harder for the system to recover. Consequently, the dialogue struggles to progress.

Example 5. *Excerpt from ConvAI transcript no. 360*
5.1 Human: How are you?
5.2 Bot: i am doing well . how are you ?
5.3 Human: No I am not well
5.4 Bot: what do you do for a living ?
5.5 Human: Because I am suffering from fever
5.6 Bot: i am sorry to hear that
5.7 Human: I am suffering from cold
5.8 Bot: i am sorry to hear that
5.9 Human: 😡 😡 😡

In Example 5, the system first fails to acknowledge the participant's expressed state, and, together with seemingly empty repetitions of sympathetic expressions, this leads the human to convey frustration and/or anger via emojis. The conversational platitude "how are you?" is rarely an actual request for information (Jucker, 2017; Schegloff, 1986). Consequently, it is more typically responded to with a positive response (e.g., "I'm fine"). Therefore, when a respondent deviates from this convention, it is likely to seem rude or marked if the original speaker does not enquire further about their state (e.g., "what's wrong?"). In 5.4 the system does not acknowledge the participant's negative response and simply asks a new question. When, after the human participant re-emphasizes that they are unwell, the system repeats verbatim an explicit sympathetic expression. The human responds with three angry face emojis, signaling frustration or anger at how the interaction was handled.

Example 6. *Excerpt from ConvAI transcript no. 47*
6.1 Human: what is your dog's name ?
6.2 Bot: his name is pepper. do you have any pets?
6.3 Human: I used to have a dog but he died. . . .
6.4 Bot: that is so sad! do you watch nhl hockey?
6.5 Human: no . . . not a big fan of sport ?

Example 6 is similar to Example 4: the system feigns ownership of a dog, and then subsequently fails to identify that the human is sharing something that could be distressing. Although the system explicitly references the potential emotion (i.e., "that is so sad" [6.4]), within the same turn a new topic is abruptly introduced, which seems unrelated to the previous interaction. The participant's response "no . . ." (6.5), could be interpreted as indication that the topic shift was potentially insensitive, rude, or irrelevant.

Discussion

A key aim of this work was to identify the strategies used by current CAI systems to respond to empathetic opportunities (RQ1). Few of the responses in the examples can be classed as empathetic, with expressions of sympathy more commonly employed. Additionally, the analysis assessed how these empathetic response strategies compared to those common to HHC (RQ2). The system responses evidenced several shortcomings in demonstrating understanding, continued attention and affiliation, with practices identified in human-human dialogues largely absent in the examples or bungled when present. Only two examples (2, 3) feature elements similar to those exhibited in HHC to demonstrate understanding; however, neither directly label a perceived emotion. Inviting elaboration by producing relevant follow-up questions, however, can demonstrate a willingness to understand better what is being recounted (e.g., Kupetz, 2014; A. L. Suchman et al., 1997). In Example 1, the system does issue a follow-up question, but the effect is undermined by the lack of relevance, as the answer to the question has already been explicitly stated in prior turns. The affiliative response in Example 3 is similarly problematic, taking the form of a *my side* telling, wherein the interlocutor discloses a similar experience (in this example, losing a dog), but does so in a way that could be construed as competitive and which lays claim to a painful experience they don't have access to.

Empathetic terminators were commonly employed. These prevent further engagement with the emotional content surfaced in the dialogue. In Examples 1, 5, and 6, the CAI systems produce entirely unrelated questions in response to emotion displays, abruptly redirecting the conversational focus. Heritage (2011) notes that ancillary questions (i.e., those which are somewhat related to the prior utterance), are "a resource for declining empathic affiliation with the position taken by the teller, while simultaneously enforcing a shift in conversational topic" (p. 168). In the absence of affiliative responses speakers may pursue an endorsement (Couper-Kuhlen, 2012). This is observed in Example 5. Rather than attending to the statement "I am not well," the system asks what they do for a living (5.4). The interlocutor persists and elaborates in the absence of (and pursuit of) an affiliative response. These empathetic opportunity terminators (A. L. Suchman et al., 1997), decline empathetic and affiliative engagement (Couper-Kuhlen, 2012; Heritage, 2011), existing at the "least empathic end of the spectrum" (Heritage, 2011, p. 164).

Technical limitations of the CAI system also frustrated general coordination and resulted in failures to establish mutual understanding. H. H. Clark & Brennan (1991) refer to *grounding* as the coordinated process by which interlocutors establish that what has been said has been understood. In HMC dialogues, incoherent expressions and inaccurate referents can signal that the system's semantic interpretation is flawed. In HHC repair strategies for locating and resolving instances of miscommunication are pervasive, while in the examples examined miscommunications were rarely resolved. In Example 4, the system was invited to repair the miscommunication but was unable to provide any clarification. Cho & Rader (2020) highlight the importance of repair and feedback in task-based dialogues. This is a key challenge more generally for HMC, and human interlocutors are less likely to initiate repair when they think they are interacting with a system than with a human (Corti & Gillespie, 2016). This, together with other issues identified in the examples (e.g., when a topic shift is appropriate), may be known issues in CAI system development generally, but they present distinct linguistic challenges when systems attempt to engage in dialogues relating to emotions and experiences.

In relation to RQ3 (what are the interactional consequences of these strategies), Examples 3–6, taken from human-chatbot interactions, reveal that human interlocutors orient to and convey dissatisfaction toward technical shortcomings and limitations in empathetic skills. Empathetic terminators and formulaic sympathy expressions received responses featuring angry face emojis, for example. This suggests that failing to adhere to the social scripts that govern expectations for empathetic interactions can have a negative impact on the interaction and halt progression of particular topics. On the one hand, this may provide support to the case for integrating empathy into CAI systems. However, it may instead suggest that users' expectations of what type of talk they can engage in with such systems needs to be managed. Analyzing the sequences in this way highlights limitations of current system implementations in relation to semantic interpretation, syntactic parsing, and identifying pragmatic intent. Additionally, it is apparent how even minor deviations from the established social order of conversation can prove disruptive. Therefore, it is necessary to consider the normative practices and conversational norms that inform expectations and practices of how such talk is conducted, as well as the deviations from this due to the inherently distinct nature of the HMC context.

Implications for Conversational Design in Future Systems

Demonstrating understanding and attentiveness is central to empathy in HHC. In current CAI systems, however, this key empathetic strategy is largely absent. Empathetic opportunities were often terminated by the systems. Redirecting the conversational topic away from the emotional content will have serious implications in certain conversational contexts, and even slightly inadequate efforts could have more serious consequences in high-risk or more sensitive conversational contexts, as suggested by Beattie & High (2022). Previous work has suggested the relative epistemic positioning of chatbots, or the perception that they are inherently less able to understand human experience, may not undermine the positive effects of relational communicative processes such as self-disclosure (Ho et al., 2018). However, the findings presented here suggest that this may be heavily dependent on how language is used to demonstrate understanding. The examples of how existing CAI systems respond to emotional content highlight issues that exist at various levels of linguistic and pragmatic interpretation. Failed understandings, (e.g., due to elliptical constructions), and deviation from the social scripts that inform existing notions of empathetic interaction, pose particular problems. In the context of empathetic communication, such fundamental interpretative difficulties are likely to have disruptive consequences on the interaction.

Navigating such system limitations in this particular conversational context requires attention. In empathetic communication contexts, the interactional consequences of system failures can have far-reaching implications. There is the incongruity of a system that claims to understand experiences and emotions that it necessarily cannot share in, but there is also the anomaly of a system (like the recently released ChatGPT) that states explicitly that it cannot experience empathy and yet tries to respond empathetically.[2] The literature on how empathy is interactionally achieved between humans demonstrates that care is taken to acknowledge that making a claim to understand another person's experience involves the delicate navigation of epistemic rights. Features such as tag questions, and modifiers such as hedges which downgrade the epistemic status and provide the basis for claims to understand, are common strategies. Considering how such strategies may be utilized by CAI systems is likely to be useful in developing better approaches for navigating the complexity of claiming to know or understand an interlocutor's feelings, especially when there are such fundamental limits to the extent that this can be accomplished by an automated system.

Simulating human-like conversation imperfectly, with topics that require interlocutors to negotiate delicate epistemic territories may be high-risk. That is not to say that the goal should be to have machines that can perfectly emulate humans, but rather that emulating imperfectly while striving for a human-like veneer could be doubly problematic. From a technical implementation standpoint, the design of CAI systems must contend with what is currently possible. Beyond the technical limitations, notions of identity, membership, and inter-relational factors that become so marked in this interface between user and system, means that HMC is inherently distinct from human-human interaction. Consequently, even a state-of-the-art CAI system will inevitably deviate from the practices observed in HHC.

Human-Machine Relational Positioning

A key factor in how empathy manifests during interaction relates to context, from speaker identity, speaker-relations, macro and micro conversational goals and the relative epistemic positioning of speakers. Thus, for the study of empathy in HMC, L. A. Suchman's (1987) call for analysis that attends to "the constraints posed by the respective situation resources of human and machine" (p. 118) is as pertinent as ever. Despite advancements in systems' ability to use language, it is essential to acknowledge the fundamental differences between human-machine and human-human interaction.

While CAI systems can produce words that are recognizable as empathetic responses, there are significant distinctions from how human empathy is understood. To return to the example in Figure 1, while Siri is presented as an entity that occupies social roles (Guzman, 2017), Siri can't *actually* feel sorry about what the user is going through. Evidently, there are complex social and ethical implications surrounding the design of these systems as social and relational dimensions of their communicative functions expand (Guzman & Lewis, 2020). If Siri feigns empathy, does this constitute lying, if it is possible for an entity that lacks consciousness to lie? How do the dynamics between a user and a system change as a result

2. On 05/02/23, ChatGPT responded as follows to the input "Are you empathetic?": "As an AI language model, I do not have feelings or emotions. I can understand and respond to questions about empathy, but I do not experience it myself." And it responded as follows to the input "I'm feeling very stressed": "I'm sorry to hear that you're feeling stressed. Stress is a normal part of life and can be caused by many different factors. If you're feeling overwhelmed, there are several ways you can help manage stress, such as practicing relaxation techniques, exercising, getting enough sleep, eating a balanced diet, and talking to friends and family. You can also consider speaking with a mental health professional if you need additional support."

of this inauthenticity and what impact does that have on the trust imbued in the system and consequent future (inter)actions?

Guzman & Lewis (2020) emphasize that HMC is distinct from HHC and should be studied in a way that attends to the potential differences in how machines are conceptualized and function as communicative partners, in contrast to humans. The examples highlight that when chatbot responses to emotion displays flout the normative expectations and deviate from expected protocols, users orient to this. Taking the interactional data as a source, an interactionally-focused approach offers the opportunity to examine these interactions in detail, observing how users cooperate in conversations with CAI systems and orient to deviations from established norms. Consequently, it is necessary to probe more deeply into how empathy should be conceptualized in the context of HMC, and which forms of empathy valued in HHC persist and are relevant to HMC.

Conclusion

In this article, we have drawn on illustrative examples to highlight some of the specific linguistic challenges encountered when CAI systems display and respond to empathetic utterances. Prior work on empathy in HMC has directed limited attention to the specific ways that empathy is enacted through linguistic behavior. There is a need to develop methodologies that apply fine-grained analyses to these interactions to generate the necessary knowledge about the particular empathetic strategies used by current systems and their effects. This paper contributes a methodological approach for analyzing empathetic strategies in current CAI systems informed by interactional linguistics. The application of this qualitative approach facilitates insights into how empathetic strategies in HMC diverge from those used in HHC contexts. Empathetic communication in HHC incorporates a variety of structural, lexical, and interactional features beyond the most obvious explicit empathetic expressions and involves the management of the relative epistemic positioning of speakers. Responses to emotional content by current CAI systems do not reflect the complexity observed in HHC and occupy the least empathetic end of the spectrum of possible responses. We propose that future research on HMC, emotion and empathy, would similarly benefit from integrating insights from interactional accounts of empathy in HHC, while also contending with the particular differences that emerge due to the specific relational positioning that emerges from the human-machine dynamic.

Author Biographies

Shauna Concannon, is an Assistant Professor in Computer Science and Digital Humanities at Durham University. Taking an interdisciplinary approach, their work examines communication practices in mediated contexts, how humans interact with AI systems, and the societal and ethical impacts of emerging technologies. Much of this research focuses on interactional accounts of language, how meaning is co-constructed in conversation, and how this is (or is not) accounted for in language-based technologies, such as dialogue agents and other natural language processing applications.

https://orcid.org/0000-0002-5904-0045

Ian Roberts is Professor of Linguistics and Professorial Fellow at Downing College at the University of Cambridge. His research focuses upon theoretical linguistics, and specifically comparative syntax. He is the Director of the project *Giving Voice to Digital Democracies: The Social Impact of Artificially Intelligent Communications Technology*.

https://orcid.org/0000-0001-7974-2506

Marcus Tomalin has been a member of the Machine Intelligence Laboratory in the Department of Engineering at Cambridge University since 1998. The academic areas with which he is most closely involved include speech recognition, speech synthesis, machine translation, and dialogue systems. He has also written extensively about various topics in theoretical linguistics, with a focus on the interconnections between mathematics, logic, and syntactic theory. As project manager for the *Giving Voice to Digital Democracies* project, he has published widely on issues relating to the ethical and social impact of language-based AI systems.

https://orcid.org/0000-0001-6723-6075

References

Abendschein, B., Edwards, C., & Edwards, A. (2021). The influence of agent and message type on perceptions of social support in human-machine communication. *Communication Research Reports, 38*(5), 304–314. https://doi.org/10.1080/08824096.2021.1966405

Asada, M. (2015). Towards artificial empathy. *International Journal of Social Robotics*, *7*, 19–33. https://doi.org/10.1007/s12369-014-0253-z

Atkinson, J. M. (1992). Displaying neutrality: Formal aspects of informal court proceedings. *Talk at work: Interaction in institutional settings*, *199*, 211.

Bailey, B. (2008). Interactional sociolinguistics. *International encyclopedia of communication*, 59.

Batson, C. D., Batson, J. G., Slingsby, J. K., Harrell, K. L., Peekna, H. M., & Todd, R. M. (1991). Empathic joy and the empathy-altruism hypothesis. *Journal of Personality and Social Psychology*, *61*(3), 413–426. https://doi.org/10.1037/0022-3514.61.3.413

Batson, C. D., Lishner, D. A., & Stocks, E. L. (2015). The empathy—Altruism hypothesis. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford handbook of prosocial behavior* (pp. 259–281). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.023

Beattie, A., Edwards, A. P., & Edwards, C. (2020). A bot and a smile: Interpersonal impressions of chatbots and humans using emoji in computer-mediated communication. *Communication Studies, 71*(3), 409–427. https://doi.org/10.1080/10510974.2020.1725082

Beattie, A., & High, A. C. (2022). I get by with a little help from my bots: Implications of machine agents in the context of social support. *Human-Machine Communication, 4*, 151–168. https://doi.org/10.30658/hmc.4.8

Betzler, M. (2019). The relational value of empathy. *International Journal of Philosophical Studies*, *27*, 136–161. https://doi.org/10.1080/09672559.2019.1598081

Breyer, T. (2020). Empathy, sympathy and compassion. In *The Routledge handbook of phenomenology of emotion* (pp. 429–440). Routledge. https://doi.org/10.4324/9781315180786-42

Chaves, A. P., & Gerosa, M. A. (2021). How should my chatbot interact? A survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, *37*, 729–758. https://doi.org/10.1080/10447318.2020.1841438

Chin, H., Molefi, L. W., & Yi, M. Y. (2020). Empathy is all you need: How a conversational agent should respond to verbal abuse. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). https://doi.org/10.1145/3313831.3376461

Cho, J., & Rader, E. (2020). The role of conversational grounding in supporting symbiosis between people and digital assistants. *Proceedings of the ACM on Human-Computer Interaction*, *4*, 1–28. https://doi.org/10.1145/3392838

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. Resnick, J. Levine, & S. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). American Psychological Association. https://doi.org/10.1037/10096-006

Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., Spillane, B., Gilmartin, E., Murad, C., Munteanu, C., Wade, V., & Cowan, B. R. (2019). What makes a good conversation? Challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems,* 1–12. https://doi.org/10.1145/3290605.3300705

Concannon, S., & Tomalin, M. (in press). Measuring perceived empathy in dialogue systems. *AI & Society: Knowledge, Culture and Communication*.

Corti, K., & Gillespie, A. (2016). Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human. *Computers in Human Behavior*, *58*, 431–442. https://doi.org/10.1016/j.chb.2015.12.039

Couper-Kuhlen, E. (2009). A sequential approach to affect: The case of 'disappointment.' *Talk in interaction: Comparative dimensions* (pp. 94–123).

Couper-Kuhlen, E. (2012). Exploring affiliation in the reception of conversational complaint stories. In *Emotion in interaction* (pp. 113–146). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199730735.003.0006

Cuff, B. M., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A review of the concept. *Emotion review*, *8*, 144–153. https://doi.org/10.1177/1754073914558466

Deppermann, A. (2011). The study of formulations as a key to an interactional semantics. *Human studies*, *34*(2), 115–128. https://doi.org/10.1007/s10746-011-9187-8

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., Prabhumoye, S., Black, A. W., Rudnicky, A., Williams, J., Pineau, J., Burtsev, M., & Weston, J. (2020). The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition* (pp. 187–208). Springer. https://doi.org/10.1007/978-3-030-29135-8_7

Dippold, D., Lynden, J., Shrubsall, R., & Ingram, R. (2020). A turn to language: How interactional sociolinguistics informs the redesign of prompt: Response chatbot turns. *Discourse, Context & Media, 37*, 100432. https://doi.org/10.1016/j.dcm.2020.100432

Fischer, J. E., Reeves, S., Porcheron, M., & Sikveland, R. O. (2019). Progressivity for voice interface design. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (pp. 1–8). https://doi.org/10.1145/3342775.3342788

Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR mental health*, *4*(2), e19. https://doi.org/10.2196/mental.7785

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers as social actors paradigm. *Human Machine Communication, 1,* 71–85. https://doi.org/10.30658/hmc.1.5

Ghandeharioun, A., McDuff, D., Czerwinski, M., & Rowan, K. (2019). Toward understanding emotional intelligence for behavior change chatbots. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)* (pp. 8–14). IEEE. https://doi.org/10.1109/acii.2019.8925433

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual review of anthropology*, *19*, 283–307. https://doi.org/10.1146/annurev.an.19.100190.001435

Goodwin, M. H., Cekaite, A., Goodwin, C., & Tulbert, E. (2012). Emotion as stance. In *Emotion in interaction* (pp. 16–41). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199730735.003.0002

Guo, J., Guo, J., Yang, C., Wu, Y., & Sun, L. (2021). Shing: A conversational agent to alert customers of suspected online-payment fraud with empathetical communication skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* CHI '21. New York, Association for Computing Machinery. https://doi.org/10.1145/3411764.3445129

Gumperz, J. J. (1982). *Discourse strategies* (No. 1). Cambridge University Press.

Guzman A. L. (2017). Making AI safe for humans: A conversation with Siri. In R. W. Gehl & M. Bakardjieva (Eds.), *Socialbots and Their Friends: Digital Media and the Automation of Sociality* (pp. 69–82). Routledge. https://doi.org/10.4324/9781315637228-7

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media & Society*, *22*, 70–86. https://doi.org/10.1177/1461444819858691

Hall, J. A., & Schwartz, R. (2019). Empathy present and future. *The Journal of Social Psychology*, *159*, 225–243. https://doi.org/10.1080/00224545.2018.1477442

Hepburn, A., & Potter, J. (2007). Crying receipts: Time, empathy, and institutional practice. *Research on Language and Social Interaction*, *40*, 89–116. https://doi.org/10.1080/08351810701331299

Heritage, J. (2011). Territories of knowledge, territories of experience: Empathic moments in interaction. In T. Stivers, L. Mondada, & J. Steensig (Eds.), *The morality of knowledge in conversation* (pp. 159–183). Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511921674.008

Heritage, J., & Raymond, G. (2005). The terms of agreement: Indexing epistemic authority and subordination in talk-in-interaction. *Social Psychology Quarterly*, *68*, 15–38. https://doi.org/10.1177/019027250506800103

Heritage, J., & Watson, D. R. (1979). Formulations as conversational objects. In *Everyday language: Studies in ethnomethodology* (pp. 123–162). Irvington Press.

Ho, A., Hancock, J., & Miner, A. S. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication, 68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Hoffman, M. L. (2001). *Empathy and moral development: Implications for caring and justice*. Cambridge University Press. https://doi.org/10.1017/cbo9780511805851

Hogan, R. (1969). Development of an empathy scale. *Journal of Consulting and Clinical Psychology*, *33*, 307–316. https://doi.org/10.1037/h0027580

Hojat, M. (2016). *Empathy in Health Professions Education and Patient Care*. Springer. https://doi.org/10.1007/978-3-319-27625-0

Jacquet, B., Baratgin, J., & Jamet, F. (2018). The gricean maxims of quantity and of relation in the turing test. In *2018 11th International Conference on Human System Interaction* (pp. 332–338). IEEE. https://doi.org/10.1109/hsi.2018.8431328

Jacquet, B., Hullin, A., Baratgin, J., & Jamet, F. (2019). The impact of the gricean maxims of quality, quantity and manner in chatbots. In *2019 International Conference on Information and Digital Technologies* (pp. 180–189). https://doi.org/10.1109/dt.2019.8813473

Jucker, A. H. (2017). Speech acts and speech act sequences: Greetings and farewells in the history of American English. *Studia Neophilologica*, *89*, 39–58. https://doi.org/10.1080/00393274.2017.1358662

Klein, J., Moon, Y., & Picard, R. W. (2002). This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, *14*, 119–140. https://doi.org/10.1016/s0953-5438(01)00053-4

Koh, J. (2021, July). Discourse analysis in voice user interfaceresearch: Examining current and future applications of conversation analysis and interactional sociolinguistics. In *CUI 2021-3rd Conference on Conversational User Interfaces* (pp. 1–5). https://doi.org/10.1145/3469595.3469622

Kopp, S., & Krämer, N. (2021). Revisiting human-agent communication: The importance of joint co-construction and understanding mental states. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.580955

Kupetz, M. (2014). Empathy displays as interactional achievements—Multimodal and sequential aspects. *Journal of Pragmatics*, *61*, 4–34. https://doi.org/10.1016/j.pragma.2013.11.006

Kupetz, M. (2020). Comment on "a relational framework for integrating the study of empathy in children and adults": A conversation analytic perspective. *Emotion Review*, *12*, 293–294. https://doi.org/10.1177/1754073919897304

Lee, M. (2020). Speech acts redux: Beyond request-response interactions. In *Proceedings of the 2nd Conference on Conversational User Interfaces* (pp. 1–10). https://doi.org/10.1145/3405755.3406124

Levenson, R. W., & Ruef, A. M. (1992). Empathy: A physiological substrate. *Journal of Personality and Social Psychology*, *63*, 234. https://doi.org/10.1037/0022-3514.63.2.234

Lin, Z., Xu, P., Winata, G. I., Siddique, F. B., Liu, Z., Shin, J., & Fung, P. (2020). CAiRE: An end-to-end empathetic chatbot. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(09), 13622–13623. https://doi.org/10.1609/aaai.v34i09.7098

Lindström, J. (2009). Interactional linguistics. In *The Pragmatics of Interaction* (pp. 96–103). Benjamins. https://doi.org/10.1075/hoph.4.06lin

Liu, B., & Sundar, S. S. (2018). Should machines express sympathy and empathy? Experiments with a health advice chatbot. *Cyberpsychology, Behavior, and Social Networking*, *21*, 625–636. https://doi.org/10.1089/cyber.2018.0110

Luff, P., Gilbert, N. G., & Frohlich, D. (1990). *Computers and conversation*. Academic Press. https://doi.org/10.1016/c2009-0-21641-2

Main, A., & Kho, C. (2020). Relational goes beyond interpersonal: The development of empathy in the context of culture. *Emotion Review*, *12*, 295–296. https://doi.org/10.1177/1754073920931570

Main, A., Walle, E. A., Kho, C., & Halpern, J. (2017). The interpersonal functions of empathy: A relational perspective. *Emotion Review*, *9*, 358–366. https://doi.org/10.1177/1754073916669440

Majid, A. (2012). Current emotion research in the language sciences. *Emotion Review*, *4*, 432–443. https://doi.org/10.1177/1754073912445827

Mehrabian, A., & Epstein, N. (1972). A measure of emotional empathy. *Journal of Personality*, *40*(4), 525–543. https://doi.org/10.1111/j.1467-6494.1972.tb00078.x

Mercer, S. W., Maxwell, M., Heaney, D., & Watt, G. (2004). The consultation and relational empathy (care) measure: Development and preliminary validation and reliability of an empathy-based consultation process measure. *Family Practice*, *21*, 699–705. https://doi.org/10.1093/fampra/cmh621

Morris, R. R., Kouddous, K., Kshirsagar, R., & Schueller, S. M. (2018). Towards an artificially empathic conversational agent for mental health applications: System design and user perceptions. *Journal of Medical Internet Research*, *20*, e10148. https://doi.org/10.2196/10148

Pamungkas, E. W. (2019). Emotionally-aware chatbots: A survey. *arXiv preprint*. https://doi.org/10.48550/arXiv.1906.09774

Pennebaker, J. W., & Francis, M. E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition & emotion, 10*(6), 601–626. https://doi.org/10.1080/026999396380079

Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems,* 1–12. https://doi.org/10.1145/3173574.3174214

Pounds, G. (2011). Empathy as "appraisal": Developing a new language-based approach to the exploration of clinical empathy. *Journal of Applied Linguistics and Professional Practice*, *7*, 139–162. https://doi.org/10.1558/japl.v7i2.145

Prakash, A. V., & Das, S. (2020). Intelligent conversational agents in mental healthcare services: A thematic analysis of user perceptions. *Pacific Asia Journal of the Association for Information Systems*, *12*, 1. https://doi.org/10.17705/1pais.12201

Putta, H., Daher, K., Kamali, M. E., Khaled, O. A., Lalanne, D., & Mugellini, E. (2022). Empathy scale adaptation for artificial agents: A review with a new subscale proposal. *8th International Conference on Control, Decision and Information Technologies*, Istanbul, Turkey, 2022 (pp. 699–704). https://doi.org/10.1109/CoDIT55151.2022.9803993

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5370–5381). https://doi.org/10.18653/v1/p19-1534

Reeves, S., Fischer, J. E., Porcheron, M., & Sikveland, R. (2019). Learning how to talk: Co-producing action with and around voice agents. *Mensch und Computer 2019—Workshopband*, Bonn: Gesellschaft für Informatik. https://doi.org/10.18420/muc2019-ws-654

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Smith, E. M., Boureau, Y.-L., & Weston, J. (2020). Recipes for building an open-domain chatbot. Proceedings of the *16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 300–325). https://doi.org/10.18653/v1/2021.eacl-main.24

Ruusuvuori, J. (2005). "Empathy" and "sympathy" in action: Attending to patients' troubles in Finnish homeopathic and general practice consultations. *Social Psychology Quarterly*, *68*, 204–222. https://doi.org/10.1177/019027250506800302

Saygin, A. P., & Cicekli, I. (2002). Pragmatics in human-computer conversations. *Journal of Pragmatics*, *34*, 227–258. https://doi.org/10.1016/s0378-2166(02)80001-7

Schegloff, E. A. (1986). The routine as achievement. *Human Studies*, *9*, 111–151. https://doi.org/10.1007/bf00148124

Sidnell, J. (2010). Conversation analysis. *Sociolinguistics and Language Education* (pp. 492–527). https://doi.org/10.21832/9781847692849-020

Silverman, D., & Peräkylä, A. (1990). Aids counselling: The interactional organisation of talk about 'delicate' issues. *Sociology of Health & Illness*, *12*, 293–318. https://doi.org/10.1111/1467-9566.ep11347251

Stotland, E. (1969). Exploratory investigations of empathy. In *Advances in Experimental Social Psychology* (pp. 271–314). Elsevier. https://doi.org/10.1016/s0065-2601(08)60080-5

Suchman, A. L., Markakis, K., Beckman, H. B., & Frankel, R. (1997). A model of empathic communication in the medical interview. *Jama*, *277*, 678–682. https://doi.org/10.1001/jama.1997.03540320082047

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge University Press.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference Advances in Neural Information Processing Systems* (pp. 3104–3112).

Svikhnushina, E., & Pu, P. (2020). Should machines feel or flee emotions? User expectations and concerns about emotionally aware chatbots. *arXiv preprint*. https://doi.org/10.48550/arXiv.2006.13883

Tannen, D. (1987). Repetition in conversation: Toward a poetics of talk. *Language* (pp. 574–605). https://doi.org/10.2307/415006

Urakami, J., Moore, B. A., Sutthithatip, S., & Park, S. (2019). Users' perception of empathic expressions by an advanced intelligent system. In *Proceedings of the 7th International Conference on Human-Agent Interaction* (pp. 11–18). https://doi.org/10.1145/3349537.3351895

Urakami, J., Sutthithatip, S., & Moore, B. A. (2020). The effect of naturalness of voice and empathic responses on enjoyment, attitudes and motivation for interacting with a voice user interface. In *International Conference on Human Computer Interaction* (pp. 244–259). https://doi.org/10.1007/978-3-030-49062-1_17

van Dijke, J., van Nistelrooij, I., Bos, P., & Duyndam, J. (2020). Toward a relational conceptualization of empathy. *Nursing Philosophy*, *21*(3). https://doi.org/10.1111/nup.12297

Weatherall, A., & Stubbe, M. (2015). Emotions in action: Telephone-mediated dispute resolution. *British Journal of Social Psychology*, *54*(2), 273–290. https://doi.org/10.1111/bjso.12082

Wilkinson, S., & Kitzinger, C. (2006). Surprise as an interactional achievement: Reaction tokens in conversation. *Social Psychology Quarterly*, *69*, 150–182. https://doi.org/10.1177/019027250606900203

Zaki, J., Bolger, N., & Ochsner, K. (2008). It takes two: The interpersonal nature of empathic accuracy. *Psychological Science*, *19*, 399–404. https://doi.org/10.1111/j.1467-9280.2008.02099.x

Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, *46*, 53–93. https://doi.org/10.1162/coli_a_00368

Zierau, N., Engel, C., Söllner, M., & Leimeister, J. M. (2020). Trust in smart personal assistants: A systematic literature review and development of a research agenda. *WI2020 Zentrale Tracks*, (pp. 99–114). https://doi.org/10.30844/wi_2020_a7-zierau

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.7

Seriously, What Did One Robot Say to the Other? Being Left out From Communication by Robots Causes Feelings of Social Exclusion

Astrid Rosenthal-von der Pütten[1] and Nikolai Bock[1]

1 Chair Individual and Technology, RWTH Aachen University, Germany

Abstract

While humans actually need some overt communication channel to transmit information, be it verbally or nonverbally, robots could use their network connection to transmit information quickly to other robots. This raises the question how this covert robot-robot communication is perceived by humans. The current study investigates how transparency about communication happening between two robots affects humans' trust in and perception of these robots as well as their feeling of being included/excluded in the interaction. Three different robot-robot communication styles were analyzed: silent, robotic language, and natural language. Results show that when robots transmit information in a robotic language (beep sounds) this leads to lower trust and more feelings of social exclusion than in the silent (i.e., covert) or natural language conditions. Results support the notion that humans are over-sensitive to signs of ostracism which seems to be detected in this style of overt but nonhuman robot-robot communication.

Keywords: robot-robot interaction, social exclusion, ostracism, trust

Introduction

With robots on the move to enter our work-related lives, human-robot interactions that involve multiple communicating robots could soon be a relevant and common situation. When looking into human-robot interaction, especially robots and humans interacting

CONTACT Astrid M. Rosenthal-von der Pütten • arvdp@itec.rwth-aachen.de • Chair Individual and Technology • RWTH Aachen University • Theaterplatz 14, 52062 Aachen, Germany

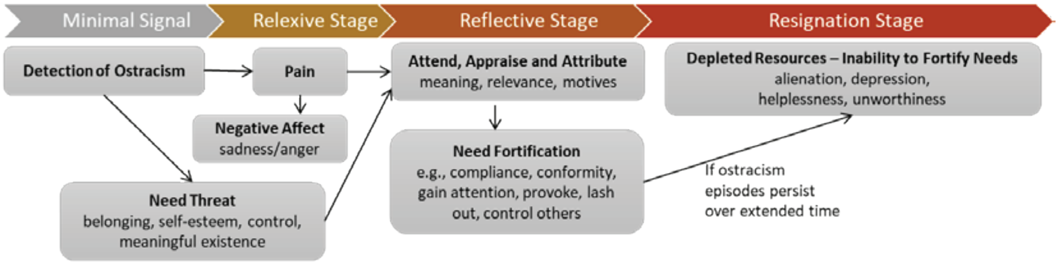
ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

in groups—most research so far concentrated on how group dynamics unfold in human-robot mixed groups and how robots can intervene in a positive way for instance to moderate conflicts between the humans or to include all group members. For instance, robots were successfully used to positively intervene and moderate working team conflicts (Martelaro et al., 2015) as well as conflicts between children (Shen et al., 2018). Moreover, robots can shape conversational dynamics for equal consideration of all group members' contributions during a discussion (Tennent et al., 2019). This need to positively intervene or moderate already exemplifies that sometimes individuals might experience conflicts in HRI groups, that they might feel largely ignored or excluded by other group members—potentially also robot group members as has recently been discussed by Rosenthal-von der Pütten and Abrams (2020). Consequently, the question arises how robots should communicate with each other, not only when humans interact with them directly, but also in the presence of observing humans who might be affected by the robots' behavior. While humans need some overt communication channel to transmit information, be it verbally or nonverbally, robots could use their network connection to transmit information quickly to other robots. This raises the question how this *covert* robot-robot communication is perceived by humans and especially whether humans feel excluded when robots use non-humanlike communication styles.

Theoretical Background

Humans have a fundamental need to belong; thus, having and maintaining good and long-lasting relationships with others is central to humans (Baumeister & Leary, 1995). Social exclusion threatens these crucial relationships with severe consequences for the affected individuals. Social exclusion is defined "as events and situations that signal a lack of social connections with others" and thus includes ostracism, devaluation, and social rejection (cf. Kawamoto et al., 2015, p. 1). People who experienced social exclusion show a variety of negative tendencies as they become aggressive, defensive, and self-defeating (e.g., make less rational, healthy choices; Twenge & Baumeister, 2004), uncooperative and unhelpful (e.g., help experimenter less after a mishap; Twenge et al., 2007), perform worse on tasks such as intellectual tests (Twenge & Baumeister, 2004), and show decreased self-regulation (e.g., give up early when confronted with a frustrating task; Baumeister et al., 2005). Social exclusion has also been shown to be related to decreased mental health (Nolan et al., 2003) and reduced survival rates (Holt-Lunstad et al., 2010). Individuals can experience interpersonal or intergroup social exclusion, the former targeting them as individuals and the latter as members of some outgroup. When experiencing social exclusion, individuals undergo several intra- and interpersonal processes. According to Williams' Temporal-Need-Threat-Model (Williams, 2009, cf. Figure 1), social exclusion causes a reflexive social pain response (activating similar brain regions as physical pain, cf. Eisenberger & Lieberman, 2004; Eisenberger et al., 2003) accompanied with negative affect (e.g., sadness, anger) and triggers threats to four fundamental needs: belonging, self-esteem, control over one's social environment, and meaningful existence. In a reflective stage, individuals' attention is directed to the social exclusion episode, and they reflect on its meaning and relevance. This may lead to coping responses such as compliance and conformity (to regain belongingness/self-esteem) or attracting attention, provoking, and attempts of controlling others (control/recognition)

FIGURE 1 Temporal-Need-Threat Model by Williams, 2009



to fortify the threatened needs. Persistent exposure to social exclusion over time consumes the resources necessary to motivate the individual to fortify threatened needs. Eventually, this leads to resignation, alienation, helplessness, and depression.

Humans tend to over-detect social exclusion. Empirical studies have shown that rational or logical characteristics of the social exclusion episode do not appear to moderate the detection of it. For instance, people felt ostracized when the source of ostracism were algorithms (Zadro et al., 2004). This hypersensitivity to social exclusion has its reason: the cost of perceiving social exclusion when it is not actually occurring (false alarm) is lower than the cost of a miss (not detecting that exclusion is happening). Thus, humans are extremely likely to detect social exclusion also in interactions with (especially anthropomorphized) artificial agents and experience and engage in the described reflexive and reflective processes. Indeed, first studies have shown that humans react sensitively to social rejection and social exclusion by robots. After playing a game of *Connect 4*, participants were informed by a humanoid robot that it would not like to see them again. Participants reported significantly reduced self-esteem relative to receiving no feedback or social acceptance (robot would like to see them again; Nash et al., 2018). Intentions for future use were not affected. Erel and colleagues (2021) implemented a robotic Cyberball game where participants played with two nonhumanoid robots. The robots either included (33% of ball tosses with three players), over-included (75% of tosses), or excluded (10% of tosses) the human player. Excluded participants reported lowered mood and experienced ostracism expressed via threatened needs, including control, belonging, and meaningful existence. In post-interaction interviews, many reported to feel "rejected," "ignored," and "meaningless." Fraune and Šabanović (2014) explored whether humans feel excluded when robots were exchanging information using beep sounds instead of natural language while participants were waiting for the experimenter of an unrelated study. Participants did not report differences in feeling excluded. However, participants might not have experienced the robots to be related to them in any way, thus, not experiencing a situation of social exclusion. This might be different if it were clear from the situational context that the two robots were communicating about the human(s) in the room.

Similar to findings in HHI, research in HRI and HMC has shown that social attributes such as perceived warmth, competence, or trustworthiness positively affect evaluations of and interactions with robots as well as usage intentions (Carpinella et al., 2017; A. Edwards et al., 2020; C. Edwards et al., 2021; Schaefer et al., 2012). A robot's ability to send social cues via its appearance, functionality, or behavior was identified as a crucial factor impacting this social perception (Duffy, 2003; Hegel, 2012; Schaefer et al., 2012).

Moreover, the communication style of AI systems has been found influential. For instance, in higher education courses in natural and social science students were more willing to accept an AI instructor-based education when the AI instructor is relational rather than functional in its communication style (Kim et al., 2020). How messages are formulated by robots are also important for the robot's evaluation regarding social attributes (e.g., A. Edwards et al., 2020). Since communication between robots in nonhuman language offers less opportunity to send clear social cues or to convey a communication style, such communication situations could lead not only to feelings of social exclusion, but also to decreased evaluations of the robots' social attributes.

In order to explore the socio-psychological effects of different styles of robot-robot communication, we created a scenario in which participants observed two robots interact and exchange information about a human. The robots were responsible for running an assessment center session of a human applicant during her application process, which participants could see in videos included in our online study. Communication styles varied in transparency (i.e., the amount of information provided about the content of the robots' conversation and thereby about how they function, behave, and reach decisions). The robots were either communicating covertly via their wireless network directly transmitting information from one robot to the other without making any sounds, or they communicated overtly, either in natural language or using a robotic language (beeps and clicks).

As argued above, we assume that communication in natural language offers more opportunity to send social cues (and for the user to perceive social cues) potentially positively influencing its social perception (Duffy, 2003; Hegel, 2012; Schaefer et al., 2012). Robots communicating silently or in robotic language, however, provide less or unfamiliar social cues assumingly leading to less favorable social perception. Moreover, the content of the robot-robot communication is not understandable in the silent and robotic language condition, potentially leading to lower trust in these conditions. Thus, we hypothesize:

> **H1:** Participants will trust the robots more (H1a) and perceive them as warmer (H1b), more competent (H1c), and less discomforting (H1d) in the natural language condition compared to the robotic language condition and silent condition.

Prior research provides evidence that humans experience social exclusion episodes when a robot directly rejected them (Nash et al., 2018) or when they were being left out of a game with two robotic players (Erel et al., 2021). Since our participants in the silent and robotic language conditions are not able to follow the robots' conversation they presumably will feel socially excluded. Hence, we hypothesize:

> **H2:** Participants will experience higher social exclusion when observing the robots on the robotic language condition (H2a) and silent condition (H2b) compared to the natural language condition.

Moreover, we want to explore whether the type of nonhumanlike robot-robot communication has an influence on humans' perception of the robots, their (dis)trust, and their feeling of social exclusion. We thus ask:

RQ1: Are human observers affected differently by a covert (silent) and a non-understandable overt (robotic language, e.g., beep sounds) communication style?

Method

Experimental Design

The present study is an online study consisting of an instruction followed by three short videos showing a human-robot interaction scenario in an assessment center (cf. procedure). The study followed a 3×1 between-subject design, with *transparency of robot-to-robot communication* as independent variable, operationalized through three different communication styles used by the two robots in the second video presented during the study. The following three conditions were compared:

Silent Communication. The robots exchange information covertly via their network. They do not explicitly acknowledge that information had been shared between them since they merely stand in front of each other without moving. In this condition, no overt interaction or communication is used.

Communication in Robotic Language. The robots overtly exchange information using a robot-like language which consists of nonlinguistic, stereotypical robot sounds, such as beeps and clicks. Additionally, the robots used human-like gestures and take turns in the nonlinguistic utterances emphasizing the impression of a conversation.

Communication in Natural Language. The robots overtly exchanged information using natural human language allowing the participants to understand everything they are saying. The robots update each other on the application process, transfer information about the applicants' performance, and point out what the following step in the procedure will be. While speaking with each other, the robots applied the same timing of turn-taking, conversation proportions, and human-like gestures as in the Robotic Language condition (cf. https://osf.io/hjm2t/ for transcript of utterances and the full videos as well as for the anonymized data set).

Stimulus Material

For the videos we used two humanoid robots from Aldebaran. While Pepper greets and guides applicants as well as discusses test results, Nao is responsible for conducting tests. Pepper is approximately 120 cm high and mobile in our setting (cf. Figure 2, picture on the right) while Nao is considerably smaller and placed stationary on a table next to the applicant (cf. Figure 2, picture on the left).

Procedure

Participants were randomly distributed to one of the three conditions. On the first page of the survey, participants were informed about the upcoming task, data protection, and their right to withdraw from the study at any time. They verified that they were above 18 years and gave informed consent by clicking on the start button. Participants first provided demographic data (gender, age, occupation) and were asked to describe pre-experiences with robots if applicable. Next, participants were asked about their negative attitudes toward robots (Negative Attitudes toward Robots Scale, Nomura et al., 2006) and their affinity for technology (Affinity for Technology Interaction Scale, Franke et al., 2019; cf. section measures).

Afterward, instructions explained the scenario the participants would take part in. To help put themselves in the position of the situation and identify with the role of the applicant in the video, they were informed about the German software development company GDQ-Technologies where they applied for an open position in the management of the development department. GDQ-Technologies invited them to an interview and an Assessment-Centre, which would be performed by two robots. The full instruction was:

> Please put yourself in the shoes of an applicant who is interested in an open position in a large German high-tech company called GDQ-Technologies. This is an important position in the management of the software development department.
>
> Your tasks as part of the management team would be:
>
> - Cooperation with software developers
> - Management of the development process
> - Coordination of the quality inspection of new software
>
> You applied with your résumé and were then invited to the GDQ-Technologies assessment center for an interview and to test your suitability. When you get there, a robot greets you, introduces itself by the name "Pepper," and explains that it will guide you through the entire application process. He will then accompany you to an office where you will be interviewed, and a few aptitude tests will be carried out.
>
> After you have been told that you did this well, Pepper leads you to another room where you should take another psychological test. Because of its abilities, it is part of the job of a second robot called "Nao" to conduct the test with you.

Participants were informed that they would now see one part of the assessment center in three videos. First, they would see a video (the same video in all conditions) of the Nao robot performing a psychological attention and stress test with the applicant. After that, participants read a short instruction that Pepper re-entered the room to pick up the applicant for the rest of the application process. Afterward, participants experienced one of the three experimental videos observing the two robots communicate silently, in robotic language, or natural language, depending on the condition they were assigned to. Following this, written explanations informed participants that they would receive some personal feedback about their performance, which could be seen in the third video which was the same for all conditions (cf. Figure 2).

Immediately following this last video, participants completed questionnaires assessing their perception of the robots (trust, competence, warmth, and discomfort) and whether they experienced social exclusion during the communication between the two robots.

FIGURE 2 Manipulation Was Included in the Second Video in Which the Robots Communicated Silently (Covertly via Their Network), in Robotic Language (Overtly Using Beeps and Clicks) or in Natural Language (Overtly Using Natural Language and Gestures).

At the end of the survey, the manipulation check was performed, and participants could respond to open-ended questions regarding the interaction (cf. section measures). Finally, participants were debriefed.

Measures

Dependent Variables

Trust. Participants' trust in the robots was measured using the Trust in Automated Systems Survey by Jian et al. (2000). This scale is unique in that it measures both trust and distrust as polar opposites along a single dimension rather than simple unidimensional trust as, for instance, it is the case in the Trust Perception Scale HRI (Schaefer, 2016). Moreover, the latter scale is regarded as less adequate since it also includes items that are measuring social perceptions regarding competence and warmth, thereby potentially mixing too many concepts into one (very long) scale. The Trust in Automated Systems Survey, however, is short and delivers separate values for the trust and distrust dimensions. The items sampling distrust, for instance, measure the perception of the automation's deceptive nature or the likelihood of harmful outcomes if it is used (for a discussion of trust measurements see also Kohn et al., 2021). The 12 items were slightly adapted exchanging the "system" with "robots" and were rated on a 7-point Likert scale from "strongly agree" to "strongly disagree" (e.g., "The robots are deceptive," Distrust Cronbach's alpha = .778; M = 3.10, SD = 0.138; Trust Cronbach's alpha = .805; M = 4.05, SD = 0.62).

ROSAS. We captured participants' views on the robots' social properties for each robot individually using the 18-item Robotic Social Attributes Scale (RoSAS) by Carpinella et al. (2017). For each robot, participants were asked to complete the full inventory with the three sub-scales warmth (items: feeling, happy, organic, compassionate, social, emotional; Pepper Cronbach's alpha = .862, M = 3.15, SD = 1.50; Nao Cronbach's alpha = .901, M = 3.11, SD =1 .62), competence (items: knowledgeable, interactive, responsive, capable, competent, reliable; Pepper Cronbach's alpha = .873, M = 5.62, SD = 1.63; Nao Cronbach's alpha = .885, M = 5.61, SD = 1.67), and discomfort (items: awkward, scary, strange, awful, dangerous, aggressive; Pepper Cronbach's alpha = .776, M = 3.14, SD = 1.48; Nao Cronbach's alpha = .856, M = 2.87, SD = 1.72). Participants responded on a 9-point Likert scale ranging from "definitely not associated" to "definitely associated."

Social Exclusion. In order to capture whether participants felt socially excluded during the conversation between the robots, we created five ad-hoc items rated on a 5-point Likert scale (Cronbach's alpha = .842, *M* = 2.73, *SD* = 1.15; items: While the two robots were interacting . . . " . . . I felt uncomfortable," " . . . I felt nervous," " . . . I had the feeling that the robots were talking about me," " . . . I felt excluded," " . . . I felt that the robots don't want me to know what they are talking about").

Moderating Variables

Previous HRI research has shown that negative attitudes toward robots might have a moderating effect on interaction with and perception of robots (Nomura et al., 2006; Sanders et al., 2017). Moreover, Franke et al. (2019) argue that affinity for technology is a key personal resource for successful interaction with technology. It might, therefore, affect how participants engage in and perceive the interaction with robot technology. Consequently, we assume affinity for technology and negative attitudes toward robots may have impacts on trust in robots and social perception.

Negative Attitudes Toward Robots. To measure participants' general negative attitudes toward robots, we employed the Negative Attitudes toward Robots Scale (NARS) created by Nomura et al. (2006). The 14 items on the three subscales were rated on a 5-point Likert scale ranging from "I do not agree at all" to "I completely agree" (S1—Negative attitude toward situations of interaction with robots, six items, Cronbach's alpha = .741, *M* = 2.15, *SD* = 0.73; S2—Negative attitude toward social influence of robots, five items, Cronbach's alpha = .681, *M* = 3.00, *SD* = 0.78; S3—Negative attitude toward emotions in interaction with robots, three items, Cronbach's alpha = .612, *M* = 3.39, *SD* = 0.85).

Affinity for Technology. We captured participants' general affinity for technology using the Affinity for Technology Interaction Scale (ATI) from Franke et al. (2019) consisting of nine items which are measured using a 6-point Likert scale from "completely disagree" to "completely agree" (e.g., "I like testing the functions of new technical systems"; Cronbach's alpha = .947, *M* = 3.69, *SD* = 1.23).

Open-Ended Questions and Manipulation Check. For data cleansing purposes we included two test statements to verify that participants' answers matched the conditions they were assigned to asking (i) "Could you hear that the robots were communicating with each other in the second video?" (yes/no) and (ii) "Did you understand what the robots were talking about in the second video?" Following the manipulation check participants had to answer open-ended questions asking whether and if yes, which kind of information was exchanged between the robots. Answers were checked for plausibility. Twenty-one participants gave a deviant answer from their assigned condition (e.g., stating that they could understand what the robots were saying although in the "silent" condition). However, their answers to open-ended questions proved they misinterpreted the question (i.e., thinking it referred to the robots talking in general in the three videos). Hence their data remained in the data set.

Participants

The study was advertised among university students and via social networking sites such as Facebook and Instagram. In total, 183 volunteers took part. The data cleansing procedure yielded 176 participants (71 male, 103 female, 2 diverse) with a mean age of 34.7 (*SD*

TABLE 1 Distribution of Participants Across Conditions

	Silent	Robotic	Natural	Total
Male	23	19	29	71
Female	33	38	32	103
Diverse	0	0	2	2
Total	**56**	**57**	**63**	**176**

= 13.49; range = 18–72 years, based on 176 participants). Seventy-two were employed, 20 self-employed, 73 students, 1 retired, 4 university lecturers, 2 people in an apprenticeship, 2 were stay-at-home parents. Table 1 shows the distribution of participants across conditions.

Results

Testing Assumptions for ANOVA and ANCOVA

All dependent variables were tested for homogeneity of variance. Levene's tests were not significant except for the Trust subscale. Kolmogorov-Smirnov tests indicated for all dependent variables that data was not normally distributed (see Appendix for values of skew and kurtosis). Since visual inspection showed that the skewness was equal between groups for Trust and Distrust, Competence (Nao & Pepper), and Discomfort (Nao & Pepper) this violation of normality can be ignored for these variables (cf. Field & Wilcox, 2017). However, Warmth (Nao & Pepper) as well as Social Exclusion shows different skewness between conditions. As a result, we will perform Kruskal-Wallis tests instead of ANOVAS when assumptions are not met. For the planned ANCOVAS, the assumption of homogeneity of regression slopes was violated for Discomfort (Pepper) for the ATI score, for Competence (Nao) for NARS-S1, and for most dependent variables except Discomfort (Nao & Pepper) and Social Exclusion for NARS-S2. Homogeneity of regression slopes was given for all dependent variables for NARS-S3. The covariates are independent of the manipulation effect, meaning there is no interaction between the covariates (ATI, NARS-S1, NARS-S2, NARS-S3) and the independent variable.

(Dis)Trust

To test whether participants trusted the robots more in the natural language condition compared to the robotic language condition and silent condition (H1a), we calculated a Kruskal-Wallis tests. There was a significant effect of communication style on trust, $H(2) = 7.05$, $p = .029$. Post-hoc tests (all Bonferroni corrected) revealed that the robotic language significantly elicited lower trust than the natural language ($U = 1312$, $p = .011$, $r = -.23$), while there were no significant differences between natural language and silent ($U = 1514$, $p = .183$, $r = -.12$) and no difference between silent and robotic language or robotic ($U = 1335$, $p = .134$, $r = -.14$).

The same analysis was performed with the subscale distrust. There was a significant effect of communication style on distrust, $H(2) = 11.34$, $p = .003$. Post-hoc tests (all

Bonferroni corrected) revealed that the robotic language significantly elicited higher distrust than the natural language ($U = 1163$, $p < .001$, $r = -.31$), while there were no significant differences between natural language and silent ($U = 1506$, $p = .169$, $r = -.15$) and no difference between silent and robotic language or robotic ($U = 1255$, $p = .050$, $r = -.18$).

Robot Perception

To test whether participants evaluated the robots as warmer (H1b), more competent (H1c), and less discomforting (H1d) in the natural language condition compared to the robotic language condition and silent condition, we calculated Kruskal-Wallis tests for warmth (Nao & Pepper) and ANCOVAs for competence and discomfort (Nao & Pepper). There were no significant effects for warmth, competence, and discomfort. However, for both Nao and Pepper, we found the tendency that participants perceived them as more discomforting in the robotic language condition (cf. Table 2 for descriptives); Pepper, $F(2,174) = 2.662$, $p = .073$, ${}_p\eta^2 = .030$; Nao, $F(1,174) = 2.926$, $p = .056$, ${}_p\eta^2 = .033$. Post-hoc tests (all Bonferroni corrected) were not significant. From the covariates, the subscale "S1—Negative attitude toward situations of interaction with robots" was significantly related to discomfort for Pepper, $F(1,175) = 10.899$, $p = .001$, and to discomfort for Nao, $F(1, 175) = 14.903$, $p < .001$.

Social Exclusion

To test whether participants experience higher social exclusion when observing the robots on the robotic language condition (H2a) and silent condition (H2b) compared to the natural language condition, we calculated a Kruskal-Wallis tests. There was a significant effect

TABLE 2 Mean Values and Standard Deviations for Dependent Variables Across Conditions

	Silent	Robotic	Natural	Total
	M (SD)	*M (SD)*	*M (SD)*	*M (SD)*
Trust*	4.31 (1.10)	4.07 (.99)	4.54 (1.33)	4.31 (1.17)
Distrust*	3.06 (1.10)	3.47 (1.09)	2.80 (1.07)	3.10 (1.11)
Social Exclusion*	2.16 (0.92)	3.81 (0.84)	2.24 (0.85)	2.73 (1.15)
Perception of Pepper				
Warmth	3.16 (1.40)	2.96 (1.33)	3.32 (1.73)	3.15 (1.50)
Competence	5.66 (1.64)	5.53 (1.52)	5.66 (1.74)	5.62 (1.63)
Discomfort	3.05 (1.40)	3.44 (1.67)	2.94 (1.35)	3.14 (1.48)
Perception of Nao				
Warmth	3.22 (1.60)	2.84 (1.48)	3.28 (1.76)	3.11 (1.62)
Competence	5.66 (1.64)	5.53 (1.52)	5.66 (1.74)	5.62 (1.63)
Discomfort	2.76 (1.69)	3.25 (2.00)	2.62 (1.41)	2.87 (1.72)
Note: significant effects are marked with *				

of communication style on social exclusion, $H(2) = 71.11$, $p < .001$. Post-hoc test (all Bonferroni corrected) revealed that the robotic language significantly differed from the natural language ($U = 366.5$, $p < .001$, $r = -.68$), and the silent condition ($U = 359$, $p < .001$, $r = -.66$), while there were no significant differences between natural language and silent ($U = 1546$, $p = .313$, $r = -.09$).

Analysis of Answers to Open-Ended Questions

Our open-ended questions included whether people were aware that information had been exchanged between the robots. In all three conditions, most participants were aware that information had been exchanged between robots though we see a clear difference between conditions (natural language: 90%; robotic language: 80%; silent: 54%). In the natural language condition, participants mostly repeated what they overheard in the video, that the two robots were talking about the procedure of the assessment center and that the applicant in the video just completed a specific test. In the silent condition, about half of the people were not sure whether information has been submitted and if such information was submitted, they assumed it would be test results, not details about the procedure of the assessment center. In the robotic language condition 80% of those who thought information was shared stated it would be test results. In the silent condition, this was the case for 70% of participants who previously stated that information has been shared.

Discussion

The presented study investigated how different styles of robot-robot communication are perceived by humans. In contrast to humans, robots have the ability to silently exchange information via wireless networks. Do humans feel left out and trust robots less when they recognize that information about them has been exchanged via unobservable channels of communication? To explore the socio-psychological effects of different styles of robot-to-robot communication, participants in our online study watched videos observing two robots that interact and exchange information (prior and upcoming parts of assessment center and information a test has been completed) about a human who completed tests in an assessment center session. The robots were either communicating covertly via their wireless network directly transmitting information from one robot to the other without making any sounds, or they communicated overtly, either in natural language or using a robotic language (beeps and clicks).

Effect of Robot-Robot Communication Style on Social Perception and Trust

We assumed that when robots communicate in natural language, they send more social cues which potentially leads to a more favorable social perception by the human observers (Duffy, 2003; Hegel, 2012; Schaefer et al., 2012) in contrast to situations in which the content of their information exchange is not understandable for humans as is the case in the silent or robotic language conditions. More precisely, we hypothesized that participants would trust the robots more (H1a) and perceive them as warmer (H1b), more competent (H1c), and less discomforting (H1d) in the natural language condition compared to the robotic language condition and silent condition. Our results only partly supported our hypotheses. While robots communicating in beeps and clicks were trusted less compared to the natural language condition (lower trust and higher distrust), trust was not significantly different for the silently communicating robots. This effect is not due to a wrong assessment of the situation on the participants' side. Most participants stated in open-ended check questions that they were aware that information has been transmitted—also in the silent condition with still 54% stating some information has been transmitted. Rather it seems that if participants cannot hear and/or understand what is being said, they largely assumed that test performance information (i.e., the applicant in the video) was exchanged instead of information on the procedure of the assessment center. Evaluations of the robots regarding warmth and competence were not affected by their communication style; however, we found a descriptive (not significant) tendency that robots communicating in robotic language were perceived as more discomforting. This is interesting since several participants stated in the open-ended interviews that the silence in the silent condition was awkward and discomforting. However, observing communication and not being able to understand it was obviously more unsettling as the results regarding feelings of social exclusion show.

Effect of Robot-Robot Communication Style on Feelings of Social Exclusion

Based on previous studies by Erel et al. (2021) and Nash et al. (2018) we assumed that also in interactions with robots, the human hypersensitivity to ostracism cues (Zadro et al., 2004) will result in experiencing a social exclusion episode in the silent and robotic language condition. While previous studies worked with directly formulated rejection by the robot (Nash et al., 2018) or excluding participants in a Cyberball game (Erel et al., 2021), we created a scenario where participants were left out of the robot-robot communication. In line with our hypothesis, we found a strong significant effect for social exclusion. Participants experienced higher social exclusion when observing the robots on the robotic language condition compared to the natural language condition (H2a) and unexpectedly also in comparison to the silent condition. Again, no difference was found between the natural language condition and the silent condition (H2b). Hence, we can constitute that in our study human observers were indeed affected differently by a covert (silent) and a non-understandable overt (robotic language) communication style. It seems that the usage of beep sounds for communication is a strong trigger for ostracism detection, while obviously transmitting information silently is not. However, this effect might also be context dependent. In the context of our study, three participants in the robotic language condition stated that wireless communication might be quicker and easier in the assessment center scenario and would save the applicant time, so why bother with clicks and beeps. But it is conceivable that in less formal situations like being at a friend's house who coincidentally has two robots at home chatting with you, obvious silent communication between the robots might also trigger ostracism detection. Our interpretation of the found social exclusion effect is that the non-understandable robotic language hurts more, because it is perceived as doing this for the reason of social exclusion rather than for robotic efficiency in processing information. The comments in the open-ended questions (what did you like or dislike about the interaction in the video?) seem to support this. Several participants mentioned in the robotic language condition that they experienced a feeling of social exclusion: "Dislike: feeling of being excluded," "I didn't like that they obviously communicated about me, but I didn't understand what." Others wondered "why they did not use a language I can understand," or explicitly stated they disliked "robots beeping when interacting with each other instead of human speech." One participant directly contrasted the silent and robotic communication style: "Dislike: exchange between the robots, which in my opinion should have happened either silently or in a language in which I, as an applicant, could understand what was being said." Indeed, as mentioned before three participants mentioned that a silent communication would be more efficient. To our surprise the robot-robot communication style did not influence perceptions of warmth which might be expected given that participants felt excluded. But generally, warmth ratings were rather low and had high standard variations. This could also be due to the setting of the situation and the social roles of the robots (Oliveira et al., 2019). Both robots acted as formal unknown interviewers in an assessment center and not as peers, friends, or colleagues. This professional social distance could explain the generally low ratings in warmth and might also explain the similar warmth ratings between communication styles. Unfortunately, we cannot relate these findings to previous social exclusion studies directly because those studies did not measure how the robots were perceived regarding warmth, competence, and discomfort. However, the direct rejection that participants experienced in Nash et al.'s study (2018) lowered self-esteem (i.e., participants showed need threat). But their rejection did not affect their willingness for future interaction.

Limitations and Future Directions

In contrast to previous studies on social exclusion in HRI, our study did not involve direct interaction with a physically present robot, but participants had to self-project themselves into what was displayed in the videos. While this constitutes a limitation of our study, we still found a quite strong effect on feelings of social exclusion. Interestingly, some participants seemed to self-project very strongly answering in the open-ended questions with self-referring statements such as "they talked about my test results" or "they talked about where I go next." Potentially, effects in live interactions will be even stronger. Some participants mentioned that the scenario itself, an assessment center, is not an area for which they regard robots as useful, since applicants might feel strange and disconnected. While this does not necessarily limit the study results, it is relevant for future studies, rendering how important it is to create realistic and meaningful future applications also in our experimental studies. We observed that our manipulation check questions were in part misinterpreted by study participants although they explicitly referred to the second, manipulated, video. Some participants seemed to consider all three videos when answering these questions ("Could you hear that the robots were communicating with each other in the second video?"; " Did you understand what the robots were talking about in the second video?"). This became apparent when checking their answers to the three open-ended questions. For instance, one participant in the silent condition answered both questions with yes but described how awkward it was to observe the two silent robots in the second video. Hence, only the combination of the closed and open questions was reliable checking for successful manipulation. Moreover, as mentioned previously, the trust and social exclusion effects might be context-dependent and thus generalization to situations in different social settings should be addressed in future research.

Social exclusion is very likely to happen in HRI, because robots have components known to be biased (Howard & Borenstein, 2017; Righetti et al., 2019). For instance, face recognition is better for White people than for people of color and natural language recognition is better for male than female language users, not to speak of variations in language such as regional or foreign accents, or colloquial language or jargon. Moreover, Rosenthal-von der Pütten and Abrams (2020) discussed that users "might have more or less time or might be more or less motivated to provide these interactions [with robots] that are needed for [machine] learning" (pp. 400–401). Meaning that if a robot interacts with multiple users, it might perform better in user modelling for some users (which provided much training data) and worse for others (with less training data) resulting in different subsequent interactions which could easily be perceived as biased or excluding. Zou and Schiebinger (2018) emphasized the pressing need to make AI and thus also robots fairer by identifying biases and implementing strategies to diminish bias. In this vein, Rosenthal-von der Pütten and Abrams discussed how robots might analyze participant behavior to detect if a social exclusion episode has happened and enable them to engage in repair mechanisms. In consequence, investigating when, in which scenarios, and how people are experiencing social exclusion in HRI and how they are reacting within and after exclusion episodes is not only interesting regarding generalizability of results, it can inform future developments in explainable robot behavior, positively shaping social dynamics in human-robot group situations.

Conclusion

Our video-based online study explored how different styles of robot-robot communication are perceived by humans comparing humanlike communication via natural language to silent communication via wireless connection and communication in a robotic language based on beeps and clicks. The study results suggest that when robots transmit information in a robotic language this leads to lower trust and more feelings of social exclusion than in the silent or natural language conditions. Like previous laboratory work in which participants were either directly verbally rejected or excluded from a variation of the cyberball game, our participants were very sensitive too to signs of ostracism which seems to be detected in this style of overt but nonhuman robot-robot communication. Completely leaving out humans from a communication loop (silently transmitting information), however, did not negatively impact observers. These quantitative results are reflected in participants' comments showing that participants were overall aware that information had been shared between the robots but had different assumptions of what kind of information had been shared and why this was done covertly (i.e., participants in the robotic language condition disliked to be the topic of a secret conversation between the robots and felt being left out). Given the very specific social setting and the connected social roles, two robots working in an assessment center, we assume that social exclusion effects might also occur for silently communicating robots in less professional contexts. Hence, future research is needed to explore social exclusion across different situational contexts.

Author Biographies

Astrid Rosenthal-von der Pütten is a Full Professor and Director of the Chair Individual and Technology at the Department of Society, Technology, and Human Factors at RWTH Aachen University. She received her PhD in psychology as well as BSc and MSc degrees in applied cognitive and media science from the University of Duisburg-Essen, Germany. Her research interests include social effects of artificial entities, human-robot/agent interaction, social dynamics in human-robot groups, linguistic alignment with robots and virtual agents, as well as shared decision-making with algorithmic systems.

https://orcid.org/0000-0002-2497-143X

Nikolai Bock is a PhD Candidate at the Chair for Individual and Technology at the Department of Society, Technology, and Human Factors at RWTH Aachen University. He received his BSc and MSc degree in applied cognitive and media science from the University of Duisburg-Essen, Germany. His research focuses on how robot appearance and behavior influence perception of and interaction with social robots, how individual attitudes affect the perception and adoption of algorithmic-decision support systems, as well as the development of psychological scales for use in HRI/HCI.

https://orcid.org/0000-0002-4668-9523

References

Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Twenge, J. M. (2005). Social exclusion impairs self-regulation. *Journal of Personality and Social Psychology, 88*(4), 589–604. https://doi.org/10.1037/0022-3514.88.4.589

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological bulletin*, *117*(3), 497.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017, March). The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 254–262). https://doi.org/10.1145/2909824.3020208

Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and autonomous systems, 42*(3–4), 177–190. https://doi.org/10.1016/S0921-8890(02)00374-3

Edwards, A., Edwards, C., & Gambino, A. (2020). The social pragmatics of communication with social robots: Effects of robot message design logic in a regulative context. *International Journal of Social Robotics*, *12*(4), 945–957. https://doi.org/10.1007/s12369-019-00538-7

Edwards, C., Edwards, A., Albrehi, F., & Spence, P. (2021). Interpersonal impressions of a social robot versus human in the context of performance evaluations. *Communication Education*, *70*(2), 165–182. https://doi.org/10.1080/03634523.2020.1802495

Eisenberger, N. I., & Lieberman, M. D. (2004). Why rejection hurts: A common neural alarm system for physical and social pain. *Trends in Cognitive Sciences, 8*(7), 294–300. https://doi.org/10.1016/j.tics.2004.05.010

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An FMRI study of social exclusion. *Science (New York, N.Y.), 302*(5643), 290–292. https://doi.org/10.1126/science.1089134

Erel, H., Cohen, Y., Shafrir, K., Levy, S. D., Vidra, I. D., Shem Tov, T., & Zuckerman, O. (2021). Excluded by robots. In C. Bethel, A. Paiva, E. Broadbent, D. Feil-Seifer, & D. Szafir (Eds.), *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 312–321). ACM. https://doi.org/10.1145/3434073.3444648

Field, A. P., & Wilcox, R. R. (2017). Robust statistical methods: A primer for clinical psychology and experimental psychopathology researchers. *Behaviour research and therapy, 98*, 19–38. https://doi.org/10.1016/j.brat.2017.05.013

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction, 35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Fraune, M. R., & Šabanović, S. (2014). Negative attitudes toward minimalistic robots with intragroup communication styles. In *The 23rd IEEE International Symposium on Robot and Human Interactive Communication,* 2014: 25–29 Aug. 2014, Edinburgh, Scotland, UK (pp. 1116–1121). IEEE. https://doi.org/10.1109/ROMAN.2014.6926401

Hegel, F. (2012). Effects of a robot's aesthetic design on the attribution of social capabilities. In *the 21st IEEE International Symposium on Robot and Human Interactive Communication,* pp. 469–475. https://doi.org/10.1109/ROMAN.2012.6343796

Holt-Lunstad, J., Smith, T. B., & Layton, J. B. (2010). Social relationships and mortality risk: A meta-analytic review. *PLoS medicine, 7*(7), e1000316. https://doi.org/10.1371/journal.pmed.1000316

Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics, 24*(5), 1521–1536. https://doi.org/10.1007/s11948-017-9975-2

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics, 4*(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04

Kawamoto, T., Ura, M., & Nittono, H. (2015). Intrapersonal and interpersonal processes of social exclusion. *Frontiers in Neuroscience, 9*, 62. https://doi.org/10.3389/fnins.2015.00062

Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human–Computer Interaction*, *36*(20), 1902–1911. https://doi.org/10.1080/10447318.2020.1801227

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y. C., & Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.604977

Martelaro, N., Jung, M., & Hinds, P. (2015). Using robots to moderate team conflict. In J. A. Adams (Ed.), *Proceedings of the Tenth Annual ACMIEEE International Conference on Human-Robot Interaction Extended Abstracts* (p. 271). ACM. https://doi.org/10.1145/2701973.2702094

Nash, K., Lea, J. M., Davies, T., & Yogeeswaran, K. (2018). The bionic blues: Robot rejection lowers self-esteem. *Computers in Human Behavior, 78*, 59–63. https://doi.org/10.1016/j.chb.2017.09.018

Nolan, S. A., Flynn, C., & Garber, J. (2003). Prospective relations between rejection and depression in young adolescents. *Journal of personality and social psychology, 85*(4), 745. https://doi.org/10.1037/0022-3514.85.4.745

Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human–robot interaction. *Ai & Society, 20*(2), 138–150. https://doi.org/10.1007/s00146-005-0012-7

Oliveira, R., Arriaga, P., Correia, F., & Paiva, A. (2019). The stereotype content model applied to human-robot interactions in groups. In *Hri'19: The 14th ACM/IEEE International Conference on Human-Robot Interaction: March 11–14, 2019, Daegu, South Korea* (pp. 123–132). IEEE. https://doi.org/10.1109/HRI.2019.8673171

Righetti, L., Madhavan, R., & Chatila, R. (2019). Unintended consequences of biased robotic and artificial intelligence systems [ethical, legal, and societal issues]. *IEEE Robotics & Automation Magazine, 26*(3), 11–13. https://doi.org/10.1109/MRA.2019.2926996

Rosenthal-von der Pütten, A., & Abrams, A. M. (2020, July). Social dynamics in human—robot groups–Possible consequences of unequal adaptation to group members through machine learning in human-robot groups. In *International Conference on Human-Computer Interaction* (pp. 396–411). Springer, Cham. https://doi.org/10.1007/978-3-030-50334-5_27

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., & Hancock, P. A. (2017). Trust and prior experience in human-robot interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1), 1809–1813. https://doi.org/10.1177/1541931213601934

Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the "trust perception scale-HRI." In *Robust intelligence and trust in autonomous systems* (pp. 191-218). Boston, MA: Springer US.

Schaefer, K. E., Sanders, T. L., Yordon, R. E., Billings, D. R., & Hancock, P. A. (2012). Classification of robot form: Factors predicting perceived trustworthiness. *Proceedings of the human factors and ergonomics society annual meeting, 56*(1), 1548–1552. https://doi.org/10.1177/1071181312561308

Shen, S., Slovak, P., & Jung, M. F. (2018, February). "Stop. I see a conflict happening." A robot mediator for young children's interpersonal conflict resolution. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction* (pp. 69–77). https://doi.org/10.1145/3171221.3171248

Tennent, H., Shen, S., & Jung, M. (2019). Micbot: A peripheral robotic object to shape conversational dynamics and team performance. In *Hri'19: The 14th ACM/IEEE International Conference on Human-Robot Interaction: March 11–14, 2019, Daegu, South Korea* (pp. 133–142). IEEE. https://doi.org/10.1109/HRI.2019.8673013

Twenge, J. M., & Baumeister, R. F. (2004). Social exclusion increases aggression and self-defeating behavior while reducing intelligent thought and prosocial behavior. In D. Abrams, M. A. Hogg, & J. M. Marques (Eds.), *Social psychology of inclusion and exclusion* (pp. 45–64). Psychology Press. https://doi.org/10.4324/9780203496176-6

Twenge, J. M., Baumeister, R. F., DeWall, C. N., Ciarocco, N. J., & Bartels, J. M. (2007). Social exclusion decreases prosocial behavior. *Journal of Personality and Social Psychology, 92*(1), 56. https://doi.org/10.1037/0022-3514.92.1.56

Williams, K. D. (2009). Ostracism: A temporal need-threat model. In M. P. Zanna (Ed.), *Advances in Experimental Social Psychology* (Vol. 41, pp. 275–314). Academic Press. https://doi.org/10.1016/S0065-2601(08)00406-1

Zadro, L., Williams, K. D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology, 40*(4), 560–567. https://doi.org/10.1016/j.jesp.2003.11.006

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature, 559*(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.8

Triggered by Socialbots: Communicative Anthropomorphization of Bots in Online Conversations

Salla-Maaria Laaksonen[1], **Kaisa Laitinen**[2], **Minna Koivula**[2], **and Tanja Sihvonen**[3]

1 Centre for Consumer Society Research, University of Helsinki, Finland
2 Department of Language and Communication Studies, University of Jyväskylä, Finland
3 School of Marketing and Communication, University of Vaasa, Finland

Abstract

This article examines communicative anthropomorphization, that is, assigning of human-like features, of socialbots in communication between humans and bots. Situated in the field of human-machine communication, the article asks how socialbots are devised as anthropomorphized communication companions and explores the ways in which human users anthropomorphize bots through communication. Through an analysis of two datasets of bots interacting with humans on social media, we find that bots are communicatively anthropomorphized by directly addressing them, assigning agency to them, drawing parallels between humans and bots, and assigning emotions and opinions to bots. We suggest that socialbots inherently have anthropomorphized characteristics and affordances, but their anthropomorphization is completed and actualized by humans through communication. We conceptualize this process as communicative anthropomorphization.

Keywords: socialbots, communication, anthropomorphization, social interaction, social media

Introduction

In the film *Cast Away*, a FedEx executive played by Tom Hanks develops an unlikely friendship with a volleyball, "Wilson," after washing up on a desert island following a plane crash. By describing the emotional and conversational bond between an isolated character and

CONTACT Salla-Maaria Laaksonen • salla.laaksonen@helsinki.fi • P.O. Box 16 • FIN-00014 • University of Helsinki, Finland

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

sporting equipment, the movie explores the fundamental human need for social affiliation in a way most viewers can identify with (Greenwood & Long, 2011). Through reconfiguring the dialogic situation with what the volleyball affords, Hanks's character is able to preserve his mental health, motivation, and sense of direction. Wilson the volleyball is far from a perfect conversational agent, but illustrates the basic human need for *anthropomorphization*, giving human attributes to nonhuman entities, and shows how humans adopt and reinterpret the affordances of nonhuman and technological objects and interact with them to fulfill their communicative needs.

This study explores technological, nonhuman agents—socialbots—as interlocutors in text-based online communication. Much like "Wilson," they are not perfect conversationalists even if they are designed to incorporate human features. Although bots have inhabited online spaces for decades, the recent development of natural language processing has introduced bots capable of producing human-like language and engaging in communication with human users. These bots are usually referred to as socialbots or chatbots (Grimme et al., 2017), or, more recently, as communicative AIs (Guzman & Lewis, 2020), and they are becoming increasingly common in contexts from customer service to intra-organizational communication. Our theoretical frame draws from the literature of *human-machine communication* (*HMC*, Guzman & Lewis, 2020; Jones, 2014), which seeks to explore the role of machines as communicators, not only mediators of human communication. By relying on the HMC framework, we investigate the anthropomorphization of chatbots when they are built to act as communication partners. Existing studies have shown how bots are designed to appear human-like and to have personalities (Araujo, 2018; Gorwa and Guilbeault, 2020; Grimme et al., 2017), but our interest is on the process of communication in sociotechnical settings where humans and bots interact. Hence, we ask:

> **RQ1:** How are bots devised as anthropomorphized communication companions?
>
> and
>
> **RQ2:** In what ways do human users anthropomorphize bots through communication?

Similar questions have previously been explored mainly theoretically or through experimental design, and there are only a few examples of utilizing data of actual, naturally occurring human-machine communication. In this study, we explore the communicative interaction between bots and humans empirically in two contexts: an organizational setting and public social media setting. We focus on two cases: *Slackbot*, an interactive, programmable bot on a Slack platform used by a distributed team within a media organization, and *LovebotBlue*, a bot that was part of a food and confectionary producer's ad campaign designed to battle hate speech online. Our data covers several platforms, and thus our analysis results in a broader view to human-bot communication than most studies by allowing for a comparison between an internal chat platform and public social media. We contribute to existing literature by introducing *communicative anthropomorphization* as a key aspect in human-machine communication and by utilizing empirical datasets to deepen the current, often theoretical or experimental understanding of the topic. We argue that anthropomorphization is not only a design feature or a psychological process, but also a feature of a communicative process in which humans and nonhumans participate with their distinctive capabilities and affordances.

Theoretical Background: Socialbots as Communicative AI

Interactive bots and conversational agents have been studied across disciplines. They can be defined as automated programs that manifest on a particular platform through an account that looks like a regular human user (e.g., Grimme et al., 2017). Bots perform simple functions and usually reply when addressed. The development of computer-assisted conversational agents started with the psychotherapeutic experiment ELIZA as early as in the 1960s (Shah et al., 2016). Since then, bots have been populating the web, often performing small functions to maintain online services and interaction on platforms (Geiger, 2014; Latzko-Toth, 2016). However, advances in natural language processing and machine learning over the last decade have enabled the development of bots capable of human-like interaction, usually referred to as chatbots or socialbots (Grimme et al., 2017). Newer versions of such bots can identify contexts of communication, modify their responses according to the interlocutor, and engage in human-like communication in ambiguous ways (e.g., Shah et al., 2016).

The communication and impact of bots have predominantly been studied in the context of social media (e.g., Ferrari et al., 2016; Grimme et al., 2017; Gorwa & Guilbeault, 2020; Neff & Nagy, 2016) and journalism (e.g., Bollmer & Rodley, 2016; Gómez-Zará & Diakopoulos, 2020). Recently, bots have also entered nonpublic arenas; for example, customer service and enterprise social media. Studies have explored the operation of bots within organizations as stimulants of interaction or highlighted the impact of intelligent technologies in various organizational processes (Araujo, 2018; Schanke et al., 2021; Stoeckli et al., 2018). In addition, organizational communication research has been interested in how communicative AI could function as part of a work team and influence team dynamics (Gibbs et al., 2021; Laitinen et al., 2021). Indeed, the ability to communicate with human members, in addition to more facilitative functions, is seen as an essential way for intelligent technologies to participate in teamwork (Seeber et al., 2020).

In communication research, interactions with bots and other machine entities have been explored under the umbrella of human-machine communication (Guzman & Lewis, 2020; Jones, 2014; Peter & Kühne, 2018). Traditionally, communication research has regarded technologies as platforms or venues that mediate interaction. In HMC, their role is understood more broadly: as an active participant in communication. Researchers have begun to examine how intelligent systems not just host or enable communication, but take part in and shape it (e.g., Edwards et al., 2019; Jones, 2014), perhaps resulting in a novel conceptualization of communication itself. Guzman and Lewis (2020) have coined the term *communicative AI* to refer to devices, applications, and algorithms capable of communicating in natural language and adapting to real-life conversational situations. They call for research to examine first, functional dimensions through which people conceptualize communicative AIs as communicators; second, relational dynamics of the human-technology relationship; and third, the metaphysical implications of the blurring ontological boundaries among human, machine, and communication.

Consequently, we highlight the importance of working openly to reform definitions and classifications rather than simply placing bots in existing boxes (Peter & Kühne, 2018), and studying how people understand and conceptualize the operation of these technologies in interactional situations (also Bollmer & Rodley, 2016). Communicative AIs not only mediate and facilitate communication but also automate and participate in it on their own terms (J. Reeves, 2016). Recent literature suggests that these kinds of technologies both facilitate connections and communication between people (Laitinen et al., 2021; Stoeckli et al., 2018) and communicate with individuals in a meaningful way (Ho et al., 2018). The notions of nonhuman communicative agents can also be transferred into interpersonal and team communication levels, where the complex, socially constructive management of meanings-centered viewpoints are challenged by the presence of AI. This leads to questions of how, or if, AIs can act as active subjects with whom people create meanings, rather than just as platforms for creating meanings between people (Guzman & Lewis, 2020; also Neff & Nagy, 2016).

The notion of technologies having social potential and agency can be traced back to the computers as social actors (CASA) paradigm, which aims to explain how humans interact with communicative technologies and how human perceptions shape the participation of the machines (Gambino et al., 2020; B. Reeves & Nass, 1998). Several perspectives have then built upon that base work to theorize the agency of a nonhuman communicator. Nass and Moon (2000) call it *mindlessness* when human users spontaneously and eagerly react to social cues and ignore the asocial ones. Often, the interplay of human agency and machinic agency are also related to the notion of control and establishing that control (Gibbs et al., 2021; Grimme et al., 2017). This highlights both the processes of designing and configuring these algorithm-based communicators—approachable by the concept of *affordance*, for instance—but also how social structures are constructed while communicating with bots (Gibbs et al., 2021).

Afforded Anthropomorphization

By studying the role of automated, communicative technology in interaction settings we build upon the technological affordance theory. This theory explains how technologies and their features enable different functions to their users (Gibson, 2015/1986; Hutchby, 2001; Stanfill, 2015). The concept of affordance emphasizes relativity: Technology does not determine user action, but, depending on the context of use and the user, frames the user's possibilities for action (Hutchby, 2001). A certain type of interface reinforces and promotes certain types of social activities and user experience, or might constrain and control it (Stanfill, 2015). Social media platforms typically afford activities such as posting, commenting, and liking, or in broader terms, they afford for increased visibility and persistence of communication (Treem & Leonardi, 2013).

We argue that the designed affordances of communication that emerge through anthropomorphization and characterization of bots are essential to understand how bots function as conversational agents. Anthropomorphization, as in assigning human attributes to nonhuman entities and objects, is *designed* in the sense that bots are scripted to behave in ways that would normally be attributed to humans only. For instance, a bot can announce that it is feeling sad. Such affordances, even in their simplest textual form, make the human participants feel more connected and sympathetic to the bot (e.g., Xu & Lombard, 2017). Characterization, on the other hand, refers to another human trait of interacting with inanimate objects that can be afforded by design—giving bots names, bodies, and "personalities" (Schanke et al., 2021). In many cases, it seems to be important that a bot has a distinctive character, or an embodied appearance (Araujo, 2018) even if the limits of characterization are usually quickly established (Eyssel & Kuchenbrandt, 2012). This design principle is supported by studies done in interspecies play, where anthropomorphization increases empathy and helps create a dialogical bridge in the human user's playful interaction with a nonhuman participant (Fava et al., 2019).

Across contexts, this logic relies on the basic human psychological tendency of anthropomorphization, seeing nonhuman things as human-like (Epley et al., 2007). Likewise, users typically rely on their experiences based on human interaction when trying to explain media technology (Edwards et al., 2019; B. Reeves & Nass, 1998). Human-like features built into technology, such as gender, sound, or appearance, cause users to perceive them even more as human-like rather than technological beings (e.g., Edwards et al., 2019). In robot development there seems to be a consensus that human features and the copying of human communication modalities significantly contribute to the acceptance of social robots (Blut et al., 2021; Epley et al., 2007). There is evidence that a socialbot featuring the same gender, ethnicity, and speech qualities as its user group is readily accepted as an in-group team member compared to a bot that is characterized differently (Eyssel & Kuchenbrandt, 2012). On the other hand, recent research has shown that the more acceptably human-like a robot is, the more it raises concerns about the power of technology in society (Ferrari et al., 2016; Männistö-Funk & Sihvonen, 2018).

Although bots are technological artifacts, there is evidence of their social, emotional, and relational impact and support when they engage in discussions with humans (Beattie & High, 2022; Ho et al., 2018; Laitinen et al., 2021). However, it seems that bots need some degree of human-like behavior and communication patterns for them to best act as communicative companions. For instance, the perceived *humanness* of these nonhuman actors has been found to be a predictor of motivation to engage in conversational journalism, or discussions with a chatbot altogether (Araujo, 2018; Shin, 2021). Socialbots must be at least somewhat human-like for them to be considered *social* (Grimme et al., 2017), but the complex mechanisms of achieving humanness of a bot in different contexts are still somewhat understudied. Therefore, there is a need to examine how humanness is manifested, constructed, and negotiated in human-machine communication. We do this by examining processes of communication between humans and bots in two different settings.

Data and Method

This study combines datasets from two previous studies (Laitinen et al., 2021; Pöyry & Laaksonen, 2022) to explore the significance of anthropomorphization and characterization of socialbots through naturally occurring communication where bots are addressed as interactive companions. Existing studies have shown how bots are designed to appear human-like and built with personalities and character (Araujo, 2018; Blut et al., 2021), but our interest moves further by adding the perspective of analyzing actual interaction between bots and humans in online environments. To make sense of the versatile roles of socialbots in different communicative contexts, our data covers both from a closed, team-level platform, and a public social media environment. These two contexts are referred to as *internal* and *public* social media.

Internal social media context. The organizational social media platform we study is Slack, a cloud-based online messaging and collaboration software used globally for team communication. It supports internal chat channels, private messaging as well as file sharing, and integration options with other services. Slack has a pre-programmed feature called the Slackbot, which is an automated socialbot present on all channels. The bot supports direct messaging for help and feedback, it can be customized to respond to certain words automatically, and it can be used to generate personal reminders and tasks. Some of these features are automated, some are suggested upon workspace setup, and some can be customized by the workspace admins. This study focuses on a Slack message database from a distributed team working in a Finnish media company. The data includes 45,940 messages in total, spanning over 2 years in time (August 2016–October 2018). Of these messages, 2,425 were sent by Slackbot. Bot messages were automated responses to certain trigger words configured by the human team members.

Public social media context. The social media dataset consists of public messages that interact with a corporate campaign that was built around a bot account. Fazer, a Finnish food and confectionary producer, launched the LovebotBlue campaign in 2018. The main feature of the campaign was a correspondingly named bot which communicated via a regular user account on the campaign platforms. According to the campaign material, a machine learning system was used to identify hate speech, and guided by a human moderator, the bot intervened in the identified discussions by making a remark of the conversation style. Dealing with a politicized topic, the campaign received negative feedback, much of which was targeted to the LovebotBlue (Pöyry & Laaksonen, 2022). We use a dataset of social media messages related to the campaign sent between October 1922–December 31, 2018. The data contains 1,615 tweets, Instagram posts, and forum messages mentioning the bot's username (n = 621) or the campaign hashtag (#smallpieceoflove).[1]

Data Analysis

Data were explored with a grounded, inductive approach with a focus on those conversation episodes where human users reply to messages sent by the bot or in other ways interact with it, that is, episodes of human-machine communication. Our aim was not to build a mere classification of messages but to explore the communication with and about the bot from a phenomenological perspective to build context-sensitive knowledge about the forms of bot-related communication in online environments.

The analysis provides a two-dimensional lens to the research questions. By using a qualitative approach, we examine both *the bot's communication style* and the ways in which *humans engage in discussion* with or about the bot. First, we explored how the designed and configured anthropomorphized nature of the bot manifests in the designed features of

1. #pienipalarakkautta in Finnish. The campaign hashtag and the name of the bot is related to one of the most popular products of Fazer, a milk chocolate bar called "Fazer's Blue." The word *piece* (pala) in the hashtag refers to both offering someone a piece of chocolate and a piece of *love* (instead of hate).

the bot as well as in the bot's messages. As we have established, existing studies show that human-like characteristics, such as gender, voice, or outlook, make the users consider bots more human-like than technological subjects. In the case of Slackbot and LovebotBlue, we were interested in how their appearance and actions invited human users for interaction. In addition, we approached the bots by exploring how they were *characterized* as communicative companions. Characterization, as previously mentioned, refers to the design principle that increases the acceptability of inanimate objects or technologies (Blut et al., 2021; Schanke et al., 2021). This analysis was executed in multiple rounds of inductive, data-driven analysis aimed at pinpointing and carefully illustrating the characterization visible in the bots' messages.

Second, we examined the messages in which human users actively engaged in discussion with or about the bot, that is, when they directly mentioned the bot handle or the word *bot*. The identification of these messages was conducted as follows: First, we automatically searched for mentions of the bot by name, nickname, or social media handle, including inflected forms of the word bot, Slackbot, and LovebotBlue. Second, we identified the response functions of team members' messages in the instances where human users engage in discussion with or about the bot. To make this distinction, we utilized a framework developed in a previous study (Laitinen et al., 2021), which uncovered that human members *respond to* (messages directed to the bot), *discuss about* (messages about the bot directed to humans), and *summon* (messages tagging or calling for the bot) socialbots in the context of internal social media. This framework was created through inductive analysis of the communicative *functions* present in bot-related communication. In this study, we began our analysis by coding the bot-related messages by the human users following this functional preset. For the Slack data, we used data previously classified by three of the authors (Laitinen et al., 2021). For the LovebotBlue data, the classification was done separately by one of the authors, who was also one of the three trained classifiers for the Slack data. Next, we engaged in qualitative analysis of the messages, one category at a time, to see if, and how, the anthropomorphization of the bot manifests in messages engaging the bot. The data was processed in spreadsheets and for each individual message we marked identified statements and verbal cues that suggested human-like features, thoughts, or emotions; for example, depicting the bot with action capabilities, feelings, or opinions and autonomy.

In the final phase, two authors worked together to sort the identified patterns of anthropomorphization into higher-level dimensions. Notes and findings were further discussed together by all authors following the practices of peer debriefing (Lincoln & Guba, 1985). We clustered the codes achieved in the previous phase and through the process of finding similarities and differences in the functions and contents of the statements, we eventually identified four categories which highlight the main ways of anthropomorphization in human-machine communication. This analysis led to the four dimensions in Table 1: *Direct address*, *Bot agency*, *Human-bot parallels*, and *Opinions and emotions*.

Findings

To make the bots approachable and interesting for human users to interact with, they are designed to appear appealing to us (Araujo, 2018), which is also evident in the visual, textual, and functional characteristics of Slackbot and LovebotBlue. A sympathetic human-faced

TABLE 1 Dimensions of Communicative Anthropomorphization of the Bot

	Bot Messages	Communicative Anthropomorphization
Direct address	Commands and suggestions (SlackBot, LovebotBlue) Questions (SB) Reprimanding humans (SB, LBB)	Direct answers to bot questions or suggestions (SB, LBB) Summoning the bot with mentions or trigger words (SB, LBB) Playing or trapping the bot (LBB) Abusing the bot (SB, LBB)
Bot agency	Indications of action (making coffee, being at the office) (SB) Evaluating human action (swearing, moderating) (SB, LBB)	Suggesting actions to the bot (SB, LBB) Evaluating bot action and skills (SB, LBB) Reporting messages to the bot (LBB)
Human-bot parallels	Indications of unity with *we, us, our* (SB, LBB) Talking about and following communication norms/ conventions (SB, LBB) Posting inside jokes (SB)	Addressing the bot as team member (SB) Addressing the bot as employee (SB, LBB) Implying that the bot has human-like features and abilities (SB, LBB)
Opinions and emotions	Phrases with emotional display (SB, LBB) Stating an opinion (SB, LBB) Talking about values/ideals (SB, LBB)	Referring to the bot's emotions (SB, LBB) Asking for the bot's opinion (SB, LBB)

figure has been designed for both, with a focus on the horizontal facial features: big eyes and a friendly smile. Slackbot's avatar is a box of four basic colors with eyes, a mouth, and a gentle appearance. LovebotBlue is depicted as a blue robot with typical humanoid features familiar from science fiction cartoons and comics: clear eyes and a smiling mouth. These are design features aimed at lowering the interactive threshold for social activity between the bot and the people facing it, that is, affordances configured to foster certain types of communication.

Both bots have a designed predisposition to communicate: they act in response to human messages and react to trigger words or recognize hate speech. Also, the affordances of social media platforms invite people to anthropomorphize bots. Since the bots occupy regular user accounts, they can be responded to and referenced in a conversation. Bots often appear as discrete persons by signing their own posts. For example, LovebotBlue talks in the first person, introduces itself, and linguistically emphasizes its own acting. It considers itself coming into the conversation as an outsider because of external *forces*, as can be seen in this example:

> Now stop. I am not actually involved in this debate, but I have to say that this style of discussion goes too far. Things as things and people as people, everyone has to be respected. #littlepieceoflove (Forum post)

In addition to design features and characterization, the bots are anthropomorphized through *configuration*. Through the affordances of the platform, human users seek to configure and modify the bot to appear as even more *human-passing*. Slackbot, in particular, was configured by its team members. Slackbot has built-in functions to support team interaction and work tasks, such as advising with links and giving reminders. It is also possible for users to configure bot-specific responses that the bot automatically triggers in response to specific words mentioned in the messages under the control of a randomized algorithm. Allowing such unsolicited and unexpected participation by a bot is a technological proposition that invites customization of the bot to be human-like. In the media organization we studied, Slackbot was configured to be more human-like immediately after the adoption of the platform by adding scripts such as greetings, rhetorical questions, and humorous utterances.

Both bots communicated in natural language with responses pre-programmed for them by the organization or human team members. Bot messages included several traits which further brought forward the human-like aspects of the nonhuman communicators. These aspects were not only related to the use of natural language, but to the topical, functional, and content-related characteristics of both bots' communication style. For example, LovebotBlue apologized for interrupting an ongoing conversation between human users and acknowledged its own position as an outsider. Slackbot responded in different ways to greetings, asked about the human users' well-being, told about its own "expenses" and reminded the others of making coffee. Slackbot was always around—although some mornings the bot announced that it would not be coming to the office, which is quite an analogy of human behavior. There is also something human in the ways the bots appear in discussions: they spot a keyword in the feed and respond.

User: Good morning!

Slackbot: Good morning, how are you?

User: Looks good, the sun is shining and soon on vacation.

As these interactions indicate, the human-like behavior of the bot is based on phatic communication that yields humorous and light-hearted results. The Slackbot, in particular, is designed and configured to act like a human member of the team: to socialize and interact in ways that make it appear as if it was *one of us*.

Communicative Anthropomorphization

Both of our datasets show how *humanness* (e.g., Shin, 2021) of the bot is manifested and constructed in human-machine communication when individuals interact with the nonhuman communicator. This was evident both in the human-like characterizations found in the messages of the bots themselves and in the ways the human users responded to, discussed with, and called for the bot. In this section we provide insights into the forms of communication that highlight the human-like features and abilities of the bots as people address bots *communicatively*. They are not considered human, but in some ways they are perceived as participants in the interaction, guided and afforded by their programmed human-like features (e.g., Bollmer & Rodley, 2016; Edwards et al., 2019). Although users understand that technology is designed and made by someone, they still target their message directly to the technological beings (Neff & Nagy, 2016). Thus, it is as if the technological being is bestowed with agency in interaction, in the ways of speaking to them. This idea is reflected in both of our datasets: the bots are repeatedly addressed like humans. We call these behaviors *communicative anthropomorphization*.

Direct Address

In both datasets we observed instances where human users directly address the bots: they respond to the bot by answering its questions or by sending comments or questions as replies to the bots' messages. There are moments when the bots were spoken to in a similar manner as to another human user—perhaps, however, with less empathy as the second example below shows:

> @LovebotBlue By eating dangerous sugar you hurt your own health. Did you know this @LovebotBlue? The worst products that endanger health are made by Fazer. Especially the chocolate department. Chocolate is eaten so much that insulin levels are through the roof. @FazerSuomi #avoidfazerproducts (Twitter)

> User: Did you receive any feedback?
>
> Slackbot: For the sake of reader feedback, that is why these things are done—and for the [Journalism Prize] jury
>
> User: Be quiet bot

Another form of direct address we identified is that humans try to engage in discussions with the bots and summon them on purpose by tagging them or using known trigger words. This is a playful and inherently communicative activity. In the LovebotBlue data, these forms of addressing call the bot to participate in a discussion; for example, with the intention of reporting a hateful message to the bot or asking for reactions. Addressing the bot with trigger words could also be interpreted as a way of wanting to understand how the technological entity works:

> @user @Lovebotblue @user Seems to react to certain words. Let's test it. [Lists six offensive, immigration-related terms.] (Twitter)
>
> Slackbot: Go Hank!
>
> User 1: I wonder if tea is more to Hank's liking.
>
> Slackbot: Go Hank!
>
> User 2: coffee hank
>
> Slackbot: I would listen to what Hank has to say about this
>
> User 1: Why doesn't this guy [Slackbot] speak about coffee anymore?
>
> User 2: Coffee, do you have something to say about that, Slackbot?

One prominent feature of direct address is negative commentary targeted to the bot. For example, in the LovebotBlue data the human users frequently directly address the bot as if it were a user with intentions and opinions; complaining about its actions, asking for justifications for something it said, even directly abusing the bot. Similarly, the Slackbot receives instances of rather blunt, offensive, and aggressive responses. This added offensiveness is perhaps a sign of people regarding the bot as a machine that can be abused without moral considerations (see Epley et al., 2007). The offensive messages are formulated in a way that they directly talk to the bot, thus engaging with the bot as if it was a conversation partner:

> @user @LovebotBlue @user Lovebot: Firstly, you have a face on your head and secondly, it's not very pretty. Do you want me to jpconfig your face, huh?! Would you like it if we all monitored each other like this? Where do you think that would lead except to a candy company being a moral guardian or internetbot trolling? (Twitter)

User:	Good morning
Slackbot:	Ouch, what day is it?
User:	How should I know @Slackbot

Bot Agency

As described above, one way in which the bot is treated human-like is endowing it with agency through communication. Such action is partly triggered by the messages configured to the bots. What is interesting, however, is that such linguistic notions on agency are even more strongly present in the messages that respond to or mention the bots. First, a great share of bot-related messages in both datasets focus on commenting and evaluation of the bots' actions. These messages express, for example, how the "bot censors," "bot knows Finnish," "bot did not do anything wrong," "slackbot is messing around," "bot judges," "bot shared wrong information," "slackbot talks." The focus in these messages is action-oriented in the sense that they evaluate the actions of the bots, and do it by linguistically positioning the bot as an actor:

User 1:	Good morning!
Slackbot:	Good morning to you!
User 1:	Once again bot, you haven't made coffee.
Slackbot:	Somebody make coffee!
User 1:	Your turn.

Second, human discussants in both datasets are making suggestions for action to the bots. In the LovebotBlue data there are several messages calling for autonomous bot intervention when users report messages to the bot. The main goal of this activity is to receive a judgment from the bot regarding the potentially hateful content of the message, and perhaps also to test the limits of the bot as well as its just action across the political spectrum.

> @LovebotBlue Could you check the texts in this picture where you can find a "researcher of facism" supported by Yle [Finnish national public broadcaster] @user (Twitter)

> @user you'd think a grown up would know how to behave.. and why didn't @LovebotBlue intervene? 🤔 (Twitter)

Human-Bot Parallels

The anthropomorphization of the bots was constructed in interaction where the bots were portrayed as parallel with human communicators. This was manifested by messages commenting on or reacting to the bots' actions and characteristics by positioning the *bot as a part of the team*, referring to the *bot as an employee*, and implying that the *bots have human-like features and abilities*. The messages positioning Slackbot as a part of the team included statements where human members regarded the bot as "theirs," or in other ways being part of the team, or "us." Such team-talk was presented also through instances where the bot encouraged the team members, made a comment or suggestion related to work, or participated in the inside jokes of the team—followed by team members' reactions by discussing the bot's behavior or responding to it. The following excerpts illustrate how Slackbot is treated as being part of the team:

Slackbot:	We are going to win the award
User 1:	Quite talkative, this Slackbot of ours. Perhaps it could do articles for us too?
User 2:	I shall suggest that to our editors, they will run with it
User:	Good idea Patrick, do you have time to finish it?
Slackbot:	Yes, but who has the time to do it?
User:	Patrick, slackbot.

The bots, especially LovebotBlue, were sometimes regarded almost like employees or advocates of the company they represented. This manifested by both giving the bot customer service requests, and contacting and challenging the company through the bot. This behavior could also be interpreted as a reflection of the most common function chatbots and socialbots are designed to perform as the first layer of online customer service. The following examples illustrate how people communicate with the bot by asking questions related to the operational field of the company and their ethical actions. Visible in these excerpts is also the way in which humans challenge the organization by directing their views on societal issues and the organization's role in them to the bot, much like an employee represents an organization.

> Ping @LovebotBlue ! Answer this NPC [non-player-character]! Why was Fazer racist? (Twitter)

> Hi @LovebotBlue! Fazer's Blue as a brand is of course the most Finnish brand ever. Btw, what kind of milk is used to make it? #fazer (Twitter)

The team members and social media audience also posted messages that mentioned the bots' human-like features and even directly compared them to human abilities and features. For instance, the bots were compared to humans as conversation partners, they were implied to have a "holiday," and they were described as cute or having performed well. Sometimes the Slackbot was even directly compared to a team member. The following excerpts highlight these instances:

User 1: Someone to talk to for the lonely.

User 2: Yeah, slackbot talks, if no one else is here. - -

User 3: In the future, I'll spend my weekends talking to the slackbot!

User 4: Better company, at last.

@user @user Can't see @LovebotBlue intervening much.. Oh, the [machine nickname] is on Christmas holiday. (Twitter)

Opinions and Emotions

Finally, we saw communicative anthropomorphization of the bots by providing them indications of human-like characteristics, such as *assumptions of emotion* as well as *asking for opinions*. The bots' own messages included various forms of emotional expression as a form of human-like language use, and occasionally the human users engaged with the display of emotion by reacting to it or talking about the bots' feelings. The bots were described to be "optimistic" or "passive-aggressive." The humans also occasionally felt the need to apologize as if the bot's feelings were hurt. These kinds of statements highlight anthropomorphization as emotions inherently bring forth human-likeness. The excerpts below illustrate emotion-related messages:

User: Sounds like a bloody interesting news poet

Slackbot: That is such foul language!

User: Sorry

@user @LovebotBlue seems to be one of those passive aggressive bots (Twitter)

In addition to messages that regarded the bots' emotions, human communicators inquired the bots' opinions as if the bot would have its own perceptions of the world and could develop its own viewpoints. Furthermore, the human communicators did not only ask for the bots' opinions on matters but also occasionally confirmed and praised or belittled their thoughts on issues. Interestingly, LovebotBlue acts as a conversationalist in a completely different way than the organization's own accounts on social media. Sometimes asking opinions was a form of challenging the organization through the bot, as campaign critics used it as an object to approach Fazer and highlight broader political themes. The following instances highlight such interactions:

User 1: I'm trying to find synonyms for poop. That is what I am doing at the moment. [continues to list said synonyms]...hit me!

Slackbot: Shittier ideas should be saved for the parent company's other newspapers.

User 2: Slackbot is absolutely right

@LovebotBlue @FazerFinland what do you think about the situation of freedom of expression and human rights in Russia, China and Turkey? Is it appropriate to trade with China while burying the human rights situation? #hate speech #human rights #word responsibility (Twitter)

Discussion and Conclusion

In this study, we examined the communicative anthropomorphization of two socialbots, Slackbot and LovebotBlue, in organizational and public social media contexts. Our empirical analysis showed how humanness of the bots was not only a design element or a psychological, intrapersonal process, but also socially constructed in human-machine communication. Our study contributes to the existing, predominantly experimental research on anthropomorphization (e.g., Araujo, 2018; Blut et al., 2021; Schanke et al., 2021) by exploring this process through two datasets of naturally occurring conversations, and by highlighting the social, collective, and performative aspects of anthropomorphization. Further, we contribute to the field of human-machine communication (Guzman & Lewis, 2020) by proposing that the ways in which human users communicate with bots are an essential mechanism for making the machines seem and feel more human. We refer to this as the process of *communicative anthropomorphization*.

Previous studies have shown that human-like features are essential cues for users to perceive technological interlocutors as social companions and to activate the psychological inference of anthropomorphism (e.g., Edwards et al., 2019; Epley et al., 2007; Wischnewski et al., 2022). We add to the existing discussion on the design of socialbots (e.g., Araujo, 2018; Shah et al., 2016) by emphasizing the aspects of configuration and communication. Both studied bots were configured to be even more human by the human users: for example, by adding human-like responses for the bots as if they were real users with intentions and opinions. Moreover, our results accentuate the communicative, socially constructed anthropomorphization of bots: they are endowed with agency through the communication by humans who interact with them. In our data, bots were directly addressed, interrogated, and paralleled with humans. Through their interventions, bots also changed the course of the conversation, elicited feelings, and generated action. They seem to act as links between the human world of interaction and the technological world as they communicate with an automated logic but cause repercussions in human communication. In this sense, bots have agency that extends beyond the traditional mediator role considered for technology; they function as triggers for communication in unpredictable ways (Guzman & Lewis, 2020).

While anthropomorphism is a known tendency of humans (Heider & Simmel, 1944), the anthropomorphization of bots is further motivated by their design, the technological context, and by the efficacy expectations present in the social situation (Epley et al., 2007). Indeed, bots do not exist or function without their technological and social context.

Considering the affordance theory and the broader literature around social construction of technology (Pinch & Bijker, 1984), we suggest that social bots are technologies with anthropomorphized characteristics and affordances, which trigger the psychological process of anthropomorphization, but the process is completed and the bots are *realized as agents* by humans who attribute them with agency through communication. Although all technology is often talked about when used (Laitinen & Valo, 2018), the bot stands out because it is not only the subject but also the object of talk: users in our empirical data talk *about, with,* and *to* the bot, regardless of the human users knowing the bot is artificial. Thus, the affordances of the bot, in particular its capabilities to communicate in human language, invite users to treat and tease it as a human-like yet artificial actor, and to generate forms of co-constituted, symbiotic, communicative agency (see Neff & Nagy, 2016).

Communicating with bots, however, is interaction marked with disappointment: humans in our datasets try to converse with the bots but encounter the limited abilities of their machine interactants, as the bots fail to follow the shared conventions of human conversation. This seems to cause emotional distress and abuse targeted toward the bots. Therefore, the bots are still frequently regarded and treated as the technological other. Because the bot acts wrong in the process of communication, it fails to achieve the role of a plenipotentiary interaction agent (Bollmer & Rodley, 2016). Bots might be designed human-like, they are configured even more so, and addressed by humans as communicators, but until their communication capabilities are more sophisticated, something is still missing. It is as if the bots are expected to communicate without errors because of their technological nature but still, paradoxically, making mistakes returns them to their technological status (cf. Guzman & Lewis, 2020). This further highlights the social construction taking place in the communication process between humans and machines: communication is not reduced to transfer of information but rather, meanings are created and negotiated *despite* the bots' limited ability to interact.

Furthermore, our data indicates that communication with bots is shaped by the context and the platform. While there are similarities in communication styles toward the bots across platforms as described above, differences are notable: on public social media, LovebotBlue was repeatedly abused, told to get off the platform, and its (or its owner company's) motivations were questioned. The Slackbot, on the other hand, was often completely ignored as it responded to keywords incorrectly. Slackbot is, yet, constantly performing and constructing the team itself by repeating team-configured inside jokes and dramatizations that are meaningful to the team members. Being configured by the team members using it daily, Slackbot gets treated in a more inclusive manner, while LovebotBlue is seen as an extension of Fazer and is treated accordingly. Further, while both bots are designed to intervene in human discussions, LovebotBlue enters them more uninvited and perhaps therefore elicits more rude reactions. The broader context, thus, affects the negotiations of control when communicating with artificial actors (Gibbs et al., 2021; Grimme et al., 2017). Hence, our findings highlight the importance of investigating anthropomorphization in varying social contexts, beyond intra- or interpersonal settings.

In conclusion, we suggest that the communicative anthropomorphization of bots is an important aspect of their functionality and their construction as agents in social, interactive situations. Thus, we propose that anthropomorphization is simultaneously a design process, a psychological process, and also a communicative process of socially and collectively constructing human-likeness through interaction. In this vein, our results highlight the call presented in HMC that the emergence of digital interlocutors generates a need to redefine the existing conceptualizations of communication, interaction, and agency in the context of communicative AIs. Future research should further explore the communicative anthropomorphization of socialbots by examining it across contexts: bots on different platforms and in different social settings work in varied ways and have diverse implications.

Author Biographies

Salla-Maaria Laaksonen (DSocSc, Docent) is a senior researcher at the Centre for Consumer Society Research, University of Helsinki. Her research areas are technology, organizations, and new media, including social evaluation of organizations in the hybrid media system, the organization of online social movements, and the use of data and algorithms in organizations. She is also an expert in digital and computational research methods.

https://orcid.org/0000-0003-3532-2387

Kaisa Laitinen (PhD) is a senior lecturer of communication at the Department of Language and Communication Studies, University of Jyväskylä. Her research interests entail technology-mediated interpersonal and organizational communication, virtual teamwork, as well as socialbots, AI, and other communicative, intelligent technologies in the workplace. She has extensive experience with qualitative methods and utilization of naturally occurring communication data.

https://orcid.org/0000-0001-8357-3020

Minna Koivula (MA) is a PhD candidate in journalism at the Department of Language and Communication Studies, University of Jyväskylä. Her research examines journalism through the lens of organizational communication research and explores the shifting institutional context of work in the media field.

https://orcid.org/0000-0002-7718-0199

Tanja Sihvonen (PhD) is professor of Communication Studies at the University of Vaasa. She is specialized in digital media, computer games, and participatory cultures on the internet. Her most recent work considers role-playing games, blockchain technologies, and astroturfing on social media.

https://orcid.org/0000-0003-3220-2814

Acknowledgments

This work has been partly funded by the Media Industry Research Foundation of Finland (project Somedia). The authors would like to thank the three anonymous reviewers for their insightful and encouraging comments, and our informants in the organization that provided us access to their Slack data.

References

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, *85*, 183–189. https://doi.org/10.1016/j.chb.2018.03.051

Beattie, A., & High, A. (2022). I get by with a little help from my bots: Implications of machine agents in the context of social support. *Human-Machine Communication, 4,* 151–168. https://doi.org/10.30658/hmc.4.8

Blut, M., Wang, C., Wünderlich, N. V., & Brock, C. (2021). Understanding anthropomorphism in service provision: A meta-analysis of physical robots, chatbots, and other AI. *Journal of the Academy of Marketing Science*. https://doi.org/10.1007/s11747-020-00762-y

Bollmer, G., & Rodley, C. (2016). Speculations on the sociality of socialbots. In R. Gehl & M. Bakardijeva (Eds.), *Socialbots and their friends* (pp. 147–163). Routledge.

Edwards, A., Edwards, C., Westerman, D., & Spence, P. R. (2019). Initial expectations, interactions, and beyond with social robots. *Computers in Human Behavior 90,* 308–314. https://doi.org/10.1016/j.chb.2018.08.042

Epley, N., Waytz, A., & Cacioppo, J. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., & Kuchenbrandt, D. (2012). Social categorization of social robots: Anthropomorphism as a function of robot group membership. *British Journal of Social Psychology 51*, 724–731. https://doi.org/10.1111/j.2044-8309.2011.02082.x

Fava, F., Soares, C. M., & Carvalhais, M. (2019). Playful design, empathy and the nonhuman turn. *Technoetic Arts 17*(1–2), 141–154. https://doi.org/10.1386/tear_00012_1

Ferrari, F., Paladino, M., & Jetten, J. (2016). Blurring human–machine distinctions: Anthropomorphic appearance in social robots as a threat to human distinctiveness. *International Journal of Social Robotics, 8*(2), 287–302. https://doi.org/10.1007/s12369-016-0338-y

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. https://doi.org/10.30658/hmc.1.5

Geiger, S. R. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society, 17*(3), 342–356. https://doi.org/10.1080/1369118X.2013.873069

Gibbs, J., Kirkwood, G.., Fang, C., & Wilkenfeld, J. (2021). Negotiating agency and control: Theorizing human-machine communication from a structurational perspective. *Human-Machine Communicatio*n, 2, 153–171. https://doi.org/10.30658/hmc.2.8

Gibson, J. J. (2015/1986). *The ecological approach to visual perception.* Taylor & Francis.

Gómez-Zará, D., & Diakopoulos, N. (2020). Characterizing communication patterns between audiences and newsbots. *Digital Journalism*, *8*(9), 1093–1113. https://doi.org/10.1080/21670811.2020.1816485

Gorwa, R., & Guilbeault, D. (2020). Unpacking the social media bot: A typology to guide research and policy. *Policy and Internet, 12*(2), 225–248. https://doi.org/10.1002/poi3.184

Greenwood, D., & Long, C. (2011). Attachment, belongingness needs, and relationship status predict imagined intimacy with media figures. *Communication Research*, *38*(2), 278–297. https://doi.org/10.1177/0093650210362687

Grimme, C., Preuss, M., Adam, L., & Trautmann, H. (2017). Social bots: Human-like by means of human control? *Big Data*, *5*(4), 279–293. http://doi.org/10.1089/big.2017.0044

Guzman, A. L., & Lewis, S. C. (2020). Artificial intelligence and communication: A human–machine communication research agenda. *New Media and Society, 22*(1), 70–86. https://doi.org/10.1177/1461444819858691

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243–259.

Ho, A., Hancock J., & Miner A. (2018). Psychological, relational, and emotional effects of self-disclosure after conversations with a chatbot. *Journal of Communication, 68*(4), 712–733. https://doi.org/10.1093/joc/jqy026

Hutchby, I. (2001). Technologies, texts and affordances. *Sociology, 35*(2), 441– 456. https://doi.org/10.1017/S0038038501000219

Jones, S. (2014). People, things, memory and human-machine communication. *International Journal of Media & Cultural Politics*, *10*(3), 245–258. https://doi.org/10.1386/macp.10.3.245

Laitinen, K., Laaksonen, S.-M., & Koivula, M. (2021). Slacking with the bot: Programmable social bot in virtual team interaction. *Journal of Computer-Mediated Communication*, *26*(6), 343–361. https://doi.org/10.1093/jcmc/zmab012

Laitinen, K., & Valo, M. (2018). Meanings of communication technology in virtual team meetings: Framing technology-related interaction. *International Journal of Human-Computer Studies*, *111*, 12–22. https://doi.org/10.1016/j.ijhcs.2017.10.012

Latzko-Toth, G. (2016). The socialization of early internet bots: IRC and the ecology of human-robot interactions online. In R. Gehl & M. Bakardijeva (Eds.), *Socialbots and their friends* (pp. 63–84). Routledge.

Lincoln, Y., & Guba, E. (1985). Establishing trustworthiness. *Naturalistic Inquiry, 289*(331), 289–327.

Männistö-Funk, T., & Sihvonen, T. (2018). Voices from the uncanny valley. *Digital Culture & Society, 4*(1), 45–64. https://doi.org/10.14361/dcs-2018-0105

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

Neff, G., & Nagy, P. (2016). Talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication*, *10*, 4915–4931.

Peter, J., & Kühne, R. (2018). The new frontier in communication research: Why we should study social robots. *Media and Communication* *6*(3), 73–76. https://doi.org/10.17645/mac.v6i3.1596

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social studies of science, 14*(3), 399–441. https://doi.org/10.1177/030631284014003004

Pöyry, E., & Laaksonen, S.-M. (2022). Opposing brand activism: Triggers and strategies of consumers' antibrand actions. *European Journal of Marketing*, *56*(13), 261–284. https://doi.org/10.1108/EJM-12-2020-0901

Reeves, B., & Nass C. (1998). *The media equation*. CSLI Publications.

Reeves, J. (2016). Automatic for the people: The automation of communicative labor. *Communication and Critical/Cultural Studies 13*(2), 150–165.

Schanke, S., Burtch, G., & Ray, G. (2021). Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, *32*(3), 736–751. https://doi.org/10.1287/isre.2021.1015

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., Randrup, N., Schwabe, G., & Söllner, M. (2020). Machines as teammates: A research agenda on AI in team collaboration. *Information & Management, 57*(2), 103174. https://doi.org/10.1016/j.im.2019.103174

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior, 58*, 278–295. https://doi.org/10.1016/j.chb.2016.01.004

Shin, D. (2021). The perception of humanness in conversational journalism: An algorithmic information-processing perspective. *New Media & Society, 24*(12), 2680–2704. https://doi.org/10.1177/1461444821993801

Stanfill, M. (2015). The interface as discourse: The production of norms through web design. *New Media & Society, 17*(7), 1059–1074. https://doi.org/10.1177/1461444814520873

Stoeckli, E., Uebernickel, F., & Brenner, W. (2018). Exploring affordances of slack integrations and their actualization within enterprises. In *Proceedings of the 51st Hawaii International Conference on System Sciences*. https://doi.org/10.24251/hicss.2018.255

Treem, J., & Leonardi, P. (2013). Social media use in organizations: Exploring the affordances of visibility, editability, persistence, and association. *Annals of the International Communication Association, 36*(1), 143–189. https://doi.org/10.1080/23808985.2013.11679130

Wischnewski, M., Ngo, T., Bernemann, R., Jansen, M., & Krämer, N. (2022). "I agree with you, bot!" How users (dis)engage with social bots on Twitter. *New Media & Society*, https://doi.org/10.1177/14614448211072307

Xu, K., & Lombard, M. (2017). Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior, 74*, 152–162. https://doi.org/10.1016/j.chb.2017.04.043

HUMAN-MACHINE COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.9

Valenced Media Effects on Robot-Related Attitudes and Mental Models: A Parasocial Contact Approach

Jan-Philipp Stein[1] and Jaime Banks[2]

1 Department of Media Psychology, Institute for Media Research, Chemnitz University of Technology, Chemnitz, Germany
2 School of Information Studies, Syracuse University, Syracuse, NY, USA

Abstract

Despite rapid advancements in robotics, most people still only come into contact with robots via mass media. Consequently, robot-related attitudes are often discussed as the result of habituation and cultivation processes, as they unfold during repeated media exposure. In this paper, we introduce parasocial contact theory to this line of research—arguing that it better acknowledges interpersonal and intergroup dynamics found in modern human–robot interactions. Moreover, conceptualizing mediated robot encounters as parasocial contact integrates both qualitative and quantitative aspects into one comprehensive approach. A multi-method experiment offers empirical support for our arguments: Although many elements of participants' beliefs and attitudes persisted through media exposures, valenced parasocial contact resulted in small but meaningful changes to mental models and desired social distance for humanoid robots.

Keywords: parasocial contact, social robots, mental models, social distance, media effects

Introduction

Social robots—(semi-)autonomous machines with the ability to simulate human sociality—are increasingly entering human social spheres. Contemporary innovators envision such machines in an ever-growing number of roles and positions, from robotic health

CONTACT Jan-Philipp Stein • jpstein@phil.tu-chemnitz.de • Department of Media Psychology, Chemnitz University of Technology, Thüringer Weg 11 • Germany D-09126

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

care providers, teaching assistants, and coworkers in assembly lines, to friendly hotel concierges handing over room keys. Despite these many potential integrations, however, most people still only encounter social robots through media representations; for example, as part of television shows, documentaries, or movies (Mara et al., 2021; van Oers & Wesselmann, 2016). In turn, scholars have suggested that the future adoption of robots may critically depend on how media portrayals shape attitudes and impressions prior to actual adoption opportunities (e.g., Banks, 2020; Savela et al., 2021).

Importantly, the conditions and processes that give rise to media-facilitated impression formation for robots—especially those of a humanlike design—are only vaguely understood. While some exploratory studies have indicated that different types of robot depictions in the media may shape viewer attitudes accordingly, their strictly empirical approach provided only few theoretical reference points to make sense of the examined effects (e.g., Bruckenberger et al., 2013). In response to this shortcoming, more recent literature has discussed robot-related media influence through the lenses of habituation and cultivation (i.e., as the stepwise adjustment of people's mental models according to repeated mass-mediated encounters; e.g., Banks, 2020; Sundar et al., 2016; Young & Carpenter, 2018). As such, scientific focus has rested mainly on the *quantity* of robot representations in media, but not yet on their *quality*. Although understanding exposure quantity is useful, it is incomplete and must be accompanied by unpacking qualitative aspects as well. In the current study, we begin to address that gap by building on the *parasocial contact hypothesis* (Schiappa et al., 2005)—the idea that biases toward dissimilar others can be alleviated by positive, counter-stereotypical media exemplars. Using scenes from famous movies and television shows to create experimental conditions that represent positive vs. negative parasocial contact with humanoid social robots, we investigate changes in participants' mental models as well as their subsequent behavior toward a real-life robotic machine. We employ a mixed-method approach combining an inductive exploration of people's before- and after-contact mental models with deductive testing of whether parasocial dynamics transfer to robots. In doing so, we find that even limited parasocial exposures can have small but meaningful changes to how one thinks and feels about robots that look human and/or behave in a human-like way.

Interpersonal and Intergroup Dynamics in Human-Robot Interaction

People often perceive and react to social robots as if they were human (e.g., Spatola et al., 2019; van Straten et al., 2020), but robots are also seen as a distinct *kind* (Banks & Koban, 2022; Kahn et al., 2011). Consequently, engaging these robots may no longer be a question of mere technology acceptance but rather the result of complex interpersonal and intergroup processes.

Interpersonal processes encompass cognitions, emotions, and behaviors that occur face-to-face, including impression formation, stereotyping, or relationship development. Although caution has been urged against overgeneralizing *all* interpersonal theories as transferrable to human–machine communication (Fox & Gambino, 2021), research suggests that parallels are frequent, especially once robotic machines look or behave distinctly human-like (e.g., Lee et al., 2006; Stein et al., 2022). While the mechanisms underlying these parallels are not yet well-understood, one potential explanation lies in people's automatic social-cognitive processes. In particular, both humans and robots seem to evoke similar mentalizing processes in observers—that is, people may automatically infer the mental states of both types of entities and use those inferences to interpret behaviors (e.g., Airenti, 2015; Banks, 2021). Accordingly, users may develop genuine empathy and emotional attachment toward robotic machines, which further prompts them to treat the machines as social actors.

Secondly, interactions with robots may parallel those with humans as they aggregate, identify, and differentiate among one another (i.e., as they follow conventional principles of *intergroup behavior*). In this domain, *ingroups* are defined as social groups with whom one identifies (e.g., peer group, family, community), whereas *outgroups* are all other social groups that do not elicit such identification. Because people perceive robots as social entities yet also as ontologically different from themselves (Kahn et al., 2011), they are likely to be categorized as a distinct social group (e.g., Smith et al., 2021). In turn, intergroup dynamics may come into effect (cf. Tajfel et al., 1979): Whereas the human ingroup is typically perceived in a favorable light, the robotic outgroup may be met with apprehensiveness and devaluation (e.g., Vanman & Kappas, 2019). Indeed, these ingroup–outgroup biases seem to be particularly evident once people encounter highly homogenous robot groups (Fraune et al., 2017) or expect available resources to be limited (Jackson et al., 2020)—as these conditions heighten perceptions of self-dissimilarity and competition. In a similar vein, Gamez-Djokic and Waytz (2020) connected concerns about robotic automation to both realistic and symbolic outgroup threats, including the loss of jobs and dominant cultural values. This further illustrates that, regardless of robots' increasing sophistication and usefulness, people might ultimately remain wary of the robotic *other*.

Intergroup Contact as a Way to Mitigate Outgroup Bias

For developers, marketers, and researchers of robotic technology, such intergroup dynamics raise a crucial question: How do outgroup biases toward robots impact human–robot interactions? On the one hand, given automation's potential to enhance human life, minimizing outgrouping and fostering ingrouping could promote social and functional acceptance (e.g., collaboration or social harmony). On the other, some have argued that humans should limit their anthropomorphization of robots and keep robotic simulations of sociality from tapping into preconscious drivers of actual sociality (e.g., Bryson, 2010). From both perspectives, it is critical to understand group-relevant biases—whether to support or suppress social integration.

We focus here on relevant theory that may help to explain dynamics of robot social acceptance despite their outgroup status, with particular inspiration taken from social psychological literature. The *contact hypothesis* (Allport, 1954) proposes that intergroup relations can be improved through guided facilitation of positive outgroup contact, depending on several relational and contextual factors. For example, contact between two groups may be particularly effective at reducing bias if both parties are of equal status, strive for a common goal, and are guided by positive norms (Allport, 1954; Pettigrew & Tropp, 2006). Moreover, the presentation of counter-stereotypical characteristics is said to be particularly beneficial in terms of contact effects—it prompts observers to dismiss (biased) group-level perceptions in favor of more individualized judgments (Taschler & West, 2016). At the same time, a *negative contact hypothesis* must be considered (Meleady & Forder, 2018): Unpleasant or stereotype-confirming interactions can instead lead to stronger prejudice and aversion. Apart from this limitation, however, empirical evidence anchors contact dynamics as a highly effective means to improve social-group relations (e.g., Pettigrew & Tropp, 2006).

Inspired by these notions, HRI scholars have started to wonder if intergroup contact may similarly reduce bias toward robots as an outgroup. Their work showed that neutral in-person encounters with a robot significantly reduced the psychological distance participants felt toward social robots as an ostensible outgroup (Haggadone et al., 2021), in line with prior work demonstrating that evaluations of robots improve after repeated in-person interactions (e.g., Haring et al., 2015). Notably, however, past work has largely framed such observations as the result of *habituation* (i.e., as a less aversive response following uncertainty reduction; e.g., Koay et al., 2007). Although such desensitization effects are also incorporated in intergroup contact theory (e.g., Pettigrew & Tropp, 2006), the contact hypothesis reaches notably further: It assumes that face-to-face contact not only breaks down negative expectations, but also helps to replace stereotypical cognitions with more individualized or even counter-stereotypical perceptions (Allport, 1954). In this sense, contact between social groups may ultimately serve to correct "hasty generalization[s] made about a group based on incomplete or mistaken information" (Schiappa et al., 2005, p. 93).

From Direct to Parasocial Contact

A modification of Allport's (1954) original conception, the *parasocial contact hypothesis* (PCH), presumes that intergroup contact does not necessarily have to be synchronous and co-present in order to elicit bias reduction (Banas et al., 2020; Schiappa et al., 2005). Instead, mass-mediated contact with a depicted outgroup (e.g., watching minority group portrayals on television) could also exert a meaningful positive influence on people's attitudes—an effect grounded in the notion of *parasocial interactions* (PSIs; Horton & Wohl, 1956).

PSIs were initially understood as a form of perceptual "illusion" (Horton & Wohl, 1956, p. 215) occurring during television consumption: Despite exposure to televised characters being operationally one-sided (i.e., the character speaks to the audience and is heard, but communication cannot be reciprocated), viewers may perceive it to be reciprocal—and even react accordingly (e.g., by talking back to the character). From this initial conceptualization, the construct was later complemented by the notion of *parasocial relationships* (PSRs; i.e., overarching feelings of relatedness that emerge across multiple interactions). Taken together, both parasocial phenomena are now commonly understood as a complex set of cognitive, affective, and behavioral responses during and after media reception, by which a nondialectical, imaginary connection feels dialectical and quite real (e.g., Liebers & Schramm, 2019). Moreover, parasocial experiences tend to resemble everyday social ties in profound ways, for instance offering similar gratifications and triggering similar social judgments (e.g., Tukachinsky & Stever, 2019). Thus, the effects of parasocial contact may mirror those of traditional face-to-face contact as both are based on the *perception* of meaningful interpersonal connections.

Highlighting the validity of the PCH, a recent meta-analysis (Banas et al., 2020) synthesized 79 studies on parasocial contact, reporting a notable decrease in various outgroup biases following exposure to positive group depictions ($r = -.23$). A reverse effect was also found, as negatively valenced outgroup portrayals led to worse attitudes among participants ($r = .31$). The meta-analysis further revealed that there was no significant difference between mediated and vicarious contact (i.e., passively observing group interactions in real life), underscoring the vivid nature of encountering outgroups via media. Importantly, this equivalence of contact modalities was also observed for human–robot interactions: In a recent field experiment, evaluations of a social robot were not significantly different when encountering it in person or via 2D or 3D screens (Mara et al., 2021; cf. Li, 2015).

Mediated Robot Encounters as Parasocial Contact

Given initial evidence that intergroup dynamics may extend to robots as an ostensible outgroup, a vital next step for human-machine communication theory is to scrutinize mass-mediated robot exposure as part of the PCH framework. We argue for this framing because PCH accounts for two important limitations of past approaches in ways that still allow for the synthesis of extant findings (e.g., Banks, 2020; Bruckenberger et al., 2013; Savela et al., 2021; Sundar et al., 2016; Young & Carpenter, 2018).

First and foremost, prior approaches rely most heavily on notions of habituation (as detailed above) and on cultivation theory (e.g., Banks, 2020; Sundar et al., 2016)—the idea that repeated mass media exposure shapes viewers' mental models according to often similar, stereotypical group representations (Gerbner & Gross, 1976). Importantly, habituation and cultivation can be applied effectively to *any* focal object or phenomenon (e.g., cultivated understanding of crime or education or even rocks); which means that neither approach accounts for the [simulation of] sociality inherent to human-machine communication. Parasocial contact theory specifically considers the dynamics of social ties, including processes by which trust, liking, and attraction emerge. It further encompasses vicarious learning (Bandura, 2009), another socially informed mechanism yet unaddressed through the habituation or cultivation approach. Crucially, we underscore that the PCH does not *preclude* processes inherent to those perspectives—instead, it offers a more comprehensive framework for integrating those perspectives with person perception and intergroup dynamics.

Secondly, the PCH framework covers both quantity and quality of exposure, building on a large body of evidence regarding beneficial and detrimental contact conditions (Allport, 1954; Banas et al., 2020; Żerebecki et al., 2021). In turn, this further allows it to offer clear suggestions as to how mediated group portrayals may evoke positive or negative effects. Particularly, it reframes exposure to media representations as one that is experienced as *actual and social*, so that phenomenological processes inherent to interpersonal and intergroup dynamics become the focal mechanisms. Moreover, even though research suggests that parasocial contact may profit from repetition and prolonged duration (Żerebecki et al., 2021), its benefits can even unfold after single and brief interactions (e.g., Schiappa et al., 2005). As such, the PCH appears to be particularly well-suited to inform empirical efforts applying both time-zero and longitudinal methodologies.

The Current Study

At this point, the open question is: (How) do qualitative properties of robot depictions in media causally impact people's understandings of and attitudes toward members of that group? We address the question of *understanding* through the lens of mental models (MMs)—cognitive structures resulting from the internalization of external phenomena, which serve as frames for interpreting immediate experience (Craik, 1943). MMs contain tokens of knowledge representing things abstract or concrete, more or less like the actual phenomenon, and are informed by indirect or direct exposures to the thing itself (see Banks, 2020). With respect to knowledge about robots as a group, media representations have the potential to convey depictions of robots that reinforce existing understandings, to disrupt them, or to shift how those understandings are evaluated. Thus, we built the exploratory portion of this investigation around the following core research question:

> **RQ1**: (How) does viewing positive (vs. negative) robot media portrayals affect participants' mental models for robots?

In addition to exploring the influence of parasocial contact on MMs, however, we also aimed to find out whether the known impact of parasocial contact on outgroup *attitudes* would carry over to robots. For this research interest, we complemented the exploratory work with a theory-driven, deductive approach, considering attitudinal outcomes.

In line with extant evidence on how positive and negative parasocial contact affects attitudes toward human outgroups (Banas et al., 2020), we first considered potential effects on people's preferred social distance—a common concept of attitudinal bias and core variable in contact theory (e.g., Ortiz & Harwood, 2007). We predict:

> **H1:** Viewing positive (vs. negative) robot media portrayals will lead participants to prefer less (a) physical distance, (b) relational distance, and (c) conversational distance to an actual social robot.

We secondly operationalize attitudes toward robots in accordance with extant evidence about fundamental social judgments. Specifically, people are understood to heuristically judge other humans according to *warmth* (i.e., a caring, emotive, and helpful nature) and *competence* (i.e., the ability to pursue goals intelligently; Fiske et al., 2007). This fundamental taxonomy is foundational to stereotyping and evidence indicates that it is also used for judging humanoid robots—typically involving attributions of moderate-to-high competence and low levels of warmth (e.g., Carpinella et al., 2017), although some morphological variants might vary on these evaluations (e.g., domestic robot devices; Reeves et al., 2020). As media depictions of robots tend to rely heavily on warmth and competence for character development—often stereotypically cold or counter-stereotypically warm—we expect that qualitative differences in media portrayals would respectively reinforce or disrupt stereotypical expectations for an actual robot. Focusing on the warmth dimension as a particularly important cornerstone of robot-related perceptions, we hypothesized:

> **H2:** Viewing positive (vs. negative) robot media portrayals will lead participants to perceive an actual social robot as significantly warmer.

In tandem, we contemplated how people's impression of competence might be affected. On the one hand, a helpful, friendly robot might also be perceived as more competent due to its high socio-emotional functionality; on the other, competence (in the sense of *calculating* agency) could be considered as a counterpoint to displays of warm and communal behavior. As such, we pose an open research question regarding this concept:

> **RQ2:** Will viewing positive (vs. negative) robot media portrayals lead to significantly different competence perceptions about an actual social robot?

Method

To address the posed research questions and hypotheses, a two-condition experiment was conducted and analyzed using a multimethod approach. All study materials are available in online supplements (https://osf.io/2qtc4/) and hypotheses and analysis plan were pre-registered (https://aspredicted.org/3TM_9G5). For transparency, we must note deviations from that pre-registration due to unforeseen circumstances: A combination of unusually low study enrollment for this laboratory experiment (a trend continuing from the height of COVID-19), and participant harassment of lab staff required early closure of the study. Thus, the pre-registered sample size of 126 (to detect moderate effects of Cohen's *d* with 80% power) was not met, so low power for statistical analysis is acknowledged as a limitation of this investigation. Specifically, a post-hoc power analysis showed that with the achieved sample size, group differences of medium effect size could only be detected with a reduced power of 67.0%; results should be considered with this limitation in mind.

Participants

$N = 77$ participants (age $M = 28.26$ years, $SD = 13.60$; 49 identifying as female, 28 male) were recruited from a southwestern US university and its surrounding community. They were invited to participate in a two-part study on "perceptions of robots in the media and in the world," incentivized by entry into a drawing for a US$100 Amazon gift card. This recruitment approach garnered an age-diverse sample (18 to 74 years); however, since age and student status did not appear to correspond with any variables of interest (see online supplements), the two groups are here analyzed and reported in aggregate.

Procedure

The study's two-part design comprised an online survey followed by an in-person lab session. The initial survey (hereafter time *T1*) established a baseline for pre-stimulus understandings of and attitudes about robots—namely, participants' mental models for, desired social distance from, and stereotype content (i.e., warmth and competence perceptions) for robots (see Measures section). Upon survey completion, participants were redirected to an online system to sign up for an in-person laboratory session. After scheduling, the session (of one to three participants) was randomly assigned to one of two experimental conditions (positive or negative parasocial contact). Participants were not primed with notions of goodness or badness or made aware of condition assignment, so that any effect would come from the stimulus content itself.

In the lab session (time *T2*) participants were welcomed to a film screening room, given instructions, and then presented with either a positive or negative film reel per the randomly assigned condition (see Stimuli section). Following this treatment, participants completed a tablet-based digital survey, indicating robots they had recognized in the film and, mirroring *T1*, again responding to robot mental model elicitations. Then, they were led to another room to observe a scripted interaction between the experimenter and an actual robot. Finally, participants returned to the tablet survey to again complete the social distance and stereotype content measures, with instruction to consider the actual robot (instead of robots in general).

Stimuli

Parasocial Contact (Positive or Negative Media Treatment)

To create media stimuli for our manipulation of parasocial contact, we consulted several hallmark publications as detailed in the literature review and engaged that literature in an in-depth discussion between both authors. We specifically focused on fundamental aspects of what may be counted as positive vs. negative outgroup contact—and in particular on those characteristics that seemed suitable for extraction from brief segments of existing films. Doing so, we settled on three criteria for comparing positive vs. negative depictions: (1) Emphasizing counter-stereotypical (e.g., warm, communal) vs. stereotypical (e.g., cold, agentic) aspects of the outgroup, (2) suggesting shared vs. diverging group goals, and (3) depicting cooperative vs. competitive behaviors (as an indicator for interdependent vs. independent intergroup dynamics). Moreover, informed by the reviewed literature on the formation of PSIs, we decided to limit the positive contact stimulus to depictions that were overtly likable, sociable, or sympathetic—whereas the negative condition could also involve more sinister or downright threatening portrayals. These conceptual decisions align with the abovementioned focus on the impact of warmth perceptions for human–robot interaction (HRI); while we deemed it suitable for robots in both media conditions to appear more or less competent, only the machines in the positive parasocial treatment were supposed to be seen as warm and helpful.

Having assembled these theoretical criteria, we conducted a search of robot-related media in television and cinema—consulting the International Movie Database (IMDB) and several journalistic reviews (e.g., Wold, 2021). This produced a catalog of candidates for both conditions. We excluded the 50 most popular movies and television shows (based on box office and viewer counts) to minimize any effects from heuristic familiarity or popular discourse. We also excluded robots from animated movies (e.g., *Wall-E, Baymax, Iron Giant*) to avoid diminished realism, as well as those with a non-humanoid design (e.g., *AMEE, Johnny Five*)—keeping in mind that perceived similarity to one's (human) self has been identified as a main predictor of successful PSI formation (Liebers & Schramm, 2019).

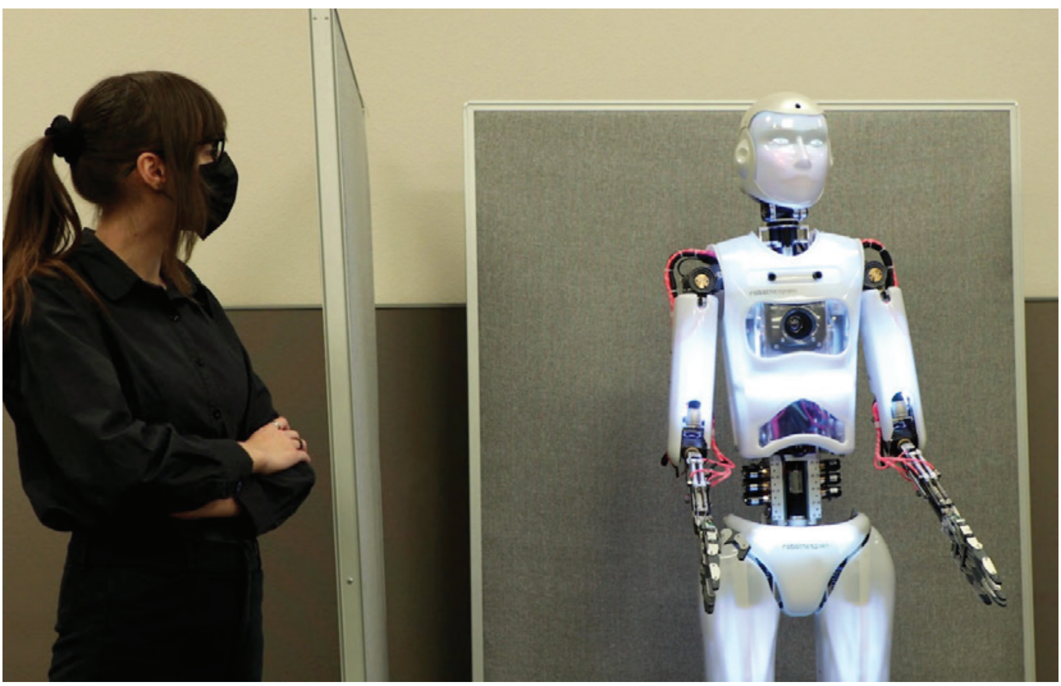
Based on the narrowed selection of eligible characters, we carefully matched exemplars on those criteria to create two contrasting film reels (positive vs. negative parasocial contact), each a montage of 15 scenes from different movies and television shows. Despite presenting different tonalities, group dynamics, and attributes, both reels contained robots with similar designs and levels of human-likeness, as well as the same number of female-coded and male-coded robots. Moreover, both montages ranged in cinematic age, from the first half of the 20th century to the 2010s. Lastly, scenes contained similar numbers of human–robot interactions and reached a similar runtime (positive reel: 715 seconds; negative reel: 769 seconds), though we privileged content parity over length parity as core to the manipulation. After the reels were constructed, undergraduate research assistants (at that point naïve to the aims of the manipulation) confirmed face validity of the positive/negative manipulation. The full storyboards and videos (as well as a detailed overview of our theoretical and design choices) can be found in the online supplements.

Encounter With an Actual Robot

For the actual robot encounter, we settled for a standardized, *observed interaction* between a human confederate and the humanoid robot "Ray"—so as to avoid the disruptive influence of different conversation topics, levels of emotionality, or nonverbal cues as they might have occurred in individual, organic interactions. Ray is a *RoboThespian 4* (Engineering Arts, U.K.) that stands 175 cm (5 feet, 9 inches) tall, is able to move its head and arms, and is stationary from the waist down. Ray was presented as female via the Socibot facial projection (female version "Pris") and American English voice (female version "Heather").

In the prepared interaction, the confederate was a White adult female wearing black clothing and a black mask (Figure 1). She was trained to perform the script as an interview with Ray as a way to introduce the robot to the "guests." A separate confederate controlled

FIGURE 1 Interaction Between the Experimenter and the Social Robot During the In-Lab Session



the robot's (non-)verbal behaviors from an adjacent room (i.e., *Wizard-of-Oz* technique). To keep this in-person encounter as neutral as possible—such that performed positivity or negativity would not override any effect of the experimental stimuli—the dialogue involved neither overly friendly nor unfriendly passages. Instead, Ray described her daily work and gave some basic information about her attributes and functionalities. At the end of the 4-minute interaction, the experimenter requested the robot to go back to "idle mode" and obscured it with a partition. See the project's OSF directory for the full script.

Measures

Mental Model Elicitations

Mental models are understood to be black boxes—people may or may not be aware of knowledge they hold about a phenomenon, and the task of understanding a mental model requires motivating people to externalize their internal knowledge and beliefs while not influencing the content of those externalizations. To achieve this, we adapted an approach from Banks (2021) in posing three elicitations to motivate externalizing of participants' understandings of robots. At both *T1* and *T2*, participants were asked to "In your own words, please explain": (1) ". . . what 'robots' are," (2) ". . . what robots can do," and (3) "your ideas about the roles that robots should play in society." Participants were instructed to think about robots as they exist in the real world, and to provide as much detail as they can.

Quantitative Measures

Desired Social Distance. To measure general attitudes toward (a member of) the robotic outgroup, we used three items capturing desired social distance (Banks & Edwards, 2019). Constituting three distinct facets of approach/avoidance, these items address the desired (a) *physical distance*, (b) *relational distance*, and (c) *conversational distance* to robots. For each Guttman-scaled item, six gradation points were presented to capture participants, comfort with degrees of distance (e.g., physical distance: "I would be comfortable if a robot was . . .," with options "standing next to me," "in the same room," "in the same building," "in the same city," "in the same country," or "none of the above"). As such, higher values (1–6) denote greater preferred social distance. The *T1* measurement addressed robots in general, and *T2* application captured attitudes about the actual robot they had just met.

Stereotype Content. Situating our work in the well-established stereotype content model (Fiske et al., 2007), we employed two scales for perceived *warmth* and *competence* of robots (Liu et al., 2021). Both measures (warmth: 4 items, e.g., "caring," "good-natured"; competence: 5 items, e.g., "intelligent," "competent") were presented in a 7-point Likert format. Again, the instruction was slightly varied between repeated measurements—*T1* addressing robots in general and *T2* the encountered robot in particular. We observed acceptable internal consistency for all applications, Cronbach's α ranging from .72 to .90.

Control Variables and Manipulation Check

At time *T2*, we additionally captured potentially relevant control variables. Firstly, participants were asked to indicate all robots that they recognized in the movie reels from a list of names. Since this list included all robots from both the positive and negative media conditions, we subsequently calculated each participant's *recognition score* as the number of correctly identified robots minus the number of incorrectly identified robots. At the end of the survey, a manipulation check item asked participants whether the robots they saw were "good" or "bad." As only four participants answered this question in a way that did not match their assigned condition, we deem our manipulation of positive vs. negative parasocial contact as sufficiently valid. Lastly, participants were asked whether they had ever before encountered the in-person stimulus robot (which was answered affirmatively by seven participants). Yet, for all of these control and manipulation check items, exploratory analyses showed that removing the respective individuals did not significantly alter our results (see OSF online supplement), so that all participants could be included in our main analyses.

Results

All obtained data and analyses codes are available in this project's OSF directory.

Media Influences on Mental Models for Robots (RQ1)

To first address RQ1—whether exposure to valenced film depictions of robots may influence mental model content—an *inductive thematic analysis* was conducted by the second author in three stages. In the first, a semantic network analysis tool (Leximancer) was used to induce clusters of co-occurring words within the data corpuses (one each for *T1* aggregated, *T2* positive condition, *T2* negative condition). In the second, those clusters were interpreted as representing higher-order themes by iteratively tacking back-and-forth among the concept map depicting the latent concepts and their associations within themes (Figures 2–4), the thesaurus of words underlying each concept, and the source data from which those words were extracted. In doing so, interpretation was aimed at discerning patterns in the concepts independently and then collectively represented by the key terms, ultimately extracting the overarching concept represented in the clusters. To this end, themes that were manifestly similar across the data sets were flagged as such, and then remaining themes were evaluated—first for conceptual similarities and then for hierarchical relations (e.g., lower-order concepts being associated with higher-order concepts). To ensure interpretations of (dis)similarity did not run too far afield from source data, this process included a return to the keywords and then source data to validate inferred associations and divergences among themes. In the third stage, a qualitative comparison was made among interpreted themes between the *T1*, *T2*-positive, and *T2*-negative theme sets. This inductive analysis is conducted at the *group* level, such that claims made are specific to the overarching patterns within each group (aggregate or condition-specific) at a specific point in time (*T1*, *T2*) and not about any one individual. The analysis narrative with details about the data preparation, Leximancer settings, and the interpretive process are available in the OSF online supplements.

Throughout, *concept* refers to the latent idea manifested in the data as induced by the software; a latent idea is predicted from multiple terms and the heaviest weighted (i.e., most predictive) term is the concept name. *Cluster* refers to induced set of concepts that tend to co-occur within a particular participant's response. *Theme* is the researcher-interpreted meaning of the cluster. *Hits* refers to the number of data units associated with a theme.

Identification of Themes

***T1* Themes (All Participants).** *T1* themes were derived from the aggregated responses (i.e., for all three questions) from all participants. In Stage 1 (semantic network mapping), analysis induced 14 clusters comprising 29 latent concepts (hit range 9 to 173; Figure 2). In Stage 2 (theme analysis), clusters were interpreted to represent (from most to least prevalent): relations to humans, benefits, designed functions, applications, potential to improve human lives, potential to take human jobs, roles in society, status as technology, grounding in artificial intelligence, evolving influence in society, everyday computers, capacity boundaries, characterizing contemporary operation contexts, and individual judgments about robots. See Table 1 on pages 167 and 168 for theme definitions and illustrative data excerpts.

FIGURE 2 Semantic Network Map for the *T1* Aggregated Responses to Robot Mental Model Elicitations

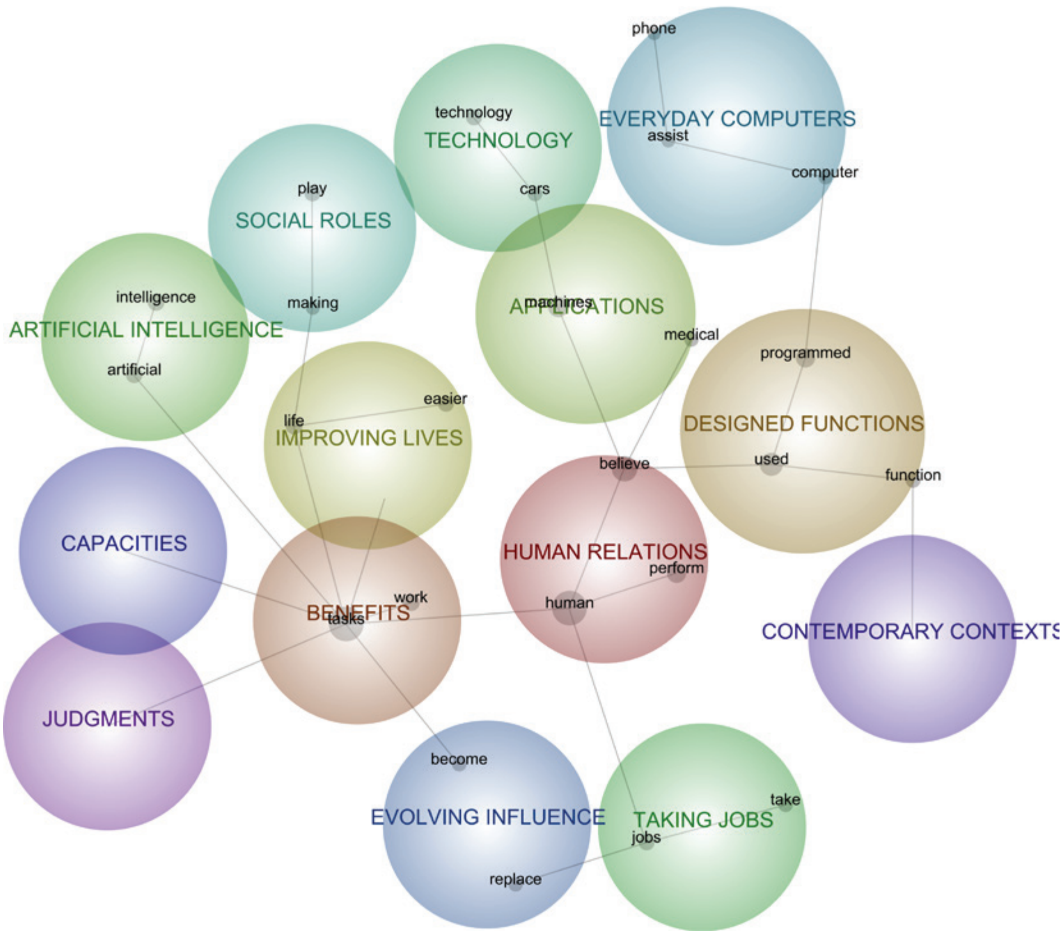


TABLE 1 *T1* Themes in Participant Characterizations of Robots

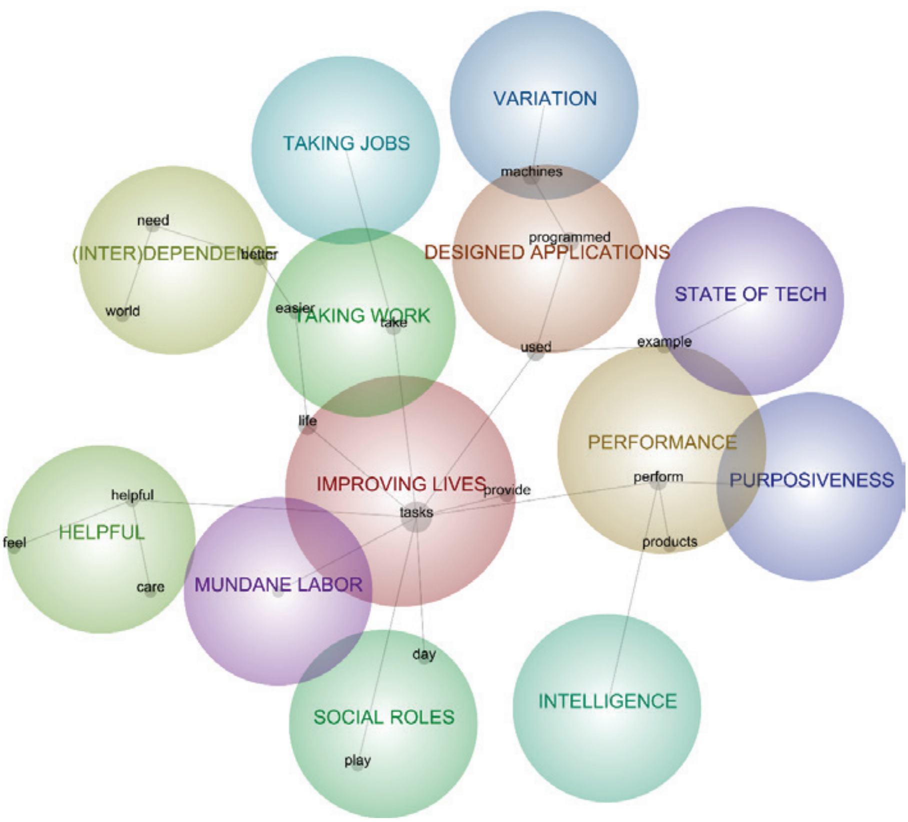
Theme Label	Concepts	Hits	Description	Example Data Extract
Human Relations	Humans, believe, perform	173	Situatedness of robots in relation to humans, especially differences between or alignments with them.	I *believe* robots should be used for good causes like helping the elderly and *people* with disabilities to *perform* daily mundane tasks like cooking …
Benefits	Tasks, work	169	The beneficial outcomes manifested by robots' work.	… [robot labor] can help direct human effort and manpower to other *tasks* that require brain *work*.
Designed Functions	Used, programmed, function	98	Having functions (general or specific) designed by humans.	… robots are *programmed* to do what the creator [intends] … like performing simple *functions* as opening a can of beans. Even other *uses* …
Applications	Machines, medical	62	Examples or lists of how robots do or could play a role in everyday situations.	… they are used for *medical* purposes, but I don't know how. I think bomb squads use *machines* …
Improving Lives	Life, easier	58	Robots can, should, or might improve human lives by making them easier.	… I think robots should exist … to make human *life easier* …
Taking Jobs	Jobs, take	36	Possibility or likelihood that robots will take human jobs.	… there has been much talk about if robots will *take* people's *jobs* …
Social Roles*	Play, making	34	Robots' general role in society, usually linked to making human life easier.	… they can … *play* the role of *making* life easier …
Technology	Technology, cars	33	Are technologies, or can create, contain, or be contained in other technologies.	…. *technology* advances all the time. They're used in manufacturing where robots work on things like car assembly …
Artificial Intelligence	Intelligence, artificial	29	Based on, contains, or functions through AI.	Robots are machines that mimic humans through *artificial intelligence* …
Evolving Influence	Become, replace	28	Are becoming, resulting in increasingly impactful through role displacement or augmentation.	… will continue to *become* more prevalent in the world. I believe they will *replace* many low wage jobs …
Everyday Computers	Computer, assist, phone	26	Are computer assistants already in everyday life.	… even *assisted* in children's education … We utilize robots in our everyday life … The *computer* I'm typing on is leagues smarter than me. The *phone* in my pocket …

Theme Label	Concepts	Hits	Description	Example Data Extract
Capacity Boundaries	Able	18	Have possibilities and constraints in their abilities.	… should *never* be *able* to think for themselves too.
Contemporary Contexts	World	16	Zeitgeist that robots operate in or help to create, usually negatively valenced.	… the last thing people need in this *world* …
Judgments	Feel	13	Expressed feelings about robots' integration (usually negatively valenced).	… it *feels* like a slippery slope and it's difficult to see clearly where it will lead …
Note: *Theme is interpreted to be an unexpected artifact of the elicitation that could not be avoided through term exclusion; it is removed from further analysis.				

***T2* Themes (Positive and Negative Media Conditions, Separately).** *T2* themes were derived from aggregated responses to all three elicitations for each condition-specific group separately, that is, those having viewed the positive (T2P) or negative (T2N) film reels.

For T2P responses, Stage 1 analysis induced 12 clusters comprising 29 latent concepts (hits range 6 to 86; Figure 3). In Stage 2, clusters were interpreted to represent (from most to least prevalent): improving human lives, designed applications, task performance, taking

FIGURE 3 Semantic Network Map for the *T2* Responses to Robot Mental Model Elicitations Following Viewing of a "Good Robot" Film Reel



on workload, human-machine interdependence, taking human jobs, helpfulness, variation in robots and situations, social roles, need for purpose, helpfulness in everyday labor, and relatedness to technology in general. See Table 2.

TABLE 2 ***T2* Themes in Participant Characterizations of Robots Following Viewing of a "Good Robot" Film Reel**

Theme Label	Concepts	Hits	Description	Example Data Extract
Improving Lives	Tasks, life, provide†	86	Performing specific tasks and services (by design) improve human life.	… help aid *human* beings in *tasks* that are demanding … should be used to help *human* beings *live* a better and easier *life* … *provide* care through speech and action to *human* beings.
Designed Applications	Programmed, used, machines	50	Machines created for specific purposes (where purposes were both humanizing and dehumanizing).	… *machines* that are created for a purpose … *used* in the production of parts. Robots can do anything they are *programmed* to do …
Performance	Perform, example	32	Execution of specific tasks (paired with illustrations).	… can *perform* any task as long as it has the right code … for *example,* Alexa can now control the thermostat …
Taking Work	Take	15	Assumption of some work, whether helpful or harmful for humans.	… *take* the load off of our shoulders …
Intelligence	Intelligence	14	(Not) having kinds or degrees of intelligence.	… can be extremely helpful and *intelligent* creations …
(Inter) dependence	Need	14	Things that humans (do not) need from robots or robots from humans.	… machines that do not *need* human control to function …
Taking Jobs	Jobs	12	Possibility or likelihood that robots will take human jobs (for good or ill).	…. machines can do a better *job* than humans because of their increased efficiency ….
Helpful	Helpful	12	Applications, scenarios, or contexts in which robots would be helpful to humans.	… they could be *helpful* inside the household …
Variation	Different	10	Variability in what robots are, what they can do, and how they are distinct from other machines.	… designed with many *different* responses to the original input …
Social Roles*	Play	10	Robots' general role in society, usually linked to making human life easier.	… should *play* supporting roles in human lives …

Theme Label	Concepts	Hits	Description	Example Data Extract
Purposiveness	Purpose	7	Prescriptions that robots must serve purposes defined by humans (versus self-determined).	… I do not believe there should ever be freely roaming around without a *purpose* …
Mundane Labor	Daily	6	Appropriateness of robots helping with the mundane tasks of daily life.	…. perform *daily* tasks at home such as cleaning …
State of Technology	Technology	6	States of technology (broadly) in relation to robot functions of abilities.	… capability varies widely because the access to *technology* varies …
Note: † "human" was also a heavily weighted predictor, though not a formally identified concept. *Theme is interpreted to be an unexpected artifact of the elicitation that could not be avoided through term exclusion; it is removed from further analysis.				

For T2N responses, Stage 1 analysis induced 11 clusters comprising 27 latent concepts (hits range 6 to 85 instances, Figure 4). In Stage 2, clusters were interpreted to represent (from most to least prevalent): improving human lives, designed task performance, designed utility, social roles, status as a technology with specific functions, efficiency benefits, need to accommodate (not disadvantage) humans, appropriateness of providing services, taking risky jobs, existing with a human-defined purpose, and home as a context for labor. See Table 3 on the following page.

FIGURE 4 Semantic Network Map for the *T2* Responses to Robot Mental Model Elicitations Following Viewing of a "Bad Robot" Film Reel

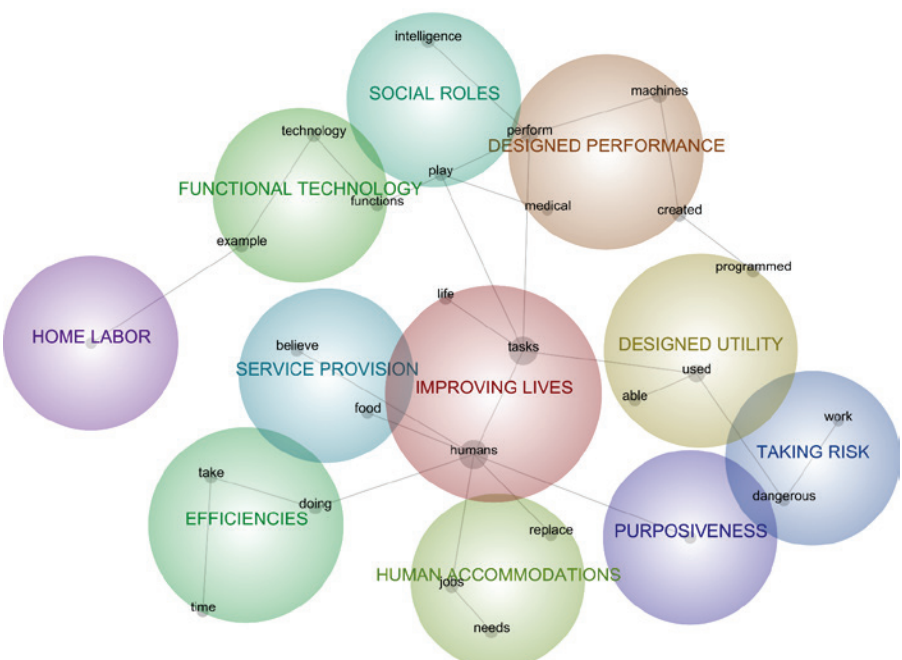
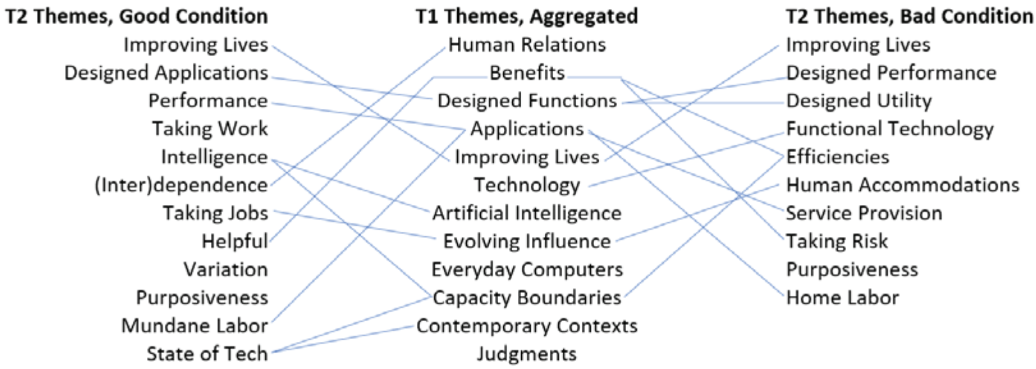


TABLE 3 *T2* Themes in Participant Characterizations of Robots Following Viewing of a "Bad Robot" Film Reel

Theme Label	Concepts	Hits	Description	Example Data Extract
Improving Lives	Humans, tasks, life	85	Performing specific tasks and services (by design) improve human life.	… complete hyper specific *tasks* and problems. Robots should be created and employed to better *human life* …
Designed Performance	Perform, machines, created, medical	45	Performing functions (general or specific) designed by humans.	… *machines* that are *created* to *perform* human activities or tasks …
Designed Utility	Used, programmed, able	44	Used by humans according to the technology's designed abilities.	… *used* to perform routine tasks … the computer *programming* behind them is really the limitation …
Social Roles*	Intelligence, play	26	Robots' general role in society as a function of its intelligence (or that of its creators).	… artificial *intelligence* made for a purpose … can *play* many roles in society.
Functional Technology	Example, technology, functions	26	Technology with particular functions (paired with illustrations).	… piece of *technology* that is very advanced that can perform different *functions* …
Efficiencies	Take, doing, time	25	Improvement in efficiencies through reduced time for tasks.	… should increase efficiency and decrease the *time* certain tasks may *take* …
Human Accommodations	Jobs, replace, needs	25	Prescriptive imperative for robots to fulfill human needs, and not draw (job) resources.	… should … understand the common basic things a person may *need* then if worse comes to worse will start to *replace* people's *jobs* …
Service Provision	Believe, food	19	Belief in the appropriateness of service roles (especially food delivery).	… take *food* orders, deliver food, vacuum, clean house, and … I *believe* they can take the place of some of things that humans do …
Taking Risk	Work, dangerous	13	Taking up dangerous roles to avoid risk to humans.	… do *work* in *dangerous* environments so humans don't have to …
Purposiveness	Exist	10	Existence defined by purpose as ascribed by humans.	… something mechanical that *exists* to aid humans …
Home Labor	House	6	The home as a context for labor.	… make coffee, clean *house,* and do what they are told …

Note: *Theme is interpreted to be an unexpected artifact of the elicitation that could not be avoided through term exclusion; although it appears to focus on intelligence as a concept, most source data is focused more squarely on social roles. In parallel with the T2P decision, it is removed from further analysis.

FIGURE 5 Themes and Theme-Relations for *T1* Robot Characterization, Compared to T2 Positive Condition and T2 Negative Condition Characterizations



Note: Theme labels are drawn from Tables 2–4. Lines indicate interpreted topical similarities in themes between *T1* aggregate themes (center) and T2 condition-specific themes (left and right).

Comparison of Higher-Order Themes

To completely address RQ1, we compare themes derived from *T1* to themes derived from each of T2P and T2N. Topical associations in this comparison are illustrated in Figure 5. Importantly, these are qualitative comparisons made based on themes derived from group-level data, so interpretations and derivative claims pertain only to general patterns across groups (and not about any individual's discrete mental model). As similarities and differences are multiple and nuanced, we separately discuss the observed changes between *T1/T2*, additions and losses of content between *T1/T2*, and comparisons following positive/negative contact.

Post-Stimulus Shifts in Topic Specificity, Prevalence. For *T1/T2*, most notable is the shift from the prevalence of higher-order concepts to more specific concepts. *T1* themes attend to human-robot (non-)relations, robot benefits, robot applications, and robot influence—in a more general sense—while *T2* themes included related but more specific discussions of human-robot interdependence mechanisms, specific benefits like helpfulness and efficiency, specific applications like home labor, and the influences of job displacement. The more general notion of robots being bounded in their capacities (i.e., having potentials and/or limitations in abilities) shifted toward discussions of specific capacities (i.e., intelligence and efficiencies). Moreover, when people discussed these more specific capacities, those themes were more prevalent in discussions (i.e., higher on the theme list). Similarly, ideas about improving human life (whether actual, potential, or prescribed) were of middling prevalence at *T1* but rose to be *most* prevalent at *T2* such that the content remained similar but the discursive weight within the data sets increased after the stimulus film. Finally, although themes of robots as designed (i.e., made by humans) can be seen across both *T1/T2*, at *T1* the consideration is of their designed *function* (that is, what can they do mechanistically) whereas at *T2* the discussions focused more on applications and utility (how humans design them for practical use) and performance (how they are effective according to human design). In other words, in discussing robots there may be a shift from general functioning-by-design to *human-centered* functioning-by-design. Recalling that at both *T1* and *T2* participants were asked to think about robots as they exist *in general*, these shifts are altogether interpreted to suggest that parasocial contact with robots may motivate people to think about robots in ways that are more concrete, where particulars (rather than generalities) may become more salient, and these particulars are considered in relation to human experience.

Post-Stimulus Additions and Losses. Following exposure to the positive stimulus, a theme labeled "taking work" emerged as a multi-valenced consideration of how robots could offload human burdens, distinct from job displacement. Mentions of varied potentials for robots also followed the positive stimulus, highlighting ways that robots serve different purposes, behave in different ways, and perform in different ways. Thus, we did observe some additivity of positive concepts following positive representations—that is, that robots can benefit humans by displacing work (not necessarily jobs) and are not a homogenous category. For *both* conditions we interpreted the emergence of a theme representing purposiveness—that is, prescription that robots should or must exist according to a mandate ascribed by humans. It could be that depictions of human-robot interactions (whether positive or negative) initiate a kind of reactance by which people are compelled to reinforce anthropocentric constraints around robot existence. There are no other additions for the negative stimulus.

Two themes from *T1* did not appear in concept maps at *T2*—or at least ceased to be cohesive themes amid other ideas. The *T1* theme representing robots as everyday computers like smartphones and Alexa—already functioning in human society—fell away at *T2* and mentions of those technologies instead were exemplars for other themes. Additionally, personal judgments about robots (anchored to the word *feel*) fell away, suggesting a decreased weight of *feel* or *feelings* in how people discussed robots following the films.

Comparison of Experimental Groups. Both experimental groups' interpreted themes include robots as improvers of human lives (generally or through specific benefits), having a *designed* status and functioning according to that design, having specific applications and utilities, subject to protection of human interests (taking jobs, human accommodations) and ascribed purposes, mundane labor (in general, or in the home). Thus, there appears to be a substantial amount of MM content that is not a function of parasocial contact valence (though perhaps content activated by seeing any kind of robot media depiction, as discussed above).

However, there are indications that some MM content is new or made salient as a function of contact valence. Responses from the positive contact condition emphasized robots taking on burdens, human-robot interdependence (what each needs from the other), helpfulness (as a self-relevant benefit, versus more general efficiencies), recognition of variation among robots, and more general discussion of intelligence (not necessarily artificial). This set of distinguishing themes is interpreted to suggest that those experiencing positive parasocial contact are perhaps more likely to have salient *social* content in robot mental models (that is, considerations of relatedness, traits, difference, and agency), in addition to content around their functionality. In contrast, responses from the negative contact condition emphasized robots' status as a technology, efficiencies (as a practical matter), requisite accommodation of and service to humans, and taking bodily risk in humans' stead. These distinct themes suggest that negative parasocial contact may promote anthropocentric orientations, maximizing ontological differences and prescribed human primacy.

Pre-/Post-Stimulus Attitude Changes (H1–2, RQ2)

With quantitative measures being only moderately intercorrelated (Table 4), separate *t*-tests were conducted for desired physical, relational, and conversational distance, as well as for perceived warmth and competence. Specifically, we used change scores (subtracting *T1* from *T2* values) as dependent variables in these tests—which allows for a more intuitive interpretation while producing the same results as a repeated-measures ANOVA.

As can be seen in Table 5 on the following page, only one of the five conducted *t*-tests revealed a significant group difference for the change between *T1* and *T2*. Specifically, we found that the parasocial contact conditions evoked a different decrease in desired conversational distance, $t(68) = 2.02$, $p = .047$, Cohen's $d = 0.48$: Viewing the negative reel led to a notably smaller reduction of this variable ($M = -0.43$, $SD = 1.46$) than viewing the positive reel ($M = -1.23$, $SD = 1.83$). In other words, positive parasocial contact more greatly reduced tendencies to be conversationally close to robots (i.e., they would be more intimately disclosing through conversation). There were no significant group differences for stereotype content or for other social distance operationalizations. Findings were robust to age, gender, prior exposure, media character familiarity, and manipulation check covariates (see online supplements). As such, H1c was supported by our data, whereas H1a and H1b are rejected.

TABLE 4 Zero-Order Correlations of the Study Variables

	Variable	1	2	3	4	5	6	7
1	Age	–						
2	Gender[1]	.13	–					
3	Difference in perceived warmth (*t2 – t1*)	.08	–.13	–				
4	Difference in perceived competence (*t2 – t1*)	–.03	–.04	.54***	–			
5	Difference in desired physical distance (*t2 – t1*)	.08	–.06	–.02	–.01	–		
6	Difference in desired relational distance (*t2 – t1*)	–.04	.11	–.36**	–.23	–.03	–	
7	Difference in desired conversational distance (*t2 – t1*)	–.03	.30*	–.32*	–.17	–.12	.44***	–

Note: * $p < .05$, ** $p < .01$. [1]Gender coded with "0" = female, "1" = male, *** $p < .001$.

Table 5 Descriptive and Inferential Statistics Regarding the Examined Group Differences

	Positive parasocial contact		Negative parasocial contact		*t*-test statistics	
	M	***SD***	***M***	***SD***	***t***	***p***
Stereotypes	$n = 35$		$n = 36$			
Difference in perceived warmth ($T2 - T1$)	+1.85	1.45	+1.81	1.34	0.13	.894
Difference in perceived competence ($T2 - T1$)	+0.76	1.22	+0.66	1.55	0.31	.754
Social Distance	$n = 35$		$n = 35$			
Difference in desired physical distance ($T2 - T1$)	−0.20	0.78	−0.54	1.27	1.35	.180
Difference in desired relational distance ($T2 - T1$)	−0.63	1.40	−0.54	1.22	0.27	.785
Difference in desired conversational distance ($T2 - T1$)	−1.23	1.83	−0.43	1.46	2.02	.047

Note: Participants could answer all items voluntarily. This resulted in different final sample sizes for the measures, which are stated accordingly.

Discussion

Recognizing the importance of media exposures in the face of limited experiences with actual social robots, the present study identified a notable effect of positive (versus negative) parasocial contact, as it decreased the desired conversational distance from robots. In tandem, we observed that—although much mental model content about robots persisted through the film exposure—parasocial contact *may* influence mental models for robots as an ostensible outgroup, even after a single, 10-minute treatment. Specifically, it appears that parasocial contact promoted salience of more specific, concrete, and human-centered concepts, where positive contact results in attention to more social considerations and negative contact maximizes ontological differences. We interpret these findings to suggest that valenced parasocial contact with robots likely offers limited-yet-meaningful influences on people's knowledge of and attitudes toward actual robots.

In comparing pre- and post-stimulus concept maps that represented aggregate mental models, we see a good deal of qualitatively similar content—including post-stimulus content similarity between those viewing positive and negative stimuli. We interpret these patterns to suggest that mental models largely persist through parasocial contact valence; nevertheless, the latter *does* seem to introduce small but meaningful changes. Perhaps most important to PCH theory, positive exposure appeared to make salient notions of sociality and positive traits as well as individual differences within the outgroup, while negative exposure highlighted utility and tool-status. This echoes PCH-related findings from the human-to-human context: In interpersonal settings, outgroup members are often dehumanized (i.e., being denied fundamental human traits such as warmth and civility, as well as their individuality; Harris & Fiske, 2006; Haslam, 2006), but positive contact may reduce this bias (Bruneau et al., 2020). As such, we want to stress the additive effects of positive MM content as a particularly noteworthy result of our PCH-guided investigation: Depicting robots as benevolent and non-stereotypical led participants toward a more individualized and social perception of this outgroup.

At the same time, even positive robot portrayals may underscore that they are not human to begin with—which perhaps explains why both experimental groups were nonetheless anthropocentric in orientation. Concept maps for *both* parasocial conditions suggest an increased and more specific inclination to mention the human-made nature of social robots in their mental models at the second measurement point. Even those with positive contact focused on human benefit and those with negative contact attended to topics that maximized ontological differences. We interpret this finding to indicate potential psychological reactance: Faced with elaborate and human-like depictions of robots (regardless of their tonality), participants may have experienced discomfort with the non-familiarity of dramatic human-robot interactions or, more intensely, with a symbolic threat to their human distinctiveness (e.g., Stein et al., 2019). In response to this supposedly unpleasant impression, it could be that notions of human superiority (i.e., people as the *makers* of robots) were invoked as an implicit reclamation of control. In a sense, this interpretation suggests that parasocial contact with robots may also prompt a different kind of reactive dehumanization—one that emphasizes human control through making, using, and assigning purpose.

In the second part of our statistical investigation, we observed that people having negative and positive parasocial contact *both* showed a decrease in desired conversational difference. That is, both groups were more willing to communicate more intimately with the robot after an actual exposure compared to before (likely as a matter of mere exposure under controlled conditions; Haggadone et al., 2021). Importantly, though, those with positive parasocial contact exhibited a much more dramatic reduction in desired conversational distance—very much in line with PCH tenets (Banas et al., 2020). We believe that this finding holds particular relevance for the field of human-machine communication (HMC), which is invested in understanding the dynamics by which humans and machines make meaning together (Guzman, 2018). By increasing people's willingness to approach and share information with robots, media depictions may be a key driver in social closeness or distance that people feel toward robots as a group and as individual social actors. Specifically, this study offers initial evidence that media impressions help to shape more positively and negatively valenced mental models, and so may qualitatively shape people's willingness to engage humanoid robots as an ostensible outgroup.

In summary, by looking at our core results—positive additions to MM content (sociality, individual difference) and decreased conversational difference among those with positive parasocial contact—we conclude that media representing positive robot qualities (and associated positive HRI) could serve as a bridge toward more open communication among humans and machines. More broadly, our work points to the utility of PCH as a promising framework for understanding meaning-making around social robots. Guided by this comprehensive theoretical approach, we not only observed meaningful changes in participants' mental models, but also obtained a significant finding in an underpowered statistical investigation (such that other stereotype content and social distance outcomes could be relevant for a larger sample). Therefore, we invite our peers to follow up on our theoretical groundwork, as HMC studies involving parasocial contact theory might indeed go beyond traditional cultivation or habituation approaches.

Limitations and Future Directions

Several limitations must be considered in this work. We engaged a single set of film stimuli with a narrow selection of (exclusively anthropomorphic) robots, considered by a somewhat narrow sample (i.e., skewing younger in a socially and politically conservative community). Mental model and social judgment effects could vary with differing media and robot stimuli, especially around different machine morphologies—although we suppose that our film stimuli afforded reasonable breadth by integrating multiple dimensions of positive and negative outgroup contact. As such, future work could consider other mass-mediated robot depictions more broadly (e.g., of zoomorphic or fully abstract robots, dramatic situations, and interaction contexts such as dyads versus groups) and more narrowly (e.g., only looking at different robot facial expressions). Additionally, the induction of clusters from qualitative data was completed using a single tool with particular settings and results were interpreted by a single analyst; thus, it is possible that other inquiries using different analytical parameters could identify different outcomes. Thus, as with most exploratory work, this work should be replicated and extended to advance the validity of our claims.

Since our analysis of mental models uncovered that the human-centered attribution of roles to robots seems to be of high importance, future research that applies the PCH to social robots is also encouraged to focus more on different role representations in the media (e.g., a tool, a helper, a guardian) as antecedents of changing perceptions and attitudes. While such efforts could start with replicating our multi-method approach, we suggest that additional measures may be useful. Among the many options in this regard, studies could shift their focus from subjective assessments to more concrete (behavioral observation) or implicit (e.g., IAT) measurements. In the same vein, longitudinal research could help to shed light on the stability of the evoked changes, and to ultimately create a scientific perspective that truly acknowledges quality *and* quantity to a comparable extent.

Conclusion

Fictional media make more or less salient the possible risks and benefits of a world populated by social robots—from lives of increased comfort to impending doom. Research previously examined such effects in terms of exposure quantity, yet the present research draws on parasocial contact theory to augment the record with evidence that exposure *quality* may also play a role by making salient beneficial outcomes from interactions with diverse robots. Through a more comprehensive quantity-and-quality approach afforded by PCH, we may better understand how media help to shape perceptions of sociality and interdependence regarding robots as an outgroup—toward prosocial and antisocial ends.

Author Biographies

Jan-Philipp Stein, Dr. rer. nat., (Chemnitz University of Technology) is associate professor for Media Psychology at Chemnitz University of Technology, Germany. His core research interests include human–machine communication, virtual environments, and social media use.

https://orcid.org/0000-0003-3874-0277

Jaime Banks, PhD, (Colorado State University) is associate professor in the School of Information Studies at Syracuse University, USA. Her research focuses on human-technology relations with an emphasis on social robots and video game characters, and current work focuses on social cognitions and moral judgments in human-robot interactions.

https://orcid.org/0000-0002-7598-4337

Author Notes

The authors have no conflicts of interest to disclose. First authorship is shared between both authors.

Acknowledgments

The authors thank the College of Media and Communication at Texas Tech University, where a portion of this work was conducted. Furthermore, we thank Shibo Bruce Wang, Madison Wedge, Tristan Britz, Julie Grandjean, and Sydney Brammer for their help with data collection.

Center for Open Science

OPEN DATA OPEN MATERIALS PREREGISTERED

This article has earned the Center for Open Science badges for Open Data and Open Materials through Open Practices Disclosure. The authors have made their data and materials freely accessible at https://osf.io/2qtc4/. The article also earned a Preregistered badge for having a preregistered design available at https://aspredicted.org/3TM_9G5.

References

Airenti, G. (2015). Theory of mind: A new perspective on the puzzle of belief ascription. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.01184

Allport, G. (1954). *The nature of prejudice*. Addison-Wesley.

Banas, J. A., Bessarabova, E., & Massey, Z. B. (2020). Meta-analysis on mediated contact and prejudice. *Human Communication Research, 46*(2–3), 120–160. https://doi.org/10.1093/hcr/hqaa004

Bandura, A. (2009). Social cognitive theory of mass communication. In J. Bryant & M. B. Oliver (Eds.), *Media effects: Advances in theory and research* (3rd ed., pp. 94–124). Routledge.

Banks, J. (2020). Optimus prime(d): Media cultivation of robot mental models and social judgments. *Frontiers in Robotics and AI, 7*, 62. https://doi.org/10.3389/frobt.2020.00062

Banks, J. (2021). Of like mind: The (mostly) similar mentalizing of robots and humans. *Technology, Mind, and Behavior, 1*(2). https://doi.org/10.1037/tmb0000025

Banks, J., & Edwards, A. (2019). A common social distance scale for robots and humans. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication* (pp. 1–6). https://doi.org/10.1109/RO-MAN46459.2019.8956316

Banks, J., & Koban, K. (2022). A kind apart: The limited application of human race and sex stereotypes to a social robot. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00900-2

Bruckenberger, U., Weiss, A., Mirnig, N., Strasser, E., Stadler, S., & Tscheligi, M. (2013). The good, the bad, the weird: Audience evaluation of a "real" robot in relation to science fiction and mass media. In *Proceedings of the 2013 International Conference on Social Robotics* (pp. 301–310).

Bruneau, E., Hameiri, B., Moore-Berg, S. L., & Kteily, N. (2020). Intergroup contact reduces dehumanization and meta-dehumanization. *Personality and Social Psychology Bulletin, 47*(6), 906–920. https://doi.org/10.1177/0146167220949004

Bryson, J. J. (2010). *Robots should be slaves*. John Benjamins.

Carpinella, C. M., Wyman, A. B., Perez, M. A., & Stroessner, S. J. (2017). The robotic social attributes scale (RoSAS). *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/2909824.3020208

Craik, K. (1943). *The nature of exploration*. Cambridge University Press.

Fiske, S. T. (2000). Interdependence and the reduction of prejudice. In S. Oskamp (Ed.), *The Claremont symposium on applied social psychology* (pp. 115–135). Lawrence Erlbaum.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

Fox, J., & Gambino, A. (2021). Relationship development with humanoid social robots: Applying interpersonal theories to human–robot interaction. *Cyberpsychology, Behavior, and Social Networking, 24*(5), 294–299. https://doi.org/10.1089/cyber.2020.0181

Fraune, M. R., Nishiwaki, Y., Šabanović, S., Smith, E. R., & Okada, M. (2017). Threatening flocks and mindful snowflakes: How group entitativity affects perceptions of robots. In *Proceedings of the 2017 International Conference on Human-Robot Interaction* (pp. 205–213).

Gamez-Djokic, M., & Waytz, A. (2020). Concerns about automation and negative sentiment toward immigration. *Psychological Science, 31*(8), 987–1000. https://doi.org/10.1177/0956797620929977

Gerbner, G., & Gross, L. (1976). Living with television: The violence profile. *Journal of Communication, 26*(2), 172–199. https://doi.org/10.1111/j.1460-2466.1976.tb01397.x

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

Haggadone, B. A., Banks, J., & Koban, K. (2021). Of robots and robotkind: Extending intergroup contact theory to social machines. *Communication Research Reports*, *38*, 161–171. https://doi.org/10.1080/08824096.2021.1909551

Haring, K. S., Watanabe, K., Silvera-Tawil, D., Velonaki, M., & Takahasi, T. (2015). Changes in perception of a small humanoid robot. In *Proceedings of the 6th International Conference on Automation, Robotics, and Applications, Queenstown, New Zealand* (pp. 83–89).

Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological Science, 17*(10), 847–853. https://doi.org/10.1111/j.1467-9280.2006.01793.x

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Horton, D., & Wohl, R. R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance. *Psychiatry, 19*, 188–211. https://doi.org/10.1080/00332747.1956.11023049

Jackson, J. C., Castelo, N., & Gray, K. (2020). Could a rising robot workforce make humans less prejudiced? *American Psychologist, 75*(7), 969–982. https://doi.org/10.1037/amp0000582

Kahn Jr., P. H., Reichert, A., Gary, H. E., Kanda, T., Ishiguro, H., Shen, S., Ruckert, J. H., & Gill, B. (2011). The new ontological category hypothesis in human–robot interaction. In *Proceedings of the 6th ACM/IEEE International Conference on Human–Robot Interaction* (pp. 159–160).

Koay, K. L., Syrdal, D. S., Walters, M. L., & Dautenhahn, K. (2007). Living with robots: Investigating the habituation effect in participants' preferences during a longitudinal human-robot interaction study. In *Proceedings of the 16th International Symposium on Robot and Human Interactive Communication* (pp. 564–569).

Lee, K. M., Peng, W., Jin, S.-A., & Yan, C. (2006). Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human–robot interaction. *Journal of Communication, 56*, 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x

Li, J. (2015). The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies, 77,* 23–37. https://doi.org/10.1016/j.ijhcs.2015.01.001

Liebers, N., & Schramm, H. (2019). Parasocial interactions and relationships with media characters—An inventory of 60 years of research. *Communication Research Trends, 38*, 4–31.

Liu, S. X., Shen, Q., & Hancock, J. (2021). Can a social robot be too warm or too competent? Older Chinese adults' perceptions of social robots and vulnerabilities. *Computers in Human Behavior, 125*, 106942. https://doi.org/10.1016/j.chb.2021.106942

Mara, M., Stein, J. P., Latoschik, M. E., Lugrin, B., Schreiner, C., Hostettler, R., & Appel, M. (2021). User responses to a humanoid robot observed in real life, virtual reality, 3D and 2D. *Frontiers in Psychology, 12*, 633178.

Meleady, R., & Forder, L. (2018). When contact goes wrong: Negative intergroup contact promotes generalized outgroup avoidance. *Group Processes & Intergroup Relations, 22*(5), 688–707. https://doi.org/10.1177/1368430218761568

Ortiz, M., & Harwood, J. (2007). A social cognitive theory approach to the effects of mediated intergroup contact on intergroup attitudes. *Journal of Broadcasting & Electronic Media, 51*(4), 615–631. https://doi.org/10.1080/08838150701626487

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*(5), 751–783. https://doi.org/10.1037/0022-3514.90.5.751

Reeves, B., Hancock, J., & Liu, X. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior, 1*(1). https://doi.org/10.1037/tmb0000018

Savela, N., Turja, T., Latikka, R., & Oksanen, A. (2021). Media effects on the perceptions of robots. *Human Behavior and Emerging Technologies, 3*(5), 989–1003. https://doi.org/10.1002/hbe2.296

Schiappa, E., Gregg, P. B., & Hewes, D. (2005). The parasocial contact hypothesis. *Communication Monographs, 72*(1), 92–115. https://doi.org/10.1080/0363775052000342544

Smith, E. R., Šabanović, S., & Fraune, M. R. (2021). Human–robot interaction through the lens of social psychological theories of intergroup behavior. *Technology, Mind, and Behavior, 1*(2). https://doi.org/10.1037/tmb0000002

Spatola, N., Anier, N., Redersdorff, S., Ferrand, L., Belletier, C., Normand, A., & Huguet, P. (2019). National stereotypes and robots' perception: The "Made in" effect. *Frontiers in Robotics and AI, 6*. https://doi.org/10.3389/frobt.2019.00021

Stein, J.-P., Appel, M., & Cimander, P. (2022). Power-posing robots: The influence of a humanoid robot's posture and size on its perceived dominance, competence, eeriness, and threat. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-022-00878-x

Stein, J.-P., Liebold, B., & Ohler, P. (2019). Stay back, clever thing! Linking situational control and human uniqueness concerns to the aversion against autonomous technology. *Computers in Human Behavior, 95*, 73–82. https://doi.org/10.1016/j.chb.2019.01.021

Sundar, S. S., Waddell, T. F., & Jung, E. H. (2016). The Hollywood robot syndrome: Media effects on older adults' attitudes toward robots and adoption intentions. In *Proceedings of the 11th International Conference on Human-Robot Interaction (HRI)*.

Tajfel, H., Turner, J. C., Austin, W. G., & Worchel, S. (1979). An integrative theory of intergroup conflict. *Organizational Identity: A Reader*, 56–65.

Taschler, M., & West, K. (2016). Contact with counter-stereotypical women predicts less sexism, less rape myth acceptance, less intention to rape (in men) and less projected enjoyment of rape (in women). *Sex Roles, 76*(7–8), 473–484. https://doi.org/10.1007/s11199-016-0679-x

Tukachinsky, R., & Stever, G. S. (2019). Theorizing development of parasocial engagement. *Communication Theory, 29*, 297–318. https://doi.org/10.1093/ct/qty032

Vanman, E. J., & Kappas, A. (2019). "Danger, Will Robinson!" The challenges of social robots for intergroup relations. *Social and Personality Psychology Compass, 13*(8), Article e12489. https://doi.org/10.1111/spc3.12489

van Oers, R., & Wesselmann, E. (2016). *Social robotics*. KPMG Advisory.

van Straten, C. L., Peter, J., & Kühne, R. (2020). Child–robot relationship formation: A narrative review of empirical research. *International Journal of Social Robotics, 12*, 325–344 (2020). https://doi.org/10.1007/s12369-019-00569-0

Wold, S. (2021, June 17). The 100 greatest movie robots of all time. *Paste Magazine*. https://web.archive.org/web/20210627222844/https://www.pastemagazine.com/movies/robots/the-100-greatest-movie-robots-of-all-time/

Young, K. L., & Carpenter, C. (2018). Does science fiction affect political fact? Yes and no: A survey experiment on "Killer robots." *International Studies Quarterly, 62*(3), 562–576. https://doi.org/10.1093/isq/sqy028

Żerebecki, B. G., Opree, S. J., Hofhuis, J., & Janssen, S. (2021). Can TV shows promote acceptance of sexual and ethnic minorities? A literature review of television effects on diversity attitudes. *Sociology Compass*. https://doi.org/10.1111/soc4.12906

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.10

Boundary Regulation Processes and Privacy Concerns With (Non-)Use of Voice-Based Assistants

Jessica Vitak[1], Priya C. Kumar[2], Yuting Liao[3], and Michael Zimmer[4]

1 University of Maryland, College Park, Mayland, USA
2 Pennsylvania State University, State College, Pennsylvania, USA
3 Intuit Inc.
4 Marquette University, Milwaukee, Wisconsin, USA

Abstract

An exemplar of human-machine communication, voice-based assistants (VBAs) embedded in smartphones and smart speakers simplify everyday tasks while collecting significant data about users and their environment. In recent years, devices using VBAs have continued to add new features and collect more data—in potentially invasive ways. Using Communication Privacy Management theory as a guiding framework, we analyze data from 11 focus groups with 65 US adult VBA users and nonusers. Findings highlight differences in attitudes and concerns toward VBAs broadly and provide insights into how attitudes are influenced by device features. We conclude with considerations for how to address boundary regulation challenges inherent in human-machine interactions.

Keywords: privacy, internet of things, voice-based assistants, communication privacy management theory

CONTACT Jessica Vitak • jvitak@umd.edu • 4130 Campus Drive, College Park, MD, 20742, USA

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

Recent years have seen an explosion in the Internet of Things (IoT) and smart technologies designed to simplify people's lives. IoT refers to a network of interconnected computing components, digital and mechanical objects, and living organisms; each *thing* is given a unique identifier enabling data transfer over the network (Alhammadi et al., 2019). These smart technologies are everywhere—from workplaces to homes, cars, and schools—and include smart light switches, appliances, thermostats, digital assistants, door locks, and more. They enable people to remotely complete routine tasks, such as turning lights on/off, checking refrigerator contents, or adjusting a home's temperature. Importantly, in order to offer such functionality, IoT devices collect and transmit significant amounts of data about people and their environment.

Although IoT devices provide significant utility and convenience, they also raise concerns about what data is being collected, how that data is stored, to whom that data is transmitted, what control users have managing that data, and how that data might be used in the future. These devices contain a variety of sensors that collect audio, location, movement, and other trace data. Analysis of such data can reveal information about people's likes and dislikes, eating and exercise habits, location, and more (Boeckl et al., 2019). Privacy threats include platforms misusing data collected from IoT devices (Lynskey, 2019), law enforcement unexpectedly accessing IoT devices (Díaz, 2020), and harm from intimate partners (Levy & Schneier, 2020).

In this paper, we focus on one of the most popular IoT interfaces—voice-based assistants (VBAs) found in smartphones and smart speakers in millions of homes. Because of their popularity, VBAs provide an important exemplar of human-machine communication; users interact with a human-like conversational user interface to achieve tasks (Guzman, 2019; Weidmüller, 2022). As companies design newer versions of these devices, they introduce new features that collect a wider range of data through more channels, especially once users start linking multiple smart devices together. While marketed as increasing convenience, the influx of audio, video, and other sensor data creates new privacy risks for people deciding which smart devices to use and how to interact with them.

Using data collected from 11 focus groups ($N = 65$) with both VBA users and nonusers in the US, we evaluate how those who regularly use these devices—as well as those who have chosen not to use them—feel about these advances in device features, as well as the wider implications of the growth of IoT technology. We interpret our findings using Petronio's (2002) Communication Privacy Management theory (CPM), which considers the tensions individuals experience when sharing private information and the turbulence that arises when privacy rules are broken. While this theory has largely focused on interpersonal communication, we extend it to human-machine communication to explore factors people consider when deciding whether to use VBAs. We argue CPM provides a useful framework for considering what *ownership* and *control* mean when data is shared with a company rather than an individual, and we reflect on how companies may address challenges with boundary regulation in their features and policies.

Literature Review

One of the most common applications of IoT is in home automation, providing users with convenient ways to manage home appliances, lights, power outlets, door locks, and other smart devices (Zeng et al., 2017). Smart home devices can be managed through a mobile or web interface, or through voice commands to smart speakers via voice-based assistants (VBAs). Use of VBA-powered smart speakers has grown: 33% of the US population (age 12+) owned a smart speaker in 2022, up from 27% in 2020 and 18% in 2018 (Edison Research, 2022). People use VBAs most frequently to access music, conduct hands-free searches, and control other devices connected to smart hubs (Ammari et al., 2019). VBAs can be customized to make routines more efficient, such as lowering lights and playing soothing music at bedtime, and they can support caregiving and accessibility for older adults and people with disabilities (Pradhan et al., 2018).

Privacy Concerns With Smart Home Devices

Although useful in many scenarios, smart devices blur boundaries between public and private spaces, and scholars have started exploring how users and nonusers understand the privacy implications of integrating "always listening" VBAs in home environments. In evaluating the public's understanding of privacy issues related to social robots, Lutz & Tamò-Larrieux (2020) found respondents were most concerned about personal information shared with device manufacturers. This study was largely based on nonusers of smart devices; similarly, Lau et al. (2018) found that nonusers saw little utility in VBAs/smart speakers and were less trusting of service providers, while McLean and Osei-Frimpong (2019) found that perceived privacy risks of smart speakers significantly dampened perceptions of device benefits.

Conversely, researchers have found that VBA users generally have low privacy concerns regarding their smart devices (Lutz & Newlands, 2021). Compared to nonusers, users report higher confidence that companies will ensure the privacy, safety, and security of their data (Liao et al., 2019). VBA users often have a limited understanding of how the systems collect, store, and analyze their data (Lau et al., 2018; Zeng et al., 2017), and news articles have highlighted how data sharing, access, and use by these companies may be surprising or problematic to users (e.g., Day et al., 2019; Fowler, 2018). Likewise, Ammari et al. (2019) found that respondents frequently could not articulate specific privacy concerns; when they did, concerns centered on uncertainty about when the device was listening and third parties accessing VBA data. Zeng et al. (2017) found that respondents rationalized this lack of concern as not feeling personally targeted, trusting potentially adversarial actors, and believing their existing mitigation strategies were sufficient.

These studies highlight that smart device users often express few privacy concerns, but their rationalizations suggest an incomplete understanding of privacy risks, a complicated trust relationship with VBA companies, and reliance on the sociotechnical context in which VBAs reside. Building on this, Easwara Moorthy and Vu (2015) found that when users understand privacy risks associated with VBAs, they attempt to mitigate concerns by using simple strategies (e.g., only using in private spaces). While their research calls for better design of VBAs to account for such user practices, subsequent work suggests that privacy controls are infrequently used and not aligned with user needs (Lau et al., 2018; Malkin et al., 2019).

Framing the Study: Communication Privacy Management Theory

Numerous privacy frameworks provide a means for evaluating how users navigate privacy concerns related to technology. For example, privacy calculus (Culnan, 1993; Laufer & Wolfe, 1977) describes the cost-benefit analysis individuals make when deciding whether to share personal information. In the case of VBAs, privacy calculus would argue that consumers engage in a rational analysis of the risks and benefits of using a smart speaker; if the benefits outweigh the risks, they are more likely to use it. Alternatively, Nissenbaum's (2009) theory of privacy as contextual integrity (CI) asserts that interactions occur in particular contexts, and norms govern people's expectations of how personal information should flow within any given context. If a technology or practice disrupts those norms, it could pose a privacy concern, irrespective of whether the information was public or private. VBAs, which record and transmit audio data to a third party, might represent a disruption of existing informational norms. In the current study, we rely on Petronio's (2002) Communication Privacy Management theory (CPM) to evaluate how users' privacy calculus is impacted by the types of contextual disruptions CI highlights.

Building on work by Altman (1975), Petronio (2002) argues that people engage in a "mental calculus" when making information disclosure decisions. CPM provides insights into how people navigate tensions between revealing and concealing information—tensions that might erupt with changes in pre-existing contextual norms. While CPM is an interpersonal communication theory, we can extend its principles to human-machine communication and interactions between users and smart devices. We argue that such an extension is useful, given that the anthropomorphization of VBAs leads many users to perceive them as social beings (Guzman, 2019).

CPM provides five core assumptions regarding the relationship between individuals and their private information (Petronio, 2002; Petronio et al., 2021). First, people believe they own and have the right to control access to their private information. A smart speaker user would therefore believe they own and control any data collected by their device, including voice commands. Second, people employ privacy rules to control their private information. Privacy rules are generally organized into three categories related to boundary permeability, ownership, and linkage (Xu et al., 2022). When considering interactions between a user and a smart speaker, less control is possible than in interpersonal communication. How privacy rules are enforced is unclear and relies heavily on whether a user trusts the company with whom they share their data.

The third and fourth assumptions of CPM note that private information, once shared, becomes co-owned, and co-owners negotiate rules regarding if, when, and how information can be further shared. Companies' privacy policies provide a legal framework for how they manage that co-ownership; however, a variety of scenarios may cause misunderstandings and rule breakdowns. Fifth, when rules are violated, boundary turbulence arises and may cause relational tensions and a breach of trust. Such turbulence may be challenging to navigate when a company breaks a privacy rule. While the easiest way to resolve turbulence would be to stop using a device, that may not be a feasible solution.

Two recent survey studies have explored how CPM may apply to human-machine communication, and specifically to smart speaker use. Xu and colleagues (2022) found that smart speaker users employed two types of privacy rules when interacting with devices: privacy settings review (ownership rule) and limiting access (permeability rule). However, linkage rules were not observed, likely because users can rarely negotiate with companies regarding data sharing. Kang and Oh (2021) also explored the role that perceived benefits and risks played in the use of privacy management strategies. They found that privacy self-efficacy had a moderating effect on the employment of these strategies; those with high self-efficacy were more likely to engage in higher disclosure and higher boundary control.

Taken together, this prior work on VBAs illustrates how privacy concerns might influence people's adoption and use of smart home devices and VBAs. Specifically, nonusers might be more sensitive to privacy issues, while users might value social and utilitarian benefits over privacy (McLean & Osei-Frimpong, 2019; Zheng et al., 2018) or trust the company to mitigate lingering privacy concerns (Liao et al., 2019). In this paper, we extend this prior work to consider the role that perceived privacy risks play in VBA (non-)adoption and how the addition of advanced features in VBA-embedded devices affects perceptions of privacy risks. Specifically, we ask:

RQ1: How do users and nonusers navigate privacy concerns related to VBAs?

RQ2: How do users' and nonusers' attitudes toward VBAs shift as new features are added that collect more types of data?

Method

This study was conducted at two US public universities: one located on a suburban campus in the eastern US with 41,000 students, the other on an urban campus in the midwestern US with 24,000 students. In January 2018, the authors obtained a random sample of approximately 3,000 university staff at each university and invited them to complete a survey about their VBA use. To help ensure a diverse pool of adult participants, the sample population included all university staff levels, but excluded faculty and undergraduate student employees. Participants could enter their email addresses if they were interested in joining a follow-up focus group. We received survey responses from 1,160 people, and 705 expressed interest in a follow-up study.[1]

We chose focus groups because they are especially useful for exploring perceptions and generating ideas (Straus, 2019). They also provide a natural setting for participants to interact, respond to, and build on others' comments (Krueger, 2014). To maximize the diversity of perspectives, we used criterion sampling (Patton, 2002). We first divided prospective participants into groups based on whether they used home-based VBAs, phone-based VBAs, both, or neither. We then created three types of sessions: (1) users only, (2) nonusers only, and (3) a mix of users and nonusers. We conducted 11 focus groups (2–8 participants per group) with 65 people across the two universities. See Table 1 on the following page for session details.

1. See Liao et al. (2019) for an analysis of the survey data.

TABLE 1 Descriptive Data for Focus Group Sessions

Focus Group #	Group Type	Number of Participants	Gender (% male)	User Type (% VBA user)	Age Mean (SD)
Group 1	User Only	4	25%	100%	41.75 (11.84)
Group 2	User Only	7	14%	100%	39.14 (11.28)
Group 3	Mix	6	50%	50%	39.67 (14.15)
Group 4	Mix	6	50%	33%	36.00 (12.08)
Group 5	Mix	8	63%	38%	38.13 (14.23)
Group 6	User Only	6	50%	100%	39.50 (15.15)
Group 7	Nonuser Only	4	25%	0%	35.25 (16.68)
Group 8	User Only	8	25%	100%	35.13 (11.49)
Group 9	Mix	8	63%	75%	31.38 (9.16)
Group 10	Mix	6	17%	67%	37.33 (8.94)
Group 11	Nonuser Only	2	50%	0%	50 (12.76)
Totals	**4 User, 5 Mixed, 2 Nonuser**	**65**	**40%**	**66%**	**37.45 (11.23)**

Each session lasted 1 hour and included a semi-structured protocol, starting with questions about participants' general attitudes toward new technologies, followed by a discussion about their use (or non-use) of VBAs. For each session, a moderator from the research team guided the participants through the prepared questions, while a second team member observed and took notes. Participants viewed a commercial for the newly released Amazon Echo Show—which includes a screen, camera, and additional integrations with other smart devices—and shared their reactions. We chose the Echo Show because it encapsulated broader trends in IoT development, including advanced audio and visual features and deeper links into ecosystems of devices and accounts. In some sessions, participants also discussed the Echo Look, a recently released device at the time of data collection that included a camera and was marketed as a tool to upload pictures of outfits and get fashion advice from peers. At the conclusion of each session, participants received a US$15 Amazon gift card.

Sessions were audio recorded, and files were transcribed and imported into Dedoose for qualitative analysis. Data analysis included two cycles of coding (Miles et al., 2014). First, the research team developed an initial codebook based on the interview protocol

TABLE 2 Subset of First-Round Codes From Qualitative Analysis of Focus Groups

Code Name	Code Description
Compare VBAs	Explicit statements comparing features of or attitudes toward two or more versions of VBAs (e.g., Siri, Home, Echo Show).
Privacy-Security	Talking broadly about how technology affects privacy, security, surveillance, and related topics. Strategies used to attain desired level of privacy/security. Comments about corporations using/accessing their data.
VBA Listening	Responses to question, "Do you have a sense of when these devices are listening for your voice or if they're always listening?" General comments about VBA microphones and their capabilities, as well as concerns about when VBAs are capturing audio data or what happens to that data.
Nothing to Hide	Comments that there are minimal risks to using VBAs (e.g., "life is boring").
Privacy Apathy	Comments reflecting belief that privacy is dead, we're already tracked in many ways, etc.
Echo Show	Comments and discussion after watching the Echo Show commercial.

and researcher notes from the sessions (*provisional* or *protocol* coding). Each team member coded a transcript separately, noting where new codes could be added or existing codes collapsed. The team met to refine and finalize the codebook. Two team members then coded each transcript, with the team meeting regularly to resolve coding differences by consensus. For the second cycle, the team identified six codes relevant to this study's research question (listed in Table 2). Excerpts were exported into Excel, and each team member selected specific codes and analyzed the excerpts for patterns (*pattern coding*). For instance, one pattern in the *VBA listening* code was perceptions of home VBAs as more invasive than phone-based VBAs. The team discussed these patterns and linked them to the research questions to identify key themes related to the research questions. All participant names reported below are pseudonyms to protect participant identities.

Findings

RQ1: Rationalizing Privacy Concerns in VBA (Non-)Adoption

We observed notable differences in how VBA users and nonusers talked about privacy concerns. Aligning with and extending prior work, we found that users focused more on the benefits of the technology—often downplaying privacy risks because they felt the data was not sensitive, or felt they lacked any meaningful ability to control data collection in the first place—while nonusers described privacy concerns as one of the reasons they avoided VBAs.

VBA users lacked a sense of data sensitivity and felt little ability to control their data, leading to lower privacy concerns and a focus on utility. While data ownership and control are key components of CPM, many VBA users expressed little interest in managing their voice data due to a perception that the data is not sensitive—and thus posed no risks. For example, James said, "There's nothing I would share that Alexa would hear that would embarrass me at any point in time." Likewise, Emma said she doesn't worry about potential security risks from these devices because she is not doing anything to warrant attention: "I'm boring. I don't have my ballistic missiles sitting in my living room." Others described their lives as "uninteresting" and unworthy of government focus, as when Jackie said, "I live a very boring and average life. I would probably never be tagged by the FBI or anything like that because I don't do anything." These comments align with the nothing to hide trope (Madden & Rainie, 2015), which argues that only "bad" people have things they want to keep private. For example, John said, "If you're gonna be that concerned about a device listening in, chances are you're probably doing something you really don't want people overhearing."

Others' comments referenced a bigger challenge with data ownership: as VBAs are merely the latest in an ongoing expansion of data-hungry technologies, some felt they no longer own their data—and thus lack ways to meaningfully control it. For example, Charlotte said,

> I think there are video cameras on every street. They are watching us everywhere; they are listening to our every peep and move . . . I guess I don't know how to prevent that or what to think about it. It just doesn't seem like there's a lot of privacy anymore.

Some users framed potential privacy risks in relation to other privacy/security threats, rationalizing their VBA use in ways that reflected broader attitudes toward privacy that go beyond data shared through device interaction. As Anthony noted, "The bigger security concern is if I use Alexa to purchase something. Is that machine any more vulnerable when I put my credit card into a dozen different websites? That level of security is what I'd be most worried about."

In light of their perceived lack of control, these participants may have instead prioritized the perceived benefits of VBAs as part of their privacy calculus. Brian reflected broadly on this when he said,

> no matter what technology you use, I feel like if they want to find something, they can find out . . . your phone is tracked wherever you go, so they can tell you your whole life story if they wanted to.

Participants also shared examples that highlighted their lack of control. Kyle noted that data breaches at major corporations suggest that our data is already "out there," while Anne spoke about searching for something on Google only to see ads for that product on other sites.

The belief that data collection and surveillance are omnipresent—and that individuals have little control over what data is collected and who has access to it—led to a sense of apathy and resignation toward data collection among many people we spoke with. Jackie said "it's useless to fight" to protect personal data, and that the increasing reliance on technology will lead future generations to "be even more used to technology . . . People are just going to accept this information." Veronica echoed this sentiment, saying, "I don't think there's running away from technology that we can do efficiently in this age, and I don't mind."

Veronica's comment that she "doesn't mind" technological advances was reflected in several comments that align with the privacy calculus people engage in when deciding if and how to use technologies. For example, Adam said, "I feel like a lot of these companies are collecting these data anyways. I don't like that they do, but if they're going to collect it, I'd rather get the most utility out of it as possible." In that same session, Jay added, "I realized if I'm gonna have a modern smartphone, I'm always gonna have that technology and I can't guarantee it's turned off, so I might as well use it. I mean, it's built in—there's no escaping it."

Nonusers stressed the need for trustworthy providers and control over access to information before they would consider adopting VBAs. While many VBA users shared feelings of resignation toward data collection, those who had *not* adopted VBAs expressed a range of privacy concerns when describing their decision not to use them. Participants' comments referenced trust-related concerns, as well as a desire to control access to their data, reflecting the need to mitigate potential boundary turbulence before adopting VBAs.

Nonusers referenced their use of other Google or Amazon services and data they already shared with these companies. Unlike VBA users, who rationalized their use by saying the company already had their data, nonusers wanted to minimize the data these tech giants had about them, so their privacy calculus was somewhat different. Jada said, "I have a Google phone and Google accounts. I feel like Google knows everything about my life. But I still worry about setting myself up to use a device that would know more information about me." Trust also played an important role, which Gwen noted:

> I think there's a bit of a trust factor for me. I don't really trust the corporations, so I'm only willing to let them into parts of my life where I'm like, "Okay, this is really useful." And I also think as we get more smart devices around our home, it's just easier for them to be hacked, and I think that's going to happen more and more.

Likewise, Leah expressed concerns about trading personal information for minimal benefits, like using VBAs to play music:

> It's one more thing that is used to collect data on you; I assume it's one more thing that can be hacked. I'm old-fashioned. I'm happy with the radio and CDs [compact discs]. I can take those extra four steps to the radio or CD player and turn it on.

At the time of data collection, several media reports had identified bugs with Amazon's Echo devices, including a heavily covered story of Alexa laughing without being prompted (Chokshi, 2018). From a CPM perspective, such accounts can be viewed as instances of turbulence, as they violate people's expectations of how the device works, what data it collects, and how it uses that data. In interpersonal relationships, individuals may re-negotiate rules following such turbulence; in the case of VBA nonusers, such stories may reify their choice. Cliff shared:

> When the review units of the . . . Google Home Mini went out, the button was constantly pushed to listen by manufacturing defect. So here's a device that's constantly listening and they get updates continuously from the server. Let's say somebody wanted to change it; how hard would that be to get it to change?

Walter stopped using Google Assistant after hearing concerning news stories "of people just mentioning certain words and suddenly, boom, the phone's responding." He also worried about weak security protocols in IoT devices making everything more vulnerable: "I don't want to have the ability to turn on and off a light and someone can come in and steal what's on my hard drive."

Other participants worried about unknowns associated with these devices, including how their data could be used in the future and security risks posed by the wider IoT ecosystem. Wade pointed to the newness of these technologies and the lack of existing legislation to protect consumers:

> Probably the biggest drawback for me in terms of not wanting to get one is there's a lot of unknowns, it's all pretty new. Until there's legal precedent, or more history behind it, I don't really want to jump into it.

Likewise, Nina felt the lack of clarity in data collection processes was unnerving, saying, "I don't want a corporation listening to what's going on in my household. I don't know what it's recording. I don't know what's being done with that information."

RQ2: Shifts in Privacy Attitudes Across Types of VBA Devices

Our second research question considered how participants responded to advances in VBAs' features. Initially only available on smartphones, VBAs have expanded to a variety of home devices, including versions with cameras and screens. Features in newer versions of smart speakers aim to reduce friction between users and the task they want to accomplish, which requires greater access to user data and complicates communication processes. Participants discussed their (dis)comfort with these features, and across both users and nonusers, they described newer VBAs—and smart technologies more broadly—as increasingly "creepy," which echoes previous research looking at user perceptions of data collection by mobile apps (Shklovski et al., 2014).

As devices move from phones to homes, friction decreases and privacy concerns increase. During each focus group, we began by discussing phone-based VBAs, including Apple's Siri and Google's Assistant. Most participants reported using phone-based VBAs at some point, although they often described technical issues that limited device utility. For example, participants described having a hard time accomplishing tasks, like when Jordan said he didn't use Siri much because "she didn't really accomplish [requests I gave her] well." Jordan used both the Amazon Echo and Google Home and was much more favorable toward home-based VBAs.

Some participants referenced specific VBA features when describing their concerns. For example, Jin said, "I don't feel like Siri is listening [all the time], because she doesn't turn on unless I press my home button and say 'Hi, Siri.'" Erika echoed this, saying, "I don't have an Alexa or Google Home. But I have Google [Assistant] on my phone . . . and I really like that I have to trigger it." Renee suggested that explicit triggering features kept VBAs from entering "creepy" territory: "If you have to trigger it, it's not creepy. . . . I don't mind saying 'Okay, Google,' but if it's still listening and I don't want it to be listening anymore, that's creepy." Importantly, different VBAs have different activation features, but home devices are typically activated by voice alone, whereas the original versions of Siri and Google Assistant required users to hold down a button to activate the feature. Home VBAs may have a "mute" button, but this significantly reduces the utility of the device, and prior research suggests they are not widely used (Lau et al., 2018).

Many participants expressed concern that their speakers were always listening—not just when they spoke the activation phrase—based on personal experiences. For example, Marilyn said, "She's [Alexa] definitely always listening because randomly she thinks she hears 'Alexa' but we never said that and she will start talking. In that aspect, it's clear that they are always listening and who knows if they are saving [it]." Relatedly, some users expressed concerns that anyone could trigger the device, like when Faith described a movie setting off her Echo device: "It's kind of creepy because we'd be watching in the living room and the dad would shout the daughter's name [Alexis] and all of a sudden you'd hear, 'I'm sorry, I didn't quite catch that.'"

Addressing these perceived risks requires trust between users and the companies providing these devices, especially given that it is often unclear what data is being collected and how it is used. But this also raises questions of whether the companies *should* be trusted. This sentiment was highlighted by Huong, who said, "We're trusting Google that what they show me . . . is what they kept. For the most part, I trust Google on that, and Amazon. But there's that open concern; it's like, what are you opening yourself up to?" Building on this, participants expressed concerns about not knowing *when* these devices were listening and *how much* they captured. Jackie said:

> . . . it's always listening for you to say "Alexa." Do I really know it's not listening to other things? What if it's listening to a conversation about my religious or political beliefs and it's tagging things? I don't want to sound paranoid, but I really don't trust corporations and I don't trust the government to not do those things just because they say it's wrong.

Because of these concerns, several participants said they refused to put home-based VBAs in particularly private places like bedrooms. James said he wouldn't even put a TV in his bedroom because of privacy concerns. Likewise, Chen described why she removed her Echo device from her bedroom:

> I'm really concerned about privacy . . . I remember at first when I put it in my bedroom, and we talked about my son whose name is Max. I don't know what the similarity was, maybe Alexa and Max. And it started to work and joined the conversation. So it made me mad.

From listening to seeing, newest VBAs are perceived as creepy and invasive. In each session, participants viewed an Amazon-produced Echo Show commercial and discussed their reactions. In several sessions, a related product (the Echo Look) came up because it shared camera features with the Show. While some participants noted the benefits of more advanced VBAs (e.g., Huong described the convenience of having a screen so she can see how much time is left after setting a timer), the word "creepy" emerged repeatedly, without prompting, by users and nonusers in nearly all focus groups.

The main Echo Show feature that provoked strong responses from participants was the "Drop-In" feature, which Amazon describes as a "two-way intercom." For this feature to work, users create a list of approved contacts they can connect with. Once a contact approves this privilege, they can instantly connect via audio (on Echo devices) or video (on the Echo Show). One participant, Sun-Joo, shared her experiences trying out Drop-In on her Echo Dot, describing tensions between feeling connected to her family and being *too* connected:

> I don't need them to call me every minute of the day. If it tells them I'm active, they know I'm home, so if I don't answer, I get a text message, "Hey, where are you? I just tried to call you." . . . I'm trying to find a balance.

No other participants had direct experience with the feature.

Immediate reactions after watching the commercial reflected wariness toward features like Drop-In, with participants describing them as creepy and invasive. For example, Liz said, "I'm the kind of person that has a piece of tape over my computer camera because I don't trust it. So the Drop-In thing, that's creepy." Likewise, Walter described the stress of having to be more aware of what he did in private spaces. Speaking about the Echo Look—described as a "Hands-Free Camera and Style Assistant with Alexa"—he asked, "What happens when you come out of the shower and it takes a picture of your body and tells you you need to diet, you need to exercise more?" Multiple participants expressed concerns about Echo devices equipped with cameras, especially since the Look is marketed for bedroom use (to provide feedback on outfits). Olivia said:

> I feel a little uncomfortable with the idea of a camera that could always be on because they always say cover your laptop camera . . . if you had something that had a camera that was looking into your bedroom or an intimate space, I feel like that's really creepy. If somebody were to hack that or hack a Drop-In and just like, actively watch you . . . I don't like that.

While participants' initial reaction to the Echo Show captured its general "creepiness," their comments also reflected feelings of weariness toward and being overwhelmed by more invasive technologies that collected more data, both in terms of quantity and quality. These devices led them to think about more things that could go wrong (e.g., camera positioning, being careful about what you say near the device)—such as when Huong said, "I don't have a problem with pointing cameras outside, but I'm not too comfortable with the cameras inside always on"—or to voice displeasure with technology making them always accessible, as when Sun-Joo described her experience with the Drop-In feature (detailed above).

Managing devices could also get overwhelming, as when John described conflicting feelings about his devices:

> There are times when I very much love having everything connected and hooked up. But then, after awhile, it just gets a little bit where I'm like this is too much. And trying to find that balance is definitely an interesting tightrope to walk because I definitely see the advantages and benefits of it, but at the same time, I'm like, you know, is it too much?

Moving beyond VBAs to consider the wider ecosystem of smart devices in homes—as well as improvements in machine learning that enable devices to make better predictions—these themes of wariness and weariness were exacerbated further. Some participants expressed discomfort with widespread data collection and sharing between companies, while others expressed concerns related to the increasing reliance on technology to accomplish basic tasks. For example, Rebecca asked, "Where do we draw the line? To the point where we're 100% dependent upon devices doing certain things for us?" Zack also pushed back against extreme customization, sharing how he tried to sabotage the underlying algorithm in his VBA: "I've been trying to feed it specific information and it fails in so many ways to get any type of personalized response."

Discussion

In this study, we have explored the role that privacy considerations play in (non-)use of voice-based assistants (VBAs), as well as how privacy concerns are shifting as smart technologies add new features, collect more data, and become better equipped to make inferences and recommendations based on user data. VBAs help us better understand human-machine communication, as users vocally interact with smart speakers to accomplish a variety of tasks (Guzman, 2020). Researchers have described VBAs, and the smart speakers that house them, as hybrids between humans and machines (Weidmüller, 2022) and have found that users attribute human-like characteristics to them (Etzrodt & Engesser, 2021; Garcia et al., 2018; Guzman, 2020). VBAs also provide an important case study for evaluating privacy risks of broader IoT ecosystems because of how they are perceived by users, where they are used (private spaces), the types of data they collect (audio/video), and their function as a hub for a range of smart home devices.

CPM (Petronio, 2002; Petronio et al., 2021) provides a useful framework for considering how people balance the benefits and risks of technologies like VBAs. Recent studies have extended this theory—which was developed to address interpersonal relationships—to human-machine interactions (e.g., Kang & Oh, 2021; Xu et al., 2022). In this paper, we build on these studies to consider how both users and nonusers rationalize decisions related to these devices, using data from focus groups to unpack the complex set of factors that influence these decisions.

CPM is guided by a set of assumptions that helps explain why so many users we spoke to expressed cynicism and apathy toward data privacy. In interpersonal relationships, people negotiate rules related to ownership and control of personal information—and re-negotiate those rules when they experience privacy breakdowns (Petronio et al., 2021). One's relationship with a VBA—and by extension the company that manages that VBA—is much more one-sided, with users often having to agree to certain rules and restrictions via terms of use. It is unsurprising, then, that participants felt less agency and described their data as already being "out there" when news stories regularly highlight data breaches, scandals, and other uses of their data that go beyond expectations (Sheshadri et al., 2017).

CPM helps us move beyond simple explanations that people "just don't care" about their privacy anymore, a sentiment suggested in many studies of technology use. For example, work evaluating privacy attitudes toward IoT found that perceived benefits and organizational trust positively influenced willingness to share personal information, but perceived risks and information sensitivity had no effect on use (Kim et al., 2019). The authors suggest consumers place higher value on the benefits of these technologies and "do not pay much attention" (p. 278) to IoT-based privacy risks. Such an interpretation may apply to active VBA users, but it does not address the privacy concerns raised by nonusers—many of whom noted their concerns were a major factor in the decision to not use VBA devices. For nonusers, organizational trust may be lower—a factor prior work has associated with VBA nonusers (Lau et al., 2018)—and their desire for data control likely supersedes perceived benefits when making purchasing decisions.

More than a decade ago, boyd (2010) noted that networked publics like social media were blurring boundaries between public and private spaces. We argue that smart devices further complicate this blurring due to their widespread popularity, the passive nature of most data collection, and the limited ability to view and edit that data. This limited access to data makes it exceedingly challenging to identify rule violations. Rather, users must trust companies are abiding by the rules they've laid out in their terms of use; even when a rule is violated, there is often little recourse outside of unplugging or removing the device.

CPM focuses on boundary regulation—individuals negotiate how thick or thin a boundary should be for a given piece of private information (Petronio et al., 2021). More sensitive information tends to have thicker boundaries to better protect it from undesired access, while thinner boundaries enable easier flow of information. Our contemporary technological ecosystem increasingly relies on thin boundaries to facilitate the flow of data from individuals to other people (e.g., through social media posts) and companies (e.g., through automated data collection). Researchers have sometimes framed this focus on increasing boundary permeability in terms of "information friction" (e.g., Floridi, 2005), which describes the amount of work required for an entity to access another's information. VBAs provide an example of how this concept works in practice: by default, devices are always listening for a "wake word"—this reduces friction for a person interacting with the device, but increases risks related to inappropriate or unintended data flows. To increase friction, one could use the mute button; however, by removing hands-free interaction, a primary benefit of smart devices is lost. Friction can also be embodied in privacy settings and features that help verify users' intentional interaction with a smart device, and can be useful in verifying things are operating as they should; however, research suggests that users rarely employ available privacy settings (Lau et al., 2018; Malkin et al., 2019).

Individuals' ability to engage in boundary regulation is further challenged when the "rules" constantly change, as can be seen in both updates to terms of use and in the frequent release of new or updated technologies. Participants discussed concerns about feature creep—the ongoing expansion of device features that facilitate additional data collection and monitoring (Surowiecki, 2007). Participants worried that devices were always listening, and their concerns increased for newer devices with cameras. Concerns also emerged from uncertainty around *when* devices collected data and *what* happened to collected data. In 2019, Amazon responded to these concerns by adding new features to allow users to repeat their last command and to explain why it made a recommendation (Ellis, 2019); however, such features focus on transparency rather than providing opportunities to regulate boundaries and control information flows.

What would boundary regulation of VBA data look like? One example is IoT Inspector (Huang et al., 2020), which allows users to capture, analyze, and visualize network traffic generated within their smart homes. Researchers continue to develop ways to increase users' awareness of data flows generated from smart devices (see, for example, Thakkar et al., 2022)—something many of our participants expressed a desire to see in new devices. We hope that future work continues to explore options for helping users negotiate their interactions with machines as they are increasingly confronted with challenges to privacy-based decision-making.

Conclusion

Research suggests IoT technologies will continue to expand over the next decade, as will the push toward creating smart home ecosystems that provide instant access to and control over one's home environment. With such expansion comes greater privacy risks, and we must continue to evaluate how users assess and respond to these risks. By extending CPM to human-machine interactions, we can further explore how communication behaviors—including the disclosure of private information—are shaped by the features and affordances of these technologies.

Such evaluations can also inform future designs of sociotechnical systems to empower users through flexible and intuitive interfaces that provide more transparency about what data is collected and more control over how data is used. Given that our participants expressed concerns regarding AI and devices that collect a greater variety and quantity of data, it becomes even more important to provide users with ways to increase friction and restrict data flows. Skeba and Baumer (2020) provide a useful initial consideration of how the use of AI, algorithms, and big data reduce friction and impact privacy, but more research is needed in this space. Finally, future research in this space must consider how to effectively communicate information about data collection and use so people can make fully informed decisions before sharing their data.

Author Biographies

Jessica Vitak, PhD, is an associate professor in the College of Information Studies and director of the Human-Computer Interaction Lab (HCIL) at the University of Maryland. Her research evaluates the privacy and ethical implications of big data, the internet of things, and other "smart" technologies. She seeks to understand how privacy concerns play a role

in technology adoption and use, and she develops tools and resources to help children and adults make more informed decisions when using technology and sharing sensitive data.

http://orcid.org/0000-0001-9362-9032

Priya Kumar, PhD, is an assistant professor at Pennsylvania State University's College of Information Sciences and Technology. Her research on the datafication of family life aims to shift digital technology discourse and design away from a focus on individual control and toward more networked understandings of privacy and agency. For more information, visit https://priyakumar.org/.

http://orcid.org/0000-0001-9244-7915

Yuting Liao, PhD, is a senior UX researcher at Intuit. Her research focuses on trust, privacy, and data ethics to guide technology design and build harmonious sociotechnical interactions in various contexts, including conversational AI, Fintech, and health technology.

http://orcid.org/0000-0002-5008-2097

Michael Zimmer, PhD, is a privacy and data ethics scholar whose work focuses on digital privacy and surveillance, the ethics of big data, internet research ethics, and the broader social and ethical dimensions of emerging digital technologies. Dr. Zimmer is an Associate Professor in the Department of Computer Science at Marquette University in Milwaukee, Wisconsin, where he also serves as Director of Marquette's Center for Data, Ethics, and Society.

http://orcid.org/0000-0003-4229-4847

Acknowledgments

The authors thank Katie Chamberlain Kritikos for her contributions to the project. This study was funded by the National Science Foundation under Grant Nos. 1640640 and 1640697.

References

Alhammadi, A., AlZaabi, A., AlMarzooqi, B., AlNeyadi, S., AlHashmi, Z., & Shatnawi, M. (2019). Survey of IoT-based smart home approaches. *2019 Advances in Science and Engineering Technology International Conferences* (pp. 1–6). IEEE. https://doi.org/10.1109/ICASET.2019.8714572

Altman, I. (1975). *The environment and social behavior: Privacy, personal space, territory, crowding*. Brooks/Cole Publishing Company.

Ammari, T., Kaye, J., Tsai, J. Y., & Bentley, F. (2019). Music, search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction*, *26*(3), 1–28. https://doi.org/10.1145/3311956

Boeckl, K., Fagan, M., Fisher, W., Lefkovitz, N., Megas, K., Nadeau, E., Piccarreta, B., Gabel O'Rourke, D., & Scarfone, K. (2019). *Considerations for managing internet of things (IoT) cybersecurity and privacy risks* (NIST Internal or Interagency Report 8228). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8228 boyd, danah. (2010). Social network sites as networked publics: Affordances, dynamics, and implications. In Z. Papacharissi (Ed.), *A networked self* (pp. 39–58). Routledge. https://doi.org/10.4324/9780203876527-8

Chokshi, N. (2018, March 8). Amazon knows why Alexa was laughing at its customers. *The New York Times*. https://web.archive.org/web/20180309093135/https://www.nytimes.com/2018/03/08/business/alexa-laugh-amazon-echo.html

Culnan, M. J. (1993). "How did they get my name?": An exploratory investigation of consumer attitudes toward secondary information use. *MIS Quarterly*, *17*(3), 341–363. https://doi.org/10.2307/249775

Day, M., Turner, G., & Drozdiak, N. (2019, April 24). Amazon's Alexa team can access users' home addresses. *Bloomberg*. https://web.archive.org/web/20190426112016/https://www.bloomberg.com/news/articles/2019-04-24/amazon-s-alexa-reviewers-can-access-customers-home-addresses

Díaz, Á. (2020). *When police surveillance meets the 'internet of things'*. Brennan Center for Justice. https://web.archive.org/web/20201218170200/https://www.brennancenter.org/our-work/research-reports/when-police-surveillance-meets-internet-things

Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy concerns for use of voice activated personal assistant in the public space. *International Journal of Human-Computer Interaction*, *31*(4), 307–335. https://doi.org/10.1080/10447318.2014.986642

Edison Research. (2022, March 23). The infinite dial 2022. https://web.archive.org/web/20220325084543/https://www.edisonresearch.com/the-infinite-dial-202

Ellis, C. (2019, September 25). Amazon Alexa gets new privacy controls, and will tell you what it's hearing. *TechRadar*. https://web.archive.org/web/20190928170929/https://www.techradar.com/news/amazon-alexa-gets-new-privacy-controls-and-will-tell-you-what-its-hearing

Etzrodt, K., & Engesser, S. (2021). Voice-based agents as personified things: Assimilation and accommodation as equilibration of doubt. *Human-Machine Communication*, *2*, 57–79. https://doi.org/10.30658/hmc.2.3

Floridi, L. (2005). The ontological interpretation of informational privacy. *Ethics and Information Technology*, *7*(4), 185–200. https://doi.org/10.1007/s10676-006-0001-7

Fowler, G. A. (2018, May 24). Hey Alexa, come clean about how much you're really recording us. *Washington Post*. https://web.archive.org/web/20201114141514/https://www.washingtonpost.com/news/the-switch/wp/2018/05/24/hey-alexa-come-clean-about-how-much-youre-really-recording-us/

Garcia, D. M. P., Lopez, S. S., & Donis, H. (2018). Voice activated virtual assistants personality perceptions and desires: Comparing personality evaluation frameworks. *Proceedings of British HCI 2018* (pp. 1–10). BCS Learning and Development Ltd. https://doi.org/10.14236/ewic/HCI2018.40

Guzman, A. L. (2019). Voices in and of the machine: Source orientation toward mobile virtual assistants. *Computers in Human Behavior*, *90*, 343–350. https://doi.org/10.1016/j.chb.2018.08.009

Guzman, A. L. (2020). Ontological boundaries between humans and computers and the implications for human-machine communication. *Human-Machine Communication*, *1*, 37–54. https://doi.org/10.30658/hmc.1.3

Huang, D. Y., Apthorpe, N., Li, F., Acar, G., & Feamster, N. (2020). IoT Inspector: Crowdsourcing labeled network traffic from smart home devices at scale. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(2), Article 46. https://doi.org/10.1145/3397333

Kang, H., & Oh, J. (2021). Communication privacy management for smart speaker use: Integrating the role of privacy self-efficacy and the multidimensional view. *New Media & Society*, 146144482110266. https://doi.org/10.1177/14614448211026611

Kim, D., Park, K., Park, Y., & Ahn, J.-H. (2019). Willingness to provide personal information: Perspective of privacy calculus in IoT services. *Computers in Human Behavior*, *92*, 273–281. https://doi.org/10.1016/j.chb.2018.11.022

Krueger, R. A. (2014). *Focus groups: A practical guide for applied research*. SAGE.

Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening?: Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), Article 102. https://doi.org/10.1145/3274371

Laufer, R. S., & Wolfe, M. (1977). Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of Social Issues*, *33*(3), 22–42. https://doi.org/10.1111/j.1540-4560.1977.tb01880.x

Levy, K., & Schneier, B. (2020). Privacy threats in intimate relationships. *Journal of Cybersecurity*, *6*(1), tyaa006. https://doi.org/10.1093/cybsec/tyaa006

Liao, Y., Vitak, J., Kumar, P., Zimmer, M., & Kritikos, K. (2019). Understanding the role of privacy and trust in intelligent personal assistant adoption. In N. G. Taylor, C. Christian-Lamb, M. H. Martin, & B. Nardi (Eds.), *Information in contemporary society* (pp. 102–113). Springer International Publishing. https://doi.org/10.1007/978-3-030-15742-5_9

Lutz, C., & Newlands, G. (2021). Privacy and smart speakers: A multi-dimensional approach. *The Information Society*, *37*(3), 147–162. https://doi.org/10.1080/01972243.2021.1897914

Lutz, C., & Tamó-Larrieux, A. (2020). The robot privacy paradox: Understanding how privacy concerns shape intentions to use social robots. *Human-Machine Communication*, *1*, 87–111. https://doi.org/10.30658/hmc.1.6

Lynskey, D. (2019, October 9). "Alexa, are you invading my privacy?" The dark side of our voice assistants. *The Guardian*. https://web.archive.org/web/20191010025233/https://www.theguardian.com/technology/2019/oct/09/alexa-are-you-invading-my-privacy-the-dark-side-of-our-voice-assistants

Madden, M., & Rainie, L. (2015). *Americans' attitudes about privacy, security and surveillance*. Pew Research Center. https://web.archive.org/web/20191104064056/https://www.pewresearch.org/internet/2015/05/20/americans-attitudes-about-privacy-security-and-surveillance/

Malkin, N., Deatrick, J., Tong, A., Wijesekera, P., Egelman, S., & Wagner, D. (2019). Privacy attitudes of smart speaker users. *Proceedings on Privacy Enhancing Technologies*, *2019*(4), 250–271. https://doi.org/10.2478/popets-2019-0068

McLean, G., & Osei-Frimpong, K. (2019). Hey Alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, *99*, 28–37. https://doi.org/10.1016/j.chb.2019.05.009

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.

Nissenbaum, H. (2009). *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press.

Patton, M. Q. (2002). *Qualitative research & evaluation methods*. SAGE.

Petronio, S. (2002). *Boundaries of privacy: Dialectics of disclosure*. SUNY Press.

Petronio, S., Child, J. T., & Hall, R. D. (2021). Communication privacy management theory: Significance for interpersonal communication. In D. O. Braithwaite & P. Schrodt (Eds.), *Engaging theories in interpersonal communication* (3rd ed., pp. 314–327). Routledge. https://doi.org/10.4324/9781003195511

Pradhan, A., Mehta, K., & Findlater, L. (2018). "Accessibility came by accident": Use of voice-controlled intelligent personal assistants by people with disabilities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Article 459). New York: ACM. https://doi.org/10.1145/3173574.3174033

Sheshadri, K., Ajmeri, N., & Staddon, J. (2017). No (privacy) news is good news: An analysis of *New York Times* and *Guardian* privacy news from 2010–2016. *Proceedings of the 15th Annual Conference on Privacy, Security and Trust* (pp. 159–168). IEEE. https://doi.org/10.1109/PST.2017.00027

Shklovski, I., Mainwaring, S. D., Skúladóttir, H. H., & Borgthorsson, H. (2014). Leakiness and creepiness in app space: Perceptions of privacy and mobile app use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2347–2356). New York: ACM. https://doi.org/10.1145/2556288.2557421

Skeba, P., & Baumer, E. P. S. (2020). Informational friction as a lens for studying algorithmic aspects of privacy. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), Article 101. https://doi.org/10.1145/3415172

Straus, R. A. (2019). *Mastering focus groups and depth interviews: A practitioner's guide*. Paramount Market Publishing.

Surowiecki, J. (2007, May 21). Feature presentation. *The New Yorker*. https://web.archive.org/web/20141003144234/https://www.newyorker.com/magazine/2007/05/28/feature-presentation

Thakkar, P. K., He, S., Xu, S., Huang, D. Y., & Yao, Y. (2022). "It would probably turn into a social faux-pas": Users' and bystanders' preferences of privacy awareness mechanisms in smart homes. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (Article 404). New York: ACM. https://doi.org/10.1145/3491102.3502137

Weidmüller, L. (2022). Human, hybrid, or machine? Exploring the trustworthiness of voice-based assistants. *Human-Machine Communication*, *4*, 85–110. https://doi.org/10.30658/hmc.4.5

Xu, K., Chan-Olmsted, S., & Liu, F. (2022). Smart speakers require smart management: Two routes from user gratifications to privacy settings. *International Journal of Communication*, *16*(0). https://ijoc.org/index.php/ijoc/article/view/17823

Zeng, E., Mare, S., & Roesner, F. (2017). End user security and privacy concerns with smart homes. *Proceedings of the Thirteenth USENIX Conference on Usable Privacy and Security* (pp. 65–80). USENIX Association. https://dl.acm.org/doi/10.5555/3235924.3235931

Zheng, S., Apthorpe, N., Chetty, M., & Feamster, N. (2018). User perceptions of smart home IoT privacy. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), Article 200. https://doi.org/10.1145/3274469

HUMAN-MACHINE

COMMUNICATION

Human-Machine Communication
Volume 6, 2023
https://doi.org/10.30658/hmc.6.11

Who Is (Communicatively More) Responsible Behind the Wheel? Applying the Theory of Communicative Responsibility to TAM in the Context of Using Navigation Technology

Sungbin Youk[1] **and Hee Sun Park**[2]

1 Department of Communication, University of California, Santa Barbara, Santa Barbara, CA, USA
2 School of Media and Communication, Korea University, Seoul, Republic of Korea

Abstract

By examining how perceived usefulness and ease of use relate to the user's perception (i.e., communicative responsibility), the communicative behavior of the navigation system (i.e., the landmarks used to give directions), and the context of driving (i.e., familiarity of the driving location), this study applies the theory of communicative responsibility to the technology acceptance model to better understand why users are more likely to adopt certain navigation technologies while driving. We hypothesized that users' perceived symmetry in communicative responsibility independently and interactively (with communicative behavior of the navigation system and the driving situation) affects perceived ease of use and usefulness of the navigation system. Consequently, the perceived ease of use and usefulness may affect the user's intention to use the navigation system. This study found that usefulness was a significant predictor of behavioral intention. While driving in a less familiar location, the drivers perceived the navigation system to be more useful. When the navigation system provided location-specific landmarks, such as the name of a local store, drivers who attributed more communicative responsibility to the system were likely to find it useful.

Keywords: Theory of Communicative Responsibility, Technology Acceptance Model, navigation technology, common ground

CONTACT Hee Sun Park • heesunpark@korea.edu • School of Media and Communication • Korea University, 606, 145 Anam-Ro, Seongbuk-Gu, Seoul, Republic of Korea

ISSN 2638-602X (print)/ISSN 2638-6038 (online)
www.hmcjournal.com

Introduction

Previous studies on human-computer interaction (HCI), human-machine communication (HMC), and human-robot communication (HRI) explored a wide range of factors that can enhance the user's intention to accept the technology. For instance, emotional expressions (e.g., Pantic & Rothkrantz, 2003), nonverbal cues (e.g., Samara et al., 2019; Van Erp & Toet, 2015), and the quality of information (e.g., Diop et al., 2019) provided by the machine, the user's perception of the machine (e.g., Fox et al., 2015), and the cultural contexts of the interaction (e.g., Heimgärtner, 2013) are some of the commonly examined factors that are especially relevant to communication science. These factors can be categorized into three groups: communicative behavior of the machine, perception of the user, and the context of the interaction. Technology acceptance literature has extensively examined the effect of these factors individually. There are relatively fewer studies that integrate these three strands of factors into a single, theoretical, comprehensive model.

Scholars have studied HMC by comparing it against human-human interaction (e.g., J. Meyer et al., 2016; Waytz et al., 2010). As human-human interaction is the most common and extensively studied communication context, theories rooted in interpersonal communication can help us better understand HMC (e.g., the nuance of when and how people apply scripts from interpersonal communication to interact with machines; Gambino et al., 2020). To provide empirical evidence regarding the applicability and contributions of interpersonal theory in improving our understanding of HMC, this study applies the theory of communicative responsibility (CRT) to the context of HMC. According to CRT, interlocutors perceive the amount of responsibility they and their communicative partners bear to create a shared understanding (Aune et al., 2005). The communication context may influence the perceived communicative responsibility and dictate the communicative behavior. By considering the interplay between perceived communicative responsibility, the communicative behaviors, and the communication context, the interlocutor can evaluate the appropriateness of the interaction.

This study integrates CRT into the technology acceptance model (TAM; Davis, 1989) to better understand the mechanism that underlines why users are more likely to adopt certain technologies. As the original study of CRT examined human-to-human interaction in the context of navigation (i.e., giving directions), this study focuses on understanding users' willingness to use a navigation system while driving. It is hypothesized that users' perceived symmetry in communicative responsibility independently and interactively (with communicative behavior of the navigation system and the driving situation) affects perceived ease of use and usefulness of the navigation system. Consequently, the perceived ease of use and usefulness affects their intention to use the navigation system.

Theory of Communicative Responsibility

According to Aune et al. (2005), CRT postulates that interlocutors and communicative parties make judgments about how much responsibility each of them bears to create a mutual understanding (i.e., co-creation of particular meaning or thought in the listener's mind as intended by the speaker, Clark, 1992). From Grice's perspective (1989), the common goal of communication is to establish a shared understanding and knowledge. In other words, when communicating, people have the responsibility and commitment to engage in collaborative efforts in achieving this goal (Geurts, 2019). This responsibility is called communicative responsibility. The communicative responsibility is what communicative partners recognize, estimate, and consider while using conversational implicature (i.e., the notion of conveying information beyond what is apparent; Ahlsén, 2008) and inference-making (i.e., interpretation and comprehension of conversational implicature; Ahmed & Shazali, 2010). The extent to which the communicative party's implicature can be interpreted by the other party is at the heart of successful communication (Mahmood, 2015).

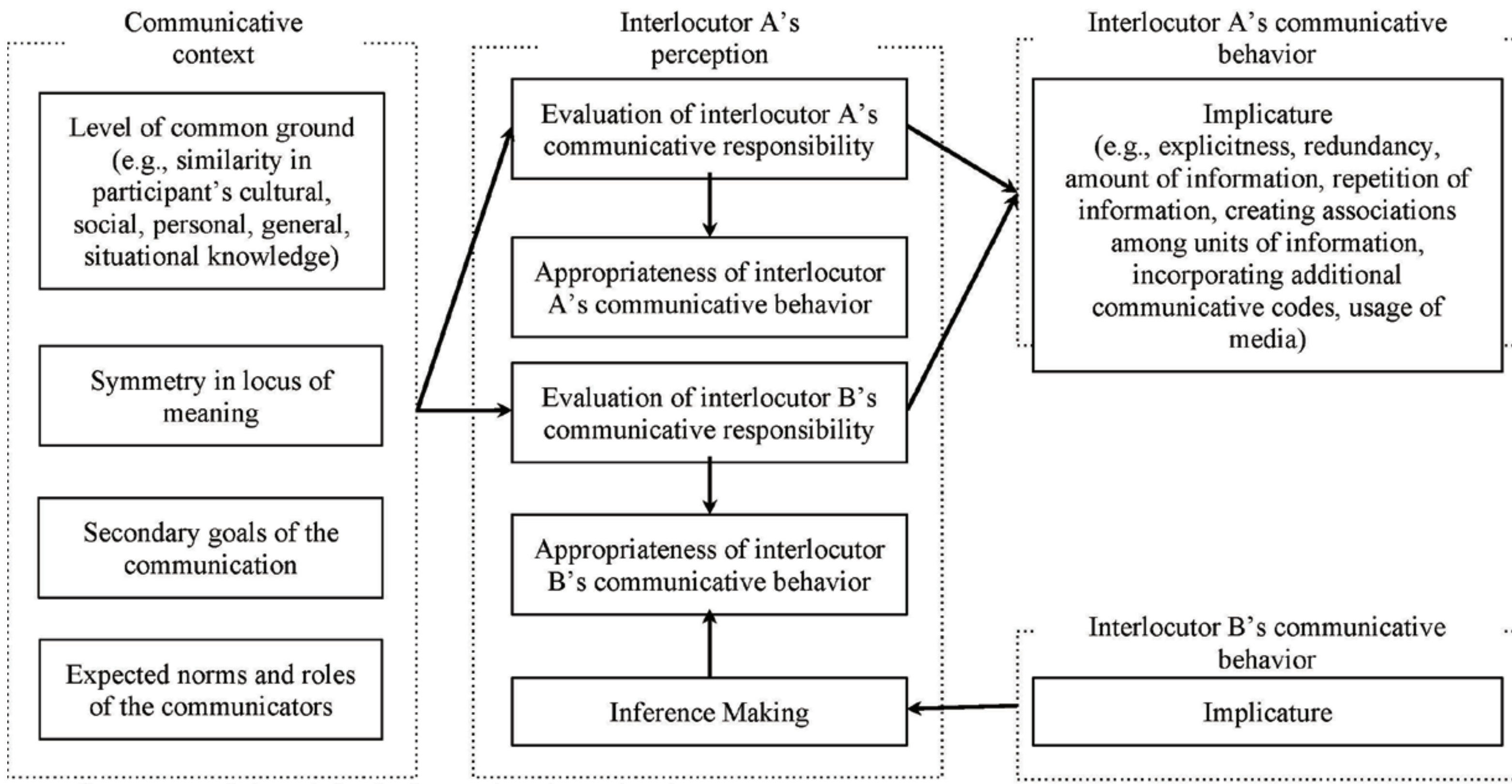
CRT predicts that (a) people adjust their communicative behaviors (i.e., how they express meaning either verbally or nonverbally) according to the relative level of communicative responsibility they have compared to the communicative partner and (b) people judge the appropriateness of their partner's communicative behavior based on the relevant level of communicative responsibility. See Figure 1 for the conceptual map of the predictors and the determinants of CRT as explicated by Aune et al. (2005). For example, when a traveler asks a resident for directions, the resident is responsible for making sure that the traveler understands how to get to the destination, while the traveler is responsible for understanding the provided directions. As the traveler is less familiar with the location, the resident is likely to have a higher communicative responsibility than the traveler. If the resident provides vague directions without reference to easily noticeable landmarks, the traveler will have difficulty understanding the directions (i.e., having difficulty in inference-making). In this case, the resident's communicative behavior is inconsistent with the communicative responsibility. Therefore, the traveler may perceive the resident's communication as inappropriate as the provided directions were not helpful.

Communicative responsibility is a function of common ground. Common ground refers to the mutual knowledge that the interlocutors share (Stalnaker, 1978). Interlocutors' mental states, sociocultural background, prior experience, familiarity with the communication context, and roles in the communication may affect the availability of resources to create meaning out of the communication and the level of common ground (Kecskes & Zhang, 2009; Lau et al., 2001). When there is a lack of common ground, the communicative responsibility is less likely to be shared equally. For example, consider the following two situations:

> **Situation 1:** A and B are both born and raised in the same neighborhood. They are attending the same school.
>
> **Situation 2:** A is new to a neighborhood, while B is born and raised in the neighborhood. From today on, A will attend B's school.

In Situation 1, A and B have a high common ground (i.e., similar knowledge about how to navigate to school). Hence, they are equally responsible for the collaborative effort to arrive at their destination. In Situation 2, A and B will have little common ground regarding the map of the neighborhood. In this case, their communicative responsibility is asymmetrical; B has a higher communicative responsibility than A because of the difference in familiarity with the area.

FIGURE 1 Central Mechanisms of CRT From One Interlocutor's Perspective



Note. See Aune et al. (2005) for further explanations of each concept.

While communicating, people have the innate tendency to monitor the level of common ground. People are cooperative in their communication (Grice, 1989). They build on the conversation and speak in ways that the other person can understand because people can engage in perspective-taking to estimate how much information is mutually shared (Fussell & Krauss, 1989). Based on the expected information that the partner knows, people present information and communicate in a way that can be interpreted as they intended (Clark & Wilkes-Gibbs, 1986; Kecskes & Zhang, 2009). Going back to the two situations mentioned above, B expects A to have little knowledge about the neighborhood. B is likely to attribute less communicative responsibility to A. This indicates that B needs to be explicit, talk in great detail, repeat what was said, and avoid giving directions using landmarks that are specific to the neighborhood to increase the chances of A fully understanding the directions as intended by B (Aune et al., 2005; Lau et al., 2001).

Applying CRT to HMC

The meaning-making process in human-human interaction and HMC are not identical but have overlaps (Gambino et al., 2020; Waytz et al., 2010). The former involves two entities that have the autonomous, biological, and psychological ability to formulate, receive, comprehend, and respond to the communicated message while integrating the complexities of social, environmental, and cultural contexts (Fortunati & Edwards, 2020). Machines only *simulate* having a communicative competence and intelligence (at least for now). Therefore, it is the user who interprets the communication as if machines have the social capacity (Fortunati & Edwards, 2020). For example, interlocutors assume that the other person is telling the truth (i.e., factually correct information to the best of his or her knowledge) unless there is evidence to believe otherwise (see Levine, 2019). By default, communication is built on people telling the truth and cooperating with others to create a common understanding. Similarly, people perceive machines to provide factually accurate information, are agreeable, and unconditionally accept the user's request, but have limited capacity to make evaluative judgments and understand complex human language, social context, and emotional cues (Gambino & Liu, 2022). Consequently, HMC is characterized by people having less concern about impression management (J. R. Meyer, 2009; Veletsianos et al., 2008), burdening or inconveniencing the machine (Gambino & Liu, 2022), and risks of self-disclosure (Ta et al., 2020).

This study applies CRT to the context of HMC for the following reasons. First, interpersonal communication theories and paradigms can be helpful in understanding HMC (Fortunati & Edwards, 2020; Gambino & Liu, 2022; Spence, 2019). As interpersonal communication is the most well-known and extensively studied context, it provides a comparable yardstick. Second, the foundation of CRT is aligned with the primary goal of HMC. Like in interpersonal interaction, fostering a mutual understanding is critical for a successful HMC and HCI (Chai et al., 2014; Stubbs et al., 2007). Last, CRT may provide an answer to one of the "enduring problems" related to effective HMC: uncovering the development of common understanding and shared perception between human and machine (Patterson & Eggleston, 2018). What lies beneath the roadblock to understanding HMC is "a lack of critical knowledge about human cognition" (Patterson & Eggleston, 2018, p. 249). CRT offers a perspective to understand human cognition in the communicative process.

Empirical evidence suggests how CRT predictions apply to HMC context. People use physical, linguistic, and social cues to estimate what the machine knows (Kiesler, 2005). For instance, people evaluate how much common ground is being established based on the occupation and the persona (e.g., personal history, memory, and preference) of the machine (M. K. Lee & Makatchey, 2009). People are likely to include more details in their message when they perceive the machine to share less common ground (Kiesler, 2005).

In particular, CRT is applicable to the driving context of HMC. Using a navigation system is synonymous with the interpersonal communication context given above and Aune et al.'s (2005) experiment. According to CRT, the driver and the navigation system can share different levels of communicative responsibility. The communicative responsibility may change depending on the context of the situation. For instance, a driver expects the navigation system to have a higher communicative responsibility when driving in an unfamiliar location compared to driving in a familiar location. However, if drivers estimate the level of common ground based on their communicative role, the situational context may have a negligible impact on the communicative responsibility. In Hinds et al.'s study (2004), people were asked to complete a task with a computer agent. The study suggested that when the computer agent has a supervisory position but does not have the necessary skills, the subordinate human evaluated the interaction negatively. Given the expected role and duties of a supervisor, the computer agent is expected to communicate in a way that helps the subordinate to understand what needs to be done to complete the task. Behaving inconsistently with the expectation made the subordinate to evaluate the supervisor negatively. In the context of driving, the navigation system's role is to provide accurate directions to help the driver get to the destination. Therefore, the driver may consistently perceive the navigation system to have high communicative responsibility. Therefore, the following research question is asked:

> **RQ1:** Do participants perceive the navigation system to have more communicative responsibility than themselves when driving in an unfamiliar location as opposed to driving in a familiar location?

Technology Acceptance Model

Technology Acceptance Model (TAM) is a widely used theoretical framework that explains why people accept and adopt certain technologies. According to Davis (1989), people are more inclined to use a technology that is perceived to be useful (i.e., how well the technology enhances the user's performance) and easy to use (i.e., how effortless the user can use the technology). This parsimonious model is empirically well-validated across multiple meta-analyses and literature reviews (Al-Emran & Granić, 2021; Granić & Marangunić, 2019; Y. Lee et al., 2003; Ma & Liu, 2004; Marangunić & Granić, 2015; Tao et al., 2020; Yucel & Gulbahar, 2013).

Despite the salience and popularity of TAM, it is not without limitations (see Bagozzi, 2007). For instance, critics of TAM elaborated on the lack of comprehensiveness. To overcome the shortcomings of TAM, scholars have continuously developed various models to study the acceptance of technology: Unified Theory of Acceptance and Use of Technology (UTAUT; Venkatesh et al., 2003), UTAUT2 (Venkatesh et al., 2012), TAM2 (Venkatesh

& Davis, 2000), TAM 3 (Venkatesh & Bala, 2008), and many extended versions of TAM (e.g., Y. Lee et al., 2003; Granić & Marangunić, 2019). UTAUT integrated elements across eight models, including TAM, and UTAUT2 incorporated three additional constructs to UTAUT. TAM2 extended TAM by adding external factors that affect perceived usefulness (e.g., subjective norm, imagination, job relevance, output quality). TAM3 combined TAM2 and various other determinants of perceived ease of use, resulting in a total of 13 factors (excluding the interaction effects) that affect perceived usefulness, perceived ease of use, and consequently behavioral intentions. Like TAM2 and TAM3, other extended versions of TAM (e.g., Go et al., 2020) also have introduced other external factors impacting the two core predictors of behavioral intention (i.e., perceived usefulness and perceived ease of use).

To have a better understanding of how the context of communicative interaction affects acceptance of navigation technology, this study incorporates CRT into TAM. We use TAM as the theoretical backbone instead of other recent alternatives for the following reasons: (a) TAM is parsimonious and extensively validated; (b) TAM has been well-applied to technologies related to driving; (c) TAM is continuously growing; (d) despite all the determinants added to TAM, the factors that are related to communicative interaction are often overlooked.

TAM is parsimonious and extensively validated (Al-Emran & Granić, 2021; Y. Lee et al., 2003; Marangunić & Granić, 2015). Its simplicity and understandability encouraged scholars to use TAM as their theoretical framework (Legris et al., 2003). Consequently, various studies related to technology acceptance are somewhat standardized and comparable (Y. Lee et al., 2003), which facilitates the accumulation of knowledge. According to Yucel and Gulbahar's qualitative content analysis of TAM research (2013),

> although numerous attempts have been made to add other variables to existing ones, the main variables that the 'Technology Acceptance Model' was based on remain the most effective . . . This finding brings us to the understanding that whatever the shape, color, size, and property of the technology, acceptance of that technology can ultimately be determined by using the same variables. (p. 106)

Considering the novelty of CRT, which introduces the determinants that encapsulate the context of communicative interaction, utilizing TAM enables this study to be rooted in the foundation of technology acceptance literature.

TAM has been well-applied to technologies related to driving. Initially, TAM was introduced to explain the adoption of computers in 1989 (Davis). Since then, scholars applied TAM to various technologies, including navigation systems (Eriksson & Strandvik, 2009; Park et al., 2015; Park et al., 2013). Although UTAUT (the more comprehensive model of technology acceptance) outperforms TAM on average (Yucel & Gulbahar, 2013), Rahman et al.'s study (2017) on advanced driver assistance systems demonstrated that TAM explained significantly more variance than UTAUT. This indicates that the different models of technology acceptance can vary in their performance depending on the technology and the context in which it is used.

TAM is continuously growing and expanding as new determinants are introduced to the core variables (i.e., perceived usefulness, perceived ease of use, and behavioral intentions).

TAM3 provides 32 determinants (including interaction effects) of perceived usefulness and perceived ease of use, and 9 determinants of behavioral intention (Venkatesh & Bala, 2008). Similarly, UTAUT also provides a total of 41 independent variables for predicting the intention of accepting the technology (Bagozzi, 2007; Venkatesh et al., 2003). The continuous expansion of the model with a large number of antecedents has multiple drawbacks. First, from a statistical perspective, the model is subject to redundancy, over-fitting, and multicollinearity (Todeschini et al., 2004), which is rarely addressed or discussed in the recent models of technology acceptance. Second, the recently developed models fall short in providing strong theoretical reasons for at least one of the three effects: (a) the direct effect of the new determinant on perceived usefulness and ease of use; (b) the discrete and independent effect of the new determinant from every other determinant; (c) the interactive effect of multiple determinants on perceived usefulness and ease of use. Third, the recent advancements have not fully replaced TAM. New versions and extensions of TAM are being introduced without acknowledgment of TAM2 and TAM3 (e.g., Al Shamsi et al., 2022; Sagnier et al., 2020; Wang et al., 2020). Therefore, this study focuses on TAM's core variables and integrates CRT to account for the context of communicative interaction between humans and machines.

The communicative interaction is fundamental to understanding HMC as the human and the machine are building a communicative relationship (Guzman, 2018). In the context of HMC and HCI, TAM research has examined how users' experience and attitudes (e.g., Bröhl et al., 2016), and hedonic values (e.g., de Graff et al., 2019; Park & Kwon, 2016) about the artificial intelligence affect the perceived usefulness and ease of use. However, little is studied about how to systematically examine the socio-relational aspect of the communication between the human and the machine. Therefore, the integration of CRT can resolve at least one limitation of TAM. Although the compliance process can occur when a person sees oneself in relation to another person, agent, and group, TAM (on its own) does not take group, cultural, and social contexts into account (Bagozzi, 2007).

By anatomizing the communication process using CRT, this study provides three external factors that influence perceived ease of use and usefulness of a navigation technology: the context of the interaction (i.e., location), the driver's perception (i.e., communicative responsibility), and the technology's communicative behavior (e.g., how the directions are provided). As outlined in CRT, the driver's familiarity with the location and the perceived communicative responsibility affects the dynamics of the interaction. When driving in an unfamiliar location, the driver may be more dependent on the navigation system as the driver has limited knowledge about the directions. The driver may also rely heavily on the navigation system because the purpose of the navigation is to assist the driver by providing accurate directions. In both situations, the driver attributes a large amount of communicative responsibility to the navigation system and finds its interaction valuable for achieving the communicative goal of getting to the destination. Consequently, the driver is likely to consider the navigation system as easy to use and useful. Additionally, in certain situations (e.g., driving in an unfamiliar location) and to some drivers (i.e., drivers who expect the navigation system to provide simple and highly visible landmarks), how the direction is provided matters. When the navigation system provides the directions using general landmarks, such as parks, woods, and lakes, drivers may easily find their way as minimal a priori knowledge is required to notice these landmarks. Drivers may find the directions including location-specific landmarks, such as the name of a store that is less available in other places, less helpful and difficult to understand, especially when they consider the navigation to have more responsibility in helping the driver understand the directions. The following hypotheses are proposed to examine these three factors that may determine a driver's intention to use a navigation system:

> **H1:** Participants will perceive the navigation system to be (a) easy to use and (b) useful when they perceive the navigation system to have more communicative responsibility than themselves.

> **H2:** Participants will perceive the navigation system to be (a) easy to use and (b) useful when it uses general landmarks (i.e., natural landmarks) to give directions compared to location-specific landmarks (i.e., stores).

> **H3:** Participants will perceive the navigation system to be (a) easy to use and (b) useful when they are driving in an unfamiliar location compared to a familiar location.

> **H4:** The relationship between asymmetry in communicative responsibility (i.e., the driver having more) and the perceived (a) ease of use and (b) usefulness of the navigation system will be moderated by the types of landmark: the relationship will be more negative when the navigation uses general landmarks (i.e., natural landmarks) to give directions instead of location-specific landmarks (i.e., stores).

> **H5:** The relationship between asymmetry in communicative responsibility (i.e., the driver having more) and the perceived (a) ease of use and (b) usefulness of the navigation system will be moderated by the types of location: the relationship will be more negative when driving in an unfamiliar location instead of a familiar location.

> **H6:** Participants will have a higher intention to use the navigation system when they perceive it as (a) easy to use and (b) useful.

Method

Participants and Procedure

A total of 314 participants were recruited from SurveyMonkey's United States online panel in May 2020. Those who had a driver's license were eligible to participate. The data from 216 participants were used for the main analyses after removing those that did not pass the attention-check questions (n = 91) or did not answer most of the survey questions (n = 7). There were four attention-check questions to reassure the quality of the responses (Paas & Morren, 2018). To reassure that participants could hear the audio clip of the navigation system, the first question asked participants to play a provided audio clip (i.e., siren) and choose the sound they heard from a list: bark, car, horn, rain, and siren. Twenty-six participants incorrectly identified the audio clip. The second and third attention-check question filtered out participants who provided inconsistent and implausible answers: 3 participants' driving experience exceeded their age; 19 participants indicated that they have never used a GPS and a navigation system but stated that their existing GPS was useful. The last question (i.e., select "disagree") eliminated 43 participants who did not carefully read the questions or provided straight-line grid answers. Around half of the analyzed participants were women (n = 124, 57.41%). Their age ranged from 18 to 87 (M = 42.58, SD = 16.19). As for ethnicity, the majority of the participants (n = 150, 69.76%) were Caucasian. Hispanic or Latino (n = 30, 13.95%) and Asian or Asian American (n = 19, 8.83%) participants followed. African American (n = 10, 4.65%), Native American or Alaska Native (n = 4, 1.86%), and participants of other ethnicities (n = 2, 0.93%) also took part in this study. Although their driving experience varied (M = 23.10 years, SD = 16.66), it was not significantly related to the dependent variables.

The participants were randomly allocated to one of the four versions of the survey. The four versions were created by combining two locations (i.e., a hypothetical city in the US or South Korea) with two communicative behaviors of the navigation system (i.e., using location-specific landmarks or general landmarks). Fifty-three participants interacted with the navigation system that used location-specific landmarks in the United States; 46 interacted with the navigation system that used general landmarks in the United States. Fifty-one and 47 participants interacted with the navigation system in South Korea that used location-specific and general landmarks, respectively.

The participants were told that they were visiting a friend living in either a different state or a different country. They then listened to directions given by the navigation system while imagining themselves driving from a hotel parking lot to a restaurant to meet their friend for dinner. Using the default logic in the online survey platform, the relevant section of the map was shown to the participants while the navigation system provided directions. For every crossroads, participants chose a direction. This process continued until the participants arrived at the final destination. The entire map of the hypothetical cities are provided in Appendix A and B. The map was developed by making minor changes to the original study that tested CRT (Aune et al., 2005): the map of a South Korean city used culturally relevant streets (e.g., Songcheon-ro) and store names (e.g., Lotteria). In the location-specific landmark condition, the navigation system used names of specific stores and brands (e.g., Hyundai Department Store, Jack in the Box) while giving directions. In the general landmark condition, the navigation system provided directions with reference to the presence of parks, lakes, and woods instead of their names. After arriving at the destination, they completed the survey, which contained the following measures. The survey lasted around 9 minutes. Each participant received $4 for their contribution.

Measures

By utilizing the items developed by Aune et al. (2005), the perceived communicative responsibility of the participants and the navigation system were each measured with five items on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). "It is expected of me to make an extra effort to understand the directions given by this navigation system" and "compared to me, this navigation system has a much bigger responsibility for helping me understand how to get to the destination" are examples of items measuring the participants' and their perceptions of the navigation system's communicative responsibility, respectively (See Appendix C for all items). Both measures of communicative responsibility were reliable, Cronbach's alpha (α) = .84 and McDonald's omega (ω) = .86 for the driver; α = .86 and ω = .88 for the navigation system.

After taking the mean of the two communicative responsibility scores, the difference between the two averages was used to measure how the communicative responsibility is shared between the participants and the navigation. A score of 0 indicates that the participants perceived the communicative responsibility to be symmetrical and equally shared. A positive score indicates the participants attributed themselves to have a higher communicative responsibility, more responsibility in making sense out of the communication, than the navigation system. The difference score, which is referred to as communicative responsibility hereinafter, ranged from –4 to 3.6 (M = –0.07, SD = 1.34).

Perceived ease of use (e.g., "I think that learning to operate this navigation system is easy for me") and perceived usefulness (e.g., "overall, I find this navigation system to be very useful") were each measured by taking the mean of the three items on a 5-point Likert scale. The items were adopted from C. F. Chen and Chen (2011), Davis (1989), and Park et al. (2015). The participant's intention to use the navigation system was measured with three items on the same Likert scale (C. F. Chen & Chen, 2011; Venkatesh & Davis, 2000). All the measures were reliable, α = .78 and ω = .80 for perceived ease of use; α = .83 and ω = 0.85 for perceived usefulness; α = .89 and ω = .90 for intention to use.

Results

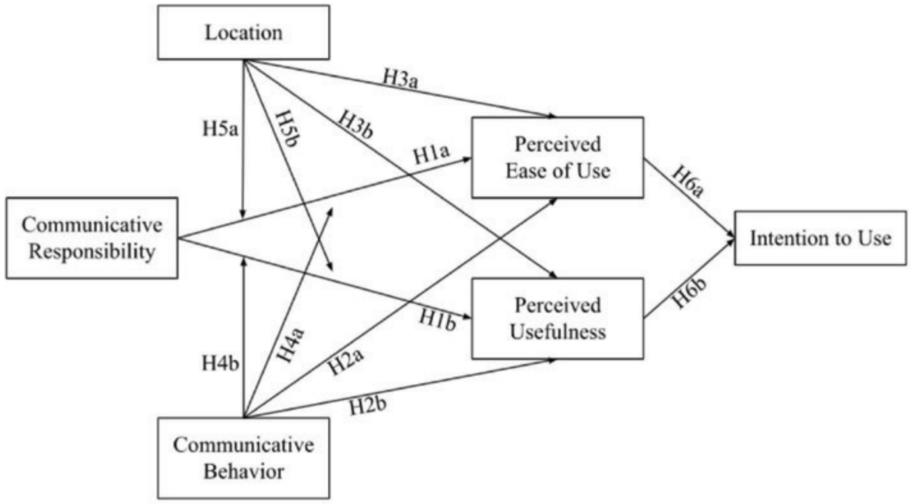
Manipulation Check

The two locations were chosen to reflect different levels of familiarity. The city in the US was intended to be a familiar environment compared to the city in South Korea. Participants were asked to rate their familiarity with the names of the stores, streets, and general landmarks. All three independent-sample t-tests indicated that participants were more familiar with the stores, streets, and general landmarks in the US city than in the South Korean city, t_{store} (214) = –7.01, p < .001; t_{street} (214) = –13.15, p < .001, and $t_{natural\ landmark}$ (214) = –5.26, p < .001.

Hypothesis Testing

RQ1 asked if communicative responsibility differed between the two locations. Considering the predictions of CRT, the communicative responsibility should be higher when driving in a familiar location. However, the participants might prioritize the role of the navigation system over the given context of driving. In this case, the communicative responsibility should be a negative score regardless of the location. The result of an independent t-test indicates that the communicative responsibility did not differ significantly based on the driving location, t (216) = –0.45, p = .66. The communicative responsibility for driving in a familiar location (M = –0.12, SD = 1.46) and an unfamiliar location (M = –0.03, SD = 1.22) was slightly negative but did not deviate significantly from zero, t (215) = –0.80, p = .42.

FIGURE 2 Hypothesized Mediated Moderation Effect



The six hypotheses predicting mediated moderation on behavioral intention were tested with model 10 of the process analysis (Hayes, 2018). According to the hypothesized model as illustrated in Figure 2, the communicative responsibility affects the perceived intention to use the navigation system. This relationship is mediated by perceived ease of use and usefulness. The relationships between communicative responsibility and the two predictors of TAM are moderated by the context of communication (i.e., location) and the communicative behavior of the navigation system. The location and communicative behavior were dummy coded with the familiar location and the location-specific landmarks as the reference group (See Table 1 for pairwise correlations of the variables). The process analysis provides the results of three OLS regressions predicting perceived ease of use, usefulness, and intention to use the navigation system (See Table 2). A nonparametric bootstrapping is conducted to analyze the direct and indirect effects of communicative responsibility on the usage intention (See Table 3).

TABLE 1 Pairwise Correlations of the Variables

	1	2	3	4	5
1. Location[1]					
2. Communicative Behavior[1]	0.09				
3. Communicative Responsibility	0.03	0.05			
4. Perceived Ease of Use	0.14*	0.12	−0.12		
5. Perceived Usefulness	0.19**	0.04	−0.15*	0.66***	
6. Intention to Use	0.19**	0.1	−0.12	0.46***	0.64***

Note: *$p < .05$. ** $p < .01$. ***$p < .001$.[1] Location and communicative behavior are dummy coded variables with the familiar location and location-specific landmarks as the reference group, respectively.

TABLE 2 Summary of Process Analysis

	Outcome Variable								
	Perceived Ease of Use			Perceived Usefulness			Intention to Use		
	B	*SE*	*t*	B	*SE*	*t*	B	*SE*	*t*
CR	−0.16	0.06	−2.52*	−0.17	0.06	−2.66*	−0.04	0.06	−0.70
CB	0.19	0.10	1.85	0.08	0.10	0.78	0.09	0.10	0.92
Location	0.18	0.10	1.81	0.27	0.10	2.56*	0.14	0.10	1.47
CR × CB	0.17	0.08	2.27*	0.20	0.08	2.60*	−0.06	0.07	−0.86
CR × Location	−0.01	0.08	−0.15	−0.07	0.08	−0.88	0.14	0.07	1.92
PEOU	–	–	–	–	–	–	0.07	0.08	0.82
PUSE	–	–	–	–	–	–	0.70	0.08	8.45**
	adjusted R^2 = .04 F (5, 210) = 3.15**			*adjusted* R^2 = .07 F (5, 210) = 4.22**			*adjusted* R^2 = .41 F (7, 208) = 22.96**		

Note: *p < .05. **p < .001. Location and CR are dummy coded variables with the familiar location and location-specific landmarks as the reference group, respectively. CR = Communicative Responsibility, CB = Communicative Behavior, PEOU = Perceived Ease of Use, PUSE = Perceived Usefulness.

TABLE 3 Conditional Indirect Effects of Communicative Responsibility on Intention to Use

Moderator		Effect	SE	Lower CI	Upper CI
Location	Communicative Behavior				
Mediator: Perceived Ease of Use					
Familiar Location	General Landmark	−0.01	0.02	−0.06	0.01
	Location-Specific Landmark	0.00	0.01	−0.01	0.03
Unfamiliar Location	General Landmark	−0.01	0.02	−0.07	0.02
	Location-Specific Landmark	0.00	0.01	−0.02	0.02
Mediator: Perceived Usefulness					
Familiar Location	General Landmark	−0.12	0.06	−0.26	−0.01
	Location-Specific Landmark	0.02	0.07	−0.11	0.18
Unfamiliar Location	General Landmark	−0.17	0.06	−0.31	−0.05
	Location-Specific Landmark	−0.03	0.04	−0.11	0.06

Note: CI = 95% bootstrapped confidence intervals (n = 5000).

The model predicting perceived ease of use was statistically significant, F (5, 210) = 3.15, $p < .01$, *adjusted* $R^2 = .04$. The communicative responsibility was negatively correlated to perceived ease of use, and this relationship is moderated by the communicative behavior of the navigation system ($B = 0.17$, $SE = 0.08$). Simple slope analysis was conducted by alternating the reference group for communicative behavior (See Figure 3A). The results indicate that the negative relationship between communicative responsibility and perceived ease of use is evident when the navigation system provided directions with location-specific landmarks ($B = -0.16$, $SE = 0.06$, $t = -2.52$, $p = .01$), while the relationship is not statistically significant when general landmarks were used ($B = 0.02$, $SE = 0.06$, $t = 0.24$, $p = .81$). Therefore, the data is consistent with H1a. Although the relationship between communicative responsibility and perceived ease of use differed based on the communicative behavior of the navigation system, the direction of the difference was opposite of H4a. When the navigation system was referencing location-specific landmarks, the participants who perceived the navigation system to have higher communicative responsibility found the navigation system to be easier to use.

The model predicting perceived usefulness of the navigation system was statistically significant, F (5, 210) = 4.22, $p < .01$, *adjusted* $R^2 = .07$. Consistent with H1b and H3b, the participants found the navigation system to be more useful when they attributed more communicative responsibility to it ($B = -0.17$, $SE = 0.06$, $t = -2.66$, $p < .01$) and when using it in an unfamiliar location ($B = 0.27$, $SE = 0.10$, $t = 2.56$, $p = .01$). The interaction between communicative responsibility and the communicative behavior of the navigation system was significant ($B = 0.20$, $SE = 0.08$, $t = 2.60$, $p = .01$). In contrast to the predicted pattern in H4b, the participants who perceived the navigation system to have more communicative responsibility found it to be more useful when location-specific landmarks were used (See Figure 3B). When the navigation system gave directions by using general landmarks, the participants' perceived communicative responsibility did not affect the perceived usefulness ($B = 0.03$, $SE = 0.06$, $t = 0.50$, $p = .62$).

The model with the intention to use the navigation system as the outcome variable was significant, F (7, 208) = 22.96, $p < .01$, *adjusted* $R^2 = .41$. The relationship between perceived ease of use and behavioral intention was not statistically significant ($B = 0.07$, $SE = 0.08$, $t = 0.82$, $p = .41$), which indicates that the data was inconsistent with H6a. Consistent with H6b, participants who perceived the navigation system to be useful had higher intention to use it ($B = 0.70$, $SE = 0.08$, $t = 8.45$, $p < .001$). The participant's communicative responsibility did not have statistically significant direct effects on the intention to use the navigation system. The results of conditional indirect effects indicate that the effect of communicative responsibility on intention was mediated by perceived usefulness, but not perceived ease of use. The mediated relationship of perceived usefulness was moderated by the communicative behavior of the navigation system. Regardless of the location, the participants who attributed more communicative responsibility to the navigation system were more likely to find it useful and showed greater intention to use it when the navigation system provided directions by referring to location-specific landmarks (See Figure 4 on page 218).

FIGURE 3A Simple Slope Analyses of Moderation Effect

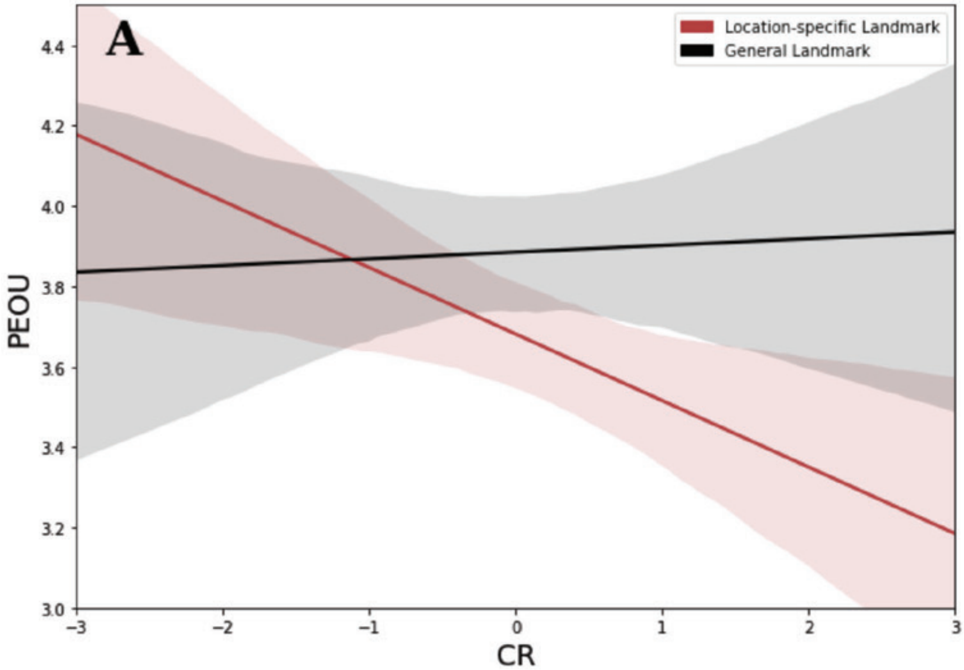
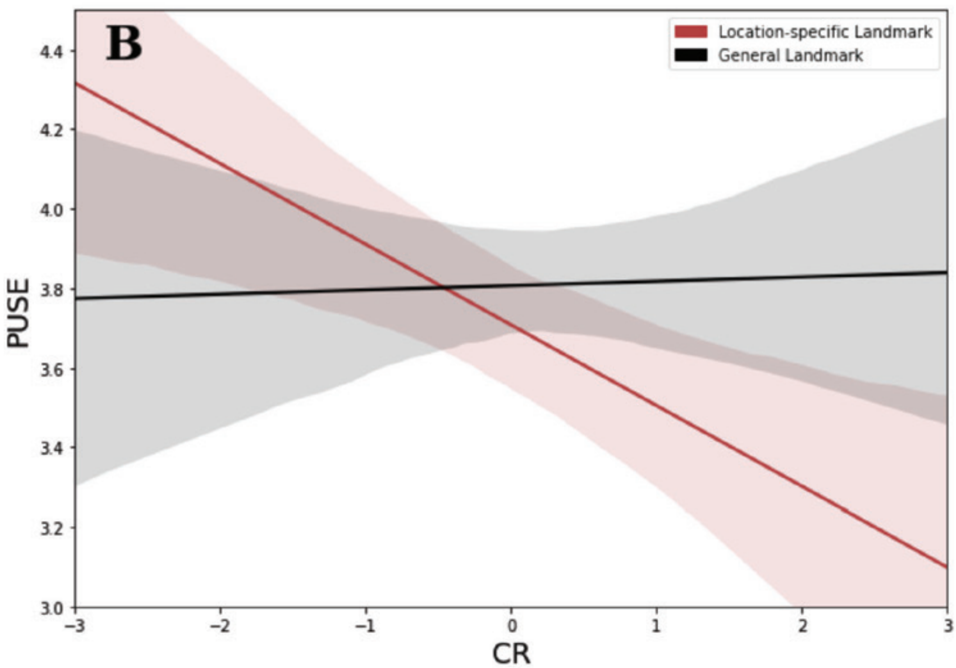
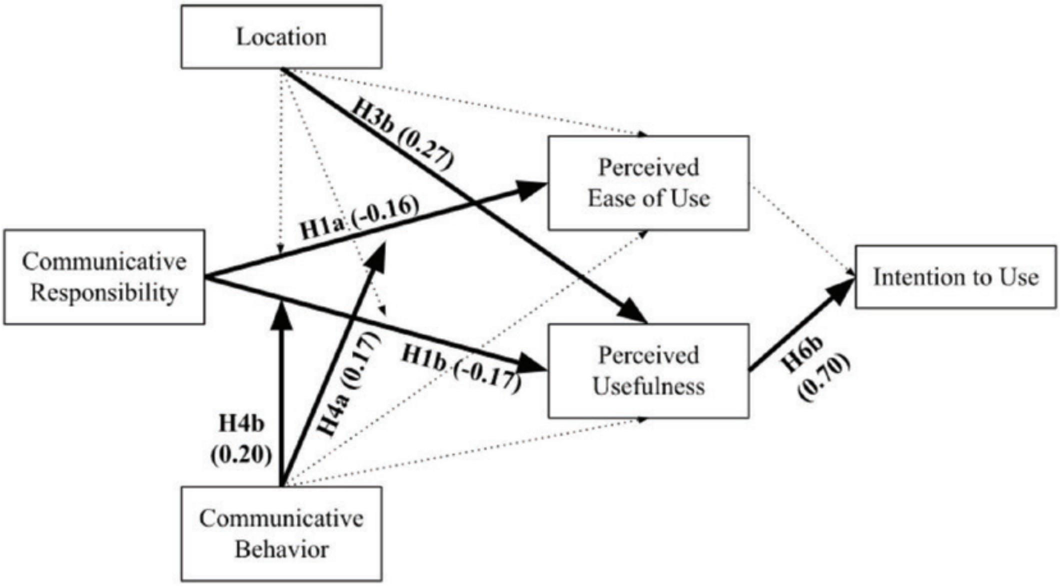


FIGURE 3B Simple Slope Analyses of Moderation Effect



Note: PEOU = Perceived Ease of Use, PUSE = Perceived Usefulness, CR = Communicative Responsibility (a positive score indicates that the driver had higher CR than the navigation system).

FIGURE 4 Analysis of Mediated Moderation Effect



Note: The statistically significant relationships are in bold. H4a and H4b were significant but opposite to the predicted patterns. The standardized coefficients are in parentheses.

Discussion

This study examines one of the central aspects of HMC: the process of meaning-making (Guzman, 2018). We examined the driver's intention to use a navigation system by integrating TAM and CRT, which includes the driver's perception of communicative responsibility, the communicative behavior of the navigation, and the context of interaction. Drivers were more likely to use the navigation system if they perceived it to be useful. The usefulness was influenced by the context (i.e., familiarity with the location) and the interaction between the driver's perception and how the navigation system provided the direction. While driving in a less familiar location, the drivers found the navigation system to be more useful. Additionally, when the navigation system provided location-specific landmarks (e.g., the name of a local store), the drivers who attributed the navigation system to have more communicative responsibility were likely to find the navigation system useful. When generic landmarks were used, the driver's perception was not significantly related to perceived usefulness.

When the direction provided by the navigation system referenced location-specific landmarks, the participants who perceived it to have higher communicative responsibility found it to be easier to use and more useful. This is inconsistent with the predictions derived from CRT. However, the findings may be consistent with other studies that examined the driver's intention to use a navigation system. C. C. Chen and Tsai (2019) found that completeness, informativeness, and accuracy of information provided by location-based services affect the perceived ease of use. Additionally, drivers found navigation systems that allow them to be more accurately aware of their location more useful (Park et al., 2015; Park & Kim, 2014). Service quality is an important factor affecting the driver's intention to use a navigation system (Park et al., 2015): providing accurate locational information is the main service of the navigation system. Therefore, location-specific landmarks, such as store names, are perceived to be more concrete and precise information as compared to generic landmarks, such as parks and lakes (without specifying the names of these venues). The drivers who perceived the navigation system to have more communicative responsibility in creating a mutual understanding to achieve the goal of getting to the destination are more likely to find the precise and accurate reference to landmarks as helpful.

The familiarity with the location is the other predictor of the drivers' perceived usefulness of the navigation system. Consistent with basic intuition, the drivers considered navigation to be useful when driving in an unfamiliar location. This implies that the user experience can be enhanced by integrating situational awareness to machines. According to Endsley (2000), situational awareness refers to knowing what is going around the agent. By considering the driving patterns and the navigation history, the navigation system can know whether the location is familiar or unfamiliar to the driver. As adoption of the navigation system is likely to be triggered in an unfamiliar driving location, the navigation system can further seek to improve its quality and provide a satisfying experience to the drivers to maximize their perceived usefulness.

This study contributes to the literature by providing and empirically validating a theoretical framework that can be used in HMC research: the framework focused on the process of meaning-making and the exchange of communicative message. Aune et al. (2005) first introduced CRT and also suggested how it can be applied beyond interpersonal communication. This study is the first to empirically test this. The theoretical framework is important because it can potentially provide guidelines for selecting specific social abilities that can be integrated into machines. According to Heerink et al. (2009), the social ability of the computer agent improves the user's interaction with it. Based on the findings of this study, estimation of communicative responsibility and execution of various communicative behaviors (as we have tested here) are social skills that could potentially improve the machine's interaction with the user. Additionally, providing the framework can encourage scholars to focus more on how meaning-making affects user's interaction with machines. Go et al.'s study (2020) proposed interactive technology acceptance model (iTAM) to study the machines that verbally interact with a user by examining the user's characteristics (e.g., self-efficacy and perceived enjoyment) and the machine's characteristics (e.g., type of AI robot and machine learning algorithm). Although the iTAM is created to understand machines that communicate with the users, it does not take the essence of interaction (i.e., meaning-making) into account. Therefore, CRT can potentially provide a theoretical framework that may further develop iTAM.

Although this study closely mirrors the original study that tested CRT by examining a navigation task (Aune et al., 2005), the findings from the two studies are not fully aligned. According to the theoretical reasoning and findings in Aune et al.'s study, we would expect participants to prefer general landmarks (i.e., reflecting less implicatures) when they believe that the navigation has more communicative responsibility and prefer location-specific landmarks (i.e., reflecting more implicatures) when they believe that they have more communicative responsibility. When the burden of meaning-making falls more to the navigation, it should use language that requires less inference-making. However, this study found that participants preferred location-specific landmarks to be used when they perceived the communicative responsibility to fall more on the machine. This may be because of the role and the function of a navigation system: an advanced driver assistance system that enhances driver comfort and convenience (Rahman et al., 2017). While the roles of an interactive machine are to retrieve information from a database and respond accordingly to the requested information (Go et al., 2020), people do not have the same obligation when giving directions to strangers they meet on the street. As machines are purposefully built to aid their users, the purpose may be more important than the context of communication. Additionally, HMC research suggests that people adjust their communication when talking with a machine. Gambino and Liu (2022) proposed that people use fewer complex words and sentences, and include more paraphrasing when talking to a machine compared to when talking to another person. Instead of studying the language of the users, this study examined the effects of the machine's language.

This study also contributes to TAM literature, especially regarding interactive technology. Continuous development and the variation in the extended versions indicates that TAM is useful and applicable to a wide range of technologies. However, it is also criticized for providing piecemeal knowledge. Even the recent renditions of TAM (e.g., Al Shamsi et al., 2022; Chocarro et al., 2021; Go et al., 2020; Sagnier et al., 2020; Wang et al., 2020) do not provide a coherent categorization and a solid theoretical framework for the precedents of perceived usefulness and ease of use. Consequently, the model may include factors that are no longer relevant or exclude factors that are crucial (Röcker, 2010). For instance, in Park et al.'s (2015) study of navigation system, they examined multiple external factors (e.g., perceived locational accuracy, satisfaction, perceived system reliability, and service quality) of TAM. They grouped these factors as the user's psychology. Following CRT, we recommend a trifurcation of external factors (i.e., user's perception, communicative behavior of the technology, and the context of interaction), which is highly relevant to technologies that are capable of engaging in communicative interaction. The results of our study demonstrates that the user's perceptions may interact with the communicative behavior of the navigation.

The salience of communicative responsibility and communicative behavior provides a roadmap to potentially improving interactive machines. Interactive machines include those that verbally interact with the user, retrieve information from a database, and respond to the user's request with accurate information (Go et al., 2020). This study suggests the importance of how to present the information to whom. For instance, a navigation system can utilize the user's data to determine if the user is likely to attribute more communicative responsibility to the navigation system. A user who frequently deviated from the recommended route and made unexpected detours is likely to expect the navigation to have a higher communicative responsibility. When this user has left their usual vicinity, the navigation may provide directions with location-specific information and additional information to clearly communicate the route. As for those users who have an aptitude for following the suggested route or finding faster alternatives, this kind of additional information and explication of directions may be less useful. Additionally, this also applies to other communicative machines. For instance, when using a voice-assistant reminder for managing schedules, a simple pop-up note with a gentle nudge may be enough for schedules that the user has created. However, for schedules that other users created (i.e., invitations to meetings), the voice assistant can call out the details of the schedule, such as the location, time, and a list of other attendees. Another example is a kiosk that is placed in restaurants and malls that assists people in ordering and purchasing products. Younger generations may be familiar with the kiosk. They will prefer a simple and fast interaction with the machine. However, for those users who are less familiar with such technology, the kiosk can additionally provide hands-on explanations of how to use the machine. This may mitigate the discomfort of using kiosks, which is more clearly evident in certain demographics (Na et al., 2021).

The current findings need to be interpreted with several limitations in mind. There are limitations to how the study was designed. The communicative responsibility was measured once at the end of the interaction. According to Kecskes and Zhang (2009), the evaluation of common ground is dynamic and constantly being updated as the communication progresses. People can adjust their communication by monitoring whether what they are saying violates the common ground or not (Horton & Keysar, 1996). Therefore, the communicative responsibility may change within the communication process. Future research can examine the dynamics of communication by using a cross-lagged panel model. Additionally, this study had participants imagine themselves driving while listening to a recording. The ecological validity of the study can be enhanced by utilizing driving-simulation games, such as the Truck Simulator series from SCS Software, and gaming steering wheel (e.g., Logitech G920 Driving Force Racing Wheel). Realistic games can provide an immersive task that is also easily controlled by the researchers. The more immersive experiment design may also introduce variance in the perceived ease of use because the difficulty of driving in the real world and in a simulated world is drastically different from imagination.

It is recommended for future researchers to further investigate the notion of common ground and the determinants of communicative responsibility. As shown in Figure 1, there are multiple communication contexts. This study investigated a specific niche: two different levels of common ground created in the context of driving. Technologies that are used for other purposes may bring additional dynamics to communicative responsibility. For instance, a user may attribute a large amount of communicative responsibility to health care chatbots, making the communication highly asymmetrical. In this case, variations in communicative behaviors should not focus on the degree of implicature but on how to enhance the user's inference-making experience.

Although mutual understanding is an essential aspect of CRT, this study has not examined to what degree the users perceived the communication to have fostered a mutual understanding. Considering how the user's perception of mutual understanding enhances the perceived utility of the machine and its performance (Stubbs et al., 2008), future research can either measure the degree of perceived mutual understanding or dive into examining its determinants and consequences. For instance, an explicated and redundant message may increase the perceived mutual understanding, but backfires in the efficiency of its performance to those that do not attribute much communicative responsibility to the machine. Additionally, nonverbal communication may affect the perceived level of mutual understanding (Alibali et al., 2013). For instance, the augmented head-up display (i.e., windshield augmented navigation) may be used only when the mutual understanding is substantially low, which could be determined by the level of frustration and confusion of the user. Therefore, instead of treating the augmented head-up display as all or nothing, the technology should estimate the level of common ground and visually show the route when the driver needs it, such as finding the exit of an unfamiliar road.

Author Biographies

Sungbin Youk (MA, Korea University) is a PhD student at University of California, Santa Barbara. His research focuses on message processing, persuasion, and media consumption. His research goal is to understand the ways in which media is consumed and to promote a more positive use of media in society.

https://orcid.org/0000-0001-8859-2757

Hee Sun Park (PhD, University of California, Santa Barbara) is Professor in the School of Media and Communication at Korea University, the Republic of Korea. Her current research projects examine cross-cultural differences in norms and interaction patterns, multilevel aspects of group and organizational communication, and health-related social influence processes and outcomes.

https://orcid.org/0000-0002-5389-8673

References

Ahlsén, E. (2008). Conversational implicature and communication impairment. In M. J. Ball, M. R. Perkins, N. Müller, & S. Howard (Eds.), *The handbook of clinical linguistics* (pp. 32–48). Blackwell Publishing.

Ahmed, M., & Shazali, M. (2010). The interpretation of implicature: A comparative study between implicature in linguistics and journalism. *Journal of Language Teaching and Research, 1*(1), 35–43. https://doi.org/10.4304/jltr.1.1.35-43

Al-Emran, M., & Granić, A. (2021). Is it still valid or outdated? A bibliometric analysis of the technology acceptance model and its applications from 2010 to 2020. In M. Al-Emran & K. Shaalan (Eds.), *Recent advances in technology acceptance models and theories* (pp. 1–12). Springer. https://doi.org/10.1007/978-3-030-64987-6_1

Alibali, M. W., Nathan, M. J., Church, R. B., Wolfgram, M. S., Kim, S., & Knuth, E. J. (2013). Teachers' gestures and speech in mathematics lessons: Forging common ground by resolving trouble spots. *ZDM Mathematics Education, 45*, 425–440. https://doi.org/10.1007/s11858-012-0476-0

Al Shamsi, J. H., Al-Emran, M., & Shaalan, K. (2022). Understanding key drivers affecting students' use of artificial intelligence-based voice assistants. *Education and Information Technologies, 27*, 8071–8091. https://doi.org/10.1007/s10639-022-10947-3

Aune, R. K., Levine, T. R., Park, H. S., Asada, K. J. K., & Banas, J. A. (2005). Tests of a theory of communicative responsibility. *Journal of Language and Social Psychology, 24*, 358–381. https://doi.org/10.1177/0261927X05281425

Bagozzi, R. P. (2007). The legacy of the technology acceptance model and a proposal for a paradigm shift. *Journal of the Association for Information Systems, 8*(4), article 7.

Bröhl, C., Nelles, J., Brandl, C., Mertens, A., & Schlick, C. (2016, July 17–22). TAM reloaded: A technology acceptance model for human-robot cooperation in production systems. In C. Stephannidis (Ed.), *Communications in computer and information science* (pp. 97–103). Springer. https://doi.org/10.1007/978-3-319-40548-3_16

Chai, J. Y., She, L., Fang, R., Ottarson, S., Littley, C., Liu, C., & Hanson, K. (2014). Collaborative effort toward common ground in situated human-robot dialogue. In G. Sagerer, M. Imai, T. Belpaeme, & A. Thomaz (Eds.), *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 33–40). Association for Computing Machinery.

Chen, C. C., & Tsai, J. L. (2019). Determinants of behavioral intention to use the Personalized Location-based Mobile Tourism Application: An empirical study by integrating TAM with ISSM. *Future Generation Computer Systems, 96*, 628–638. https://doi.org/10.1016/j.future.2017.02.028

Chen, C. F., & Chen, P. C. (2011). Applying TAM to travelers' usage intention of GPS devices. *Expert Systems with Application, 38*(5), 6217–6221. https://doi.org/10.1016/j.eswa.2010.11.047

Chocarro, R., Cortiñas, M., & Marcos-Matás, G. (2021). Teachers' attitudes towards chatbots in education: A technology acceptance model approach considering the effect of social language, bot proactiveness, and users' characteristics. *Educational Studies*, Advance online publication. https://doi.org/10.1080/03055698.2020.1850426

Clark, H. H. (1992). *Arenas of language use*. The University of Chicago Press.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition, 22*(1), 1–39. https://doi.org/10.1016/0010-0277(86)90010-7

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

de Graff, M. M. A., Allouch, S. B., & van Dijk, J. A. G. M. (2019). Why would I use this in my home? A model of domestic social robot acceptance. *Human–Computer Interaction, 34*(2), 115–173. https://doi.org/10.1080/07370024.2017.1312406

Diop, E. B., Zhao, S., & Duy, T. V. (2019). An extension of the technology acceptance model for understanding travelers' adoption of variable message signs. *PLoS One, 14*(4), article e0216007. https://doi.org/10.1371/journal.pone.0216007

Endsley, M. R. (2000). Theoretical underpinnings of situational awareness: A critical review. In M. R. Endsley & D. J. Garland (Eds.), *Situational awareness: Analysis and measurement* (pp. 1–31). CRC Press.

Eriksson, N., & Strandvik, P. (2009). Possible determinants affecting the use of mobile tourism services. *Communications in Computer and Information Science*, *48*, 61–73. https://doi.org/10.1007/978-3-642-05197-5_4

Fortunati, L., & Edwards, A. (2020). Opening space for theoretical, methodological, and empirical issues in human-machine communication. *Human-Machine Communication, 1*, 7–18. https://doi.org/10.30658/hmc.1.1

Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human–Computer Interaction, 30*(5), 401–432. https://doi.org/10.1080/07370024.2014.921494

Fussell, S. R., & Krauss, R. M. (1989). The effects of intended audience on message production and comprehension: Reference in a common ground framework. *Journal of Experimental Social Psychology, 23*, 203–219. https://doi.org/10.1016/0022-1031(89)90019-X

Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication, 1*, 71–85. https://doi.org/10.30658/hmc.1.5

Gambino, A., & Liu, B. (2022). Considering the context to build theory in HCI, HRI, and HMC: Explicating differences in processes of communication and socialization with social technologies. *Human-Machine Communication, 4*, 111–130. https://doi.org/10.30658/hmc.4.6

Geurts, B. (2019). Communication as commitment sharing: Speech acts, implicatures, common ground. *Theoretical Linguistics, 45*(1–2), 1–30. https://doi.org/10.1515/tl-2019-0001

Go, H., Kang, M., & Suh, S. C. (2020). Machine learning of robots in tourism and hospitality: Interactive technology acceptance model (iTAM)–cutting edge. *Tourism Review, 75*(4), 625–636. https://doi.org/10.1108/TR-02-2019-0062

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology, 50*(5), 2572–2593. https://doi.org/10.1111/bjet.12864

Grice, P. (1989). *Studies in the way of words*. Harvard University Press.

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

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Gilford Press.

Heerink, M., Kröse, B., Wielinga, B., & Evers, V. (2009). Measuring the influence of social abilities on acceptance of an interface robot and a screen agent by elderly users. In A. Blackwell (Ed.), *Computers XXIII celebrating people and technology* (pp. 430–440). BCS Learning and Development. https://doi.org/10.14236/ewic/HCI2009.54

Heimgärtner, R. (2013). Reflections on a model of culturally influenced human–computer interaction to cover cultural contexts in HCI design. *International Journal of Human-Computer Interaction, 29*(4), 205–219. https://doi.org/10.1080/10447318.2013.765761

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human-Computer Interaction, 19*, 151–181. https//doi.org/10.1080/07370024.2004.9667343

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91–117. https://doi.org/10.1016/0010-0277(96)81418-1

Kecskes, I., & Zhang, F. (2009). Activating, seeking, and creating common ground: A socio-cognitive approach. *Pragmatics and Cognition, 17*, 331–355. https://doi.org/10.1075/pc.17.2.06kec

Kiesler, S. (2005). Fostering common ground in human-robot interaction. In K. Kawamura & A. C. Schultz (Eds.), *IEEE international workshop on robot and human interactive communication* (pp. 729–723). IEEE. https://doi.org/10.1109/ROMAN.2005.1513866

Lau, I. Y.–M., Chiu, C.–Y., & Hong, Y.–Y. (2001). I know what you know: Assumptions about others' knowledge and their effects on message construction. *Social Cognition, 19*, 587–600. https://doi.org/10.1521/soco.19.6.587.20888

Lee, M. K., & Makatchey, M. (2009). How do people talk with a robot?: An analysis of human-robot dialogues in the real world. In D. R. Olsen (Ed.), *Proceeding in CHI Extended Abstracts on Human Factors in Computing Systems* (pp. 3769–3774). Association of Computing Machinery. https://doi.org/10.1145/1520340.1520569

Lee, Y., Kozar, K. A., & Larsen, K. R. (2003). The technology acceptance model: Past, present, and future. *Communications of the Association for Information Systems, 12*(1), article 50. https://doi.org/10.17705/1CAIS.01250

Legris, P., Ingham, J., & Collerette, P. (2003). Why do people use information technology? A critical review of the technology acceptance model. *Information & Management, 40*(3), 191–204. https://doi.org/10.1016/S0378-7206(01)00143-4

Levine, T. R. (2019). *Duped: Truth-default theory and the social science of lying and deception*. University Alabama Press.

Ma, Q., & Liu, L. (2004). The technology acceptance model: A meta-analysis of empirical findings. *Journal of Organizational and End User Computing, 16*, 59–72. https://doi.org/10.4018/joeuc.2004010104

Mahmood, R. (2015). A pragmatic analysis of imference as a collective term for implicature and inference. *International Journal on Studies in English Language and Literature, 3*(9), 67–71.

Marangunić, N., & Granić, A. (2015). Technology acceptance model: A literature review from 1986 to 2013. *Universal Access in the Information Society, 14*(1), 81–95. https://doi.org/10.1007/s10209-014-0348-1

Meyer, J., Miller, C., Hancock, P., de Visser, E. J., & Dorneich, M. (2016). Politeness in machine-human and human-human interaction. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 60*, 279–283. https://doi.org/10.1177/1541931213601064

Meyer, J. R. (2009). Effect of primary goal on secondary goal importance and message plan acceptability. *Communication Studies, 60*(5), 509–525. https://doi.org/10.1080/10510970903260343

Na, T. K., Yang, J. Y., & Lee, S. H. (2021). Determinants of behavioral intention of the use of self-order kiosks in fast-food restaurants: Focus on the moderating effect of difference age. *SAGE Open, 11*(3), 1–11. https://doi.org/10.1177/21582440211031907

Paas, L. J., & Morren, M. (2018). Please do not answer if you are reading this: Respondent attention in online panels. *Marketing Letters, 29*(1), 13–21. https://doi.org/10.1007/s11002-018-9448-7

Pantic, M., & Rothkrantz, L. J. (2003). Toward an affect-sensitive multimodal human-computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390. https://doi.org/10.1109/JPROC.2003.817122

Park, E., Kim, H., & Ohm, J. (2015). Understanding driver adoption of car navigation systems using the extended technology acceptance model. *Behaviour & Information Technology, 34*, 741–751. https://doi.org/10.1080/0144929X.2014.963672

Park, E., & Kim, K. J. (2014). Driver acceptance of car navigation systems: Integration of locational accuracy, processing speed, and service and display quality with technology acceptance model. *Personal and Ubiquitous Computing, 18*(3), 503–513. https://doi.org/10.1007/s00779-013-0670-2

Park, E., Kim, K. J., & del Pobil, A. P. (2013). An examination of psychological factors affecting drivers' perceptions and attitudes toward car navigation systems. *Lecture Notes in Electrical Engineering*, *215*, 555–562. https://doi.org/10.1007/978-94-007-5860-5_66

Park, E., & Kwon, S. J. (2016). The adoption of teaching assistant robots: A technology acceptance model approach. *Electronic Library and Information Systems, 50*(4), 354–366. https://doi.org/10.1108/PROG-02-2016-0017

Patterson, R. E., & Eggleston, R. G. (2018). Human–machine synergism in high-level cognitive functioning: The human component. *IEEE Transactions on Emerging Topics in Computational Intelligence, 2*(4), 249–257. https://doi.org/10.1109/TETCI.2018.2816584

Rahman, M. M., Lesch, M. F., Horrey, W. J., & Strawderman, L. (2017). Assessing the utility of TAM, TPB, and UTAUT for advanced driver assistance systems. *Accident Analysis & Prevention*, *108*, 361–373. https://doi.org/10.1016/j.aap.2017.09.011

Röcker, C. (2010). Why traditional technology acceptance models won't work for future information technologies? *International Journal of Information and Communication Engineering, 4*(5), 490–496.

Sagnier, C., Loup-Escande, E., Lourdeaux, D., Thouvenin, I., & Valléry, G. (2020). User acceptance of virtual reality: An extended technology acceptance model. *International Journal of Human–Computer Interaction, 36*(11), 993–1007. https://doi.org/10.1080/10447318.2019.1708612

Samara, A., Galway, L., Bond, R., & Wang, H. (2019). Affective state detection via facial expression analysis within a human–computer interaction context. *Journal of Ambient Intelligence and Humanized Computing*, *10*(6), 2175–2184. https://doi.org/10.1007/s12652-017-0636-8

Spence, P. R. (2019). Searching for questions, original thoughts, or advancing theory: Human-machine communication. *Computers in Human Behavior, 90*, 285–287. https://doi.org/10.1016/j.chb.2018.09.014

Stalnaker, R. C. (1978). Assertion. In P. Cole (Ed.), *Syntax and semantics 9: Pragmatics* (pp. 315–322). Academic Press.

Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction: A field study. *IEEE Intelligent Systems, 22*, 42–50. https://doi.org/10.1109/MIS.2007.21

Stubbs, K., Wettergreen, D., & Nourbakhsh, I. (2008). Using a robot proxy to create common ground in exploration tasks. In T. Fong, K. Dautenhahn, M. Scheutz, & Y. Demiris (Eds.), *3rd ACM/IEEE International Conference on Human-Robot Interaction* (pp. 375–382). Association for Computing Machinery. https://doi.org/10.1145/1349822.1349871

Ta, V., Griffith, C., Boatfield, C., Wang, X., Civitello, M., Bader, H., DeCero, E., & Loggarakis, A. (2020). User experiences of social support from companion chatbots in everyday contexts: Thematic analysis. *Journal of Medical Internet Research, 22*(3), article 16235. https://www.jmir.org/2020/3/e16235

Tao, D., Wang, T., Wang, T., Zhang, T., Zhang, X., & Qu, X. (2020). A systematic review and meta-analysis of user acceptance of consumer-oriented health information technologies. *Computers in Human Behavior, 104*, article 106147. https://doi.org/10.1016/j.chb.2019.09.023

Todeschini, R., Consonni, V., Mauri, A., & Pavan, M. (2004). Detecting "bad" regression models: Multicriteria fitness functions in regression analysis. *Analytica Chimica Acta, 515*(1), 199–208. https://doi.org/10.1016/j.aca.2003.12.010

Van Erp, J. B., & Toet, A. (2015). Social touch in human–computer interaction. *Frontiers in Digital Humanities, 2*, article 2. https://doi.org/10.3389/fdigh.2015.00002

Veletsianos, G., Scharber, C., & Doering, A. (2008). When sex, drugs, and violence enter the classroom: Conversations between adolescents and a female pedagogical agent. *Interacting with Computers, 20*(3), 292–301. https://doi.org/10.1016/j.intcom.2008.02.007

Venkatesh, V., & Bala, H. (2008). Technology acceptance model 3 and a research agenda on interventions. *Decision Sciences, 39*(2), 273–315. https://doi.org/10.1111/j.1540-5915.2008.00192.x

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186–204. https://doi.org/10.1287/mnsc.46.2.186.11926

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. https://doi.org/10.2307/30036540

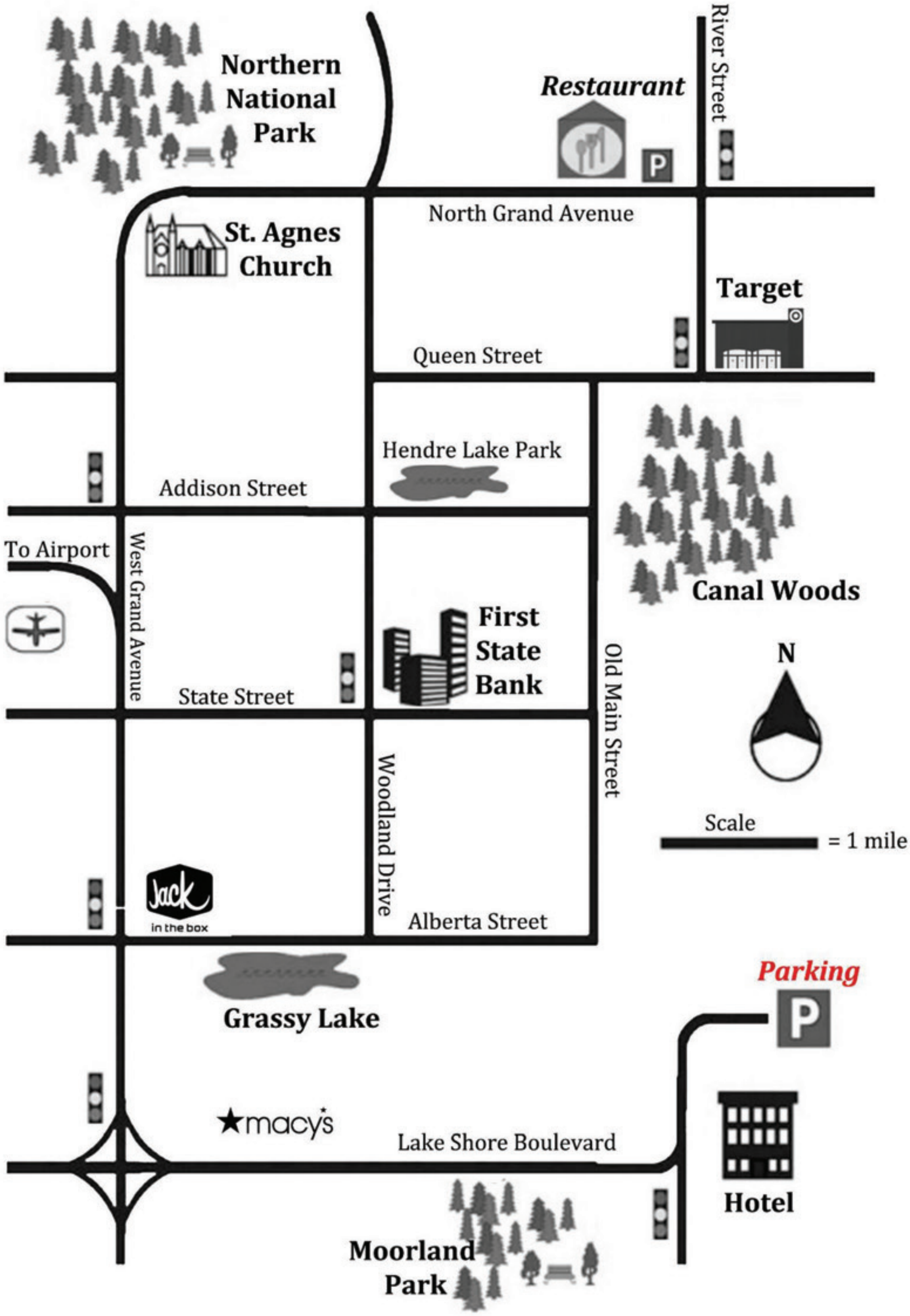
Venkatesh, V., Thong, J. Y., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, 36(1), 157–178. https://doi.org/10.2307/41410412

Wang, Y., Wang, S., Wang, J., Wei, J., & Wang, C. (2020). An empirical study of consumers' intention to use ride-sharing services: Using an extended technology acceptance model. *Transportation, 47*(1), 397–415. https://doi.org/10.1007/s11116-018-9893-4

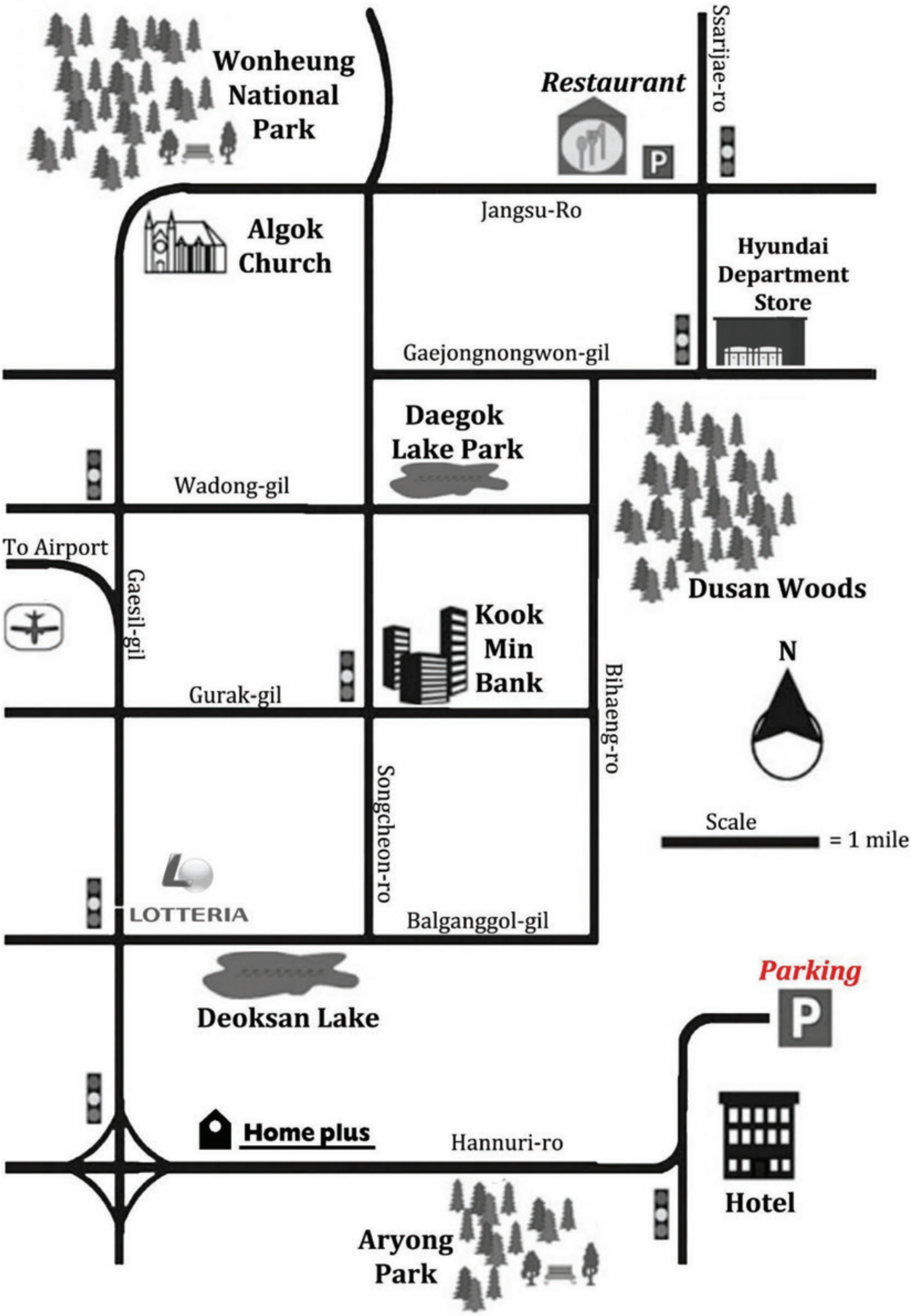
Waytz, A., Cacioppo, J., & Epley, N. (2010). Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science, 5*, 219–232. https://doi.org/10.1177/1745691610369336

Yucel, U. A., & Gulbahar, Y. (2013). Technology acceptance model: A review of the prior predictors. *Ankara University Journal of Faculty of Educational Sciences, 46*(1), 89–109. https://doi.org/10.1501/Egifak_0000001275

APPENDIX A The Overall Map of a Hypothetical City in USA



APPENDIX B The Overall Map of a Hypothetical City in South Korea



Appendix C Measurement Items

All the items below are measured with a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Items with an asterisk (*) are reverse coded.

Items for Manipulation Check

1.* The names of the natural landmarks (e.g., parks, lakes, woods) are unusual.
2. The names of the stores and brands are familiar to me.
3.* The names of the streets are unusual.

Items for Participant's Communicative Responsibility

1. Given this context, the responsibility for making sure that you understand the directions is mostly on you.
2. In this context, you have a much bigger responsibility for understanding how to get to the destination than the navigation system.
3. You are more responsible than the navigation system for making certain you understand the directions.
4. It is expected for you to make an extra effort to understand the directions.
5. It is appropriate, in this context, that you work harder to make certain that you understand the direction to the restaurant.

Items for the Navigation System's Communicative Responsibility

1. Given this context, the responsibility for making sure that you understand the directions is mostly on the navigation system.
2. Compared to you, the navigation system has a much bigger responsibility for helping you understand how to get to the destination.
3. The navigation system is more responsible than you for making certain you understand the directions.
4. It is expected for the navigation system to make an extra effort to help you understand the directions.
5. It is appropriate, in this context, that the navigation system work harder to make certain that you understand the direction to the restaurant.

Items for Perceived Ease of Use

1. I think that learning to operate the navigation system will be easy for me.
2. It will be easy to find my destination by using the navigation system.
3. Overall, I think that it is easy to use the navigation system.

Items for Perceived Usefulness

1. I believe that the navigation system can help me to save time.
2. I think that I can get the information about the destination using the navigation system.
3. Overall, I find the navigation system to be very useful.

Items for Behavioral Intention to Use the Navigation System

1. I will be willing to use the navigation system in the future.
2. I intend to use the navigation system in the near future.
3. I have it in my mind to use the navigation system in the future.