Methodological Middle Spaces: Addressing the Need for Methodological Innovation to Achieve Simultaneous Realism, Control, and Scalability in Experimental Studies of AI-Mediated Communication

ZHILA AGHAJARI, Lehigh University, USA
ERIC P. S. BAUMER, Lehigh University, USA
JESS HOHENSTEIN, Cornell University, USA
MALTE F. JUNG, Cornell University, USA
DOMINIC DIFRANZO, Lehigh University, USA

As AI-mediated communication (AI-MC) becomes more prevalent in everyday interactions, it becomes increasingly important to develop a rigorous understanding of its effects on interpersonal relationships and on society at large. Controlled experimental studies offer a key means of developing such an understanding, but various complexities make it difficult for experimental AI-MC research to simultaneously achieve the criteria of experimental realism, experimental control, and scalability. After outlining these methodological challenges, this paper offers the concept of methodological middle spaces as a means to address these challenges. This concept suggests that the key to simultaneously achieving all three of these criteria is to abandon the perfect attainment of any single criterion. This concept's utility is demonstrated via its use to guide the design of a platform for conducting text-based AI-MC experiments. Through a series of three example studies, the paper illustrates how the concept of methodological middle spaces can inform the design of specific experimental methods. Doing so enabled these studies to examine research questions that would have been either difficult or impossible to investigate using existing approaches. The paper concludes by describing how future research could similarly apply the concept of methodological middle spaces to expand methodological possibilities for AI-MC research in ways that enable contributions not currently possible.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI; Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: Artificial Intelligence (AI), Artificial Intelligence-Mediated Communication (AI-MC), Experimental research in AI-MC, Methodological Innovation

ACM Reference Format:
Zhila Aghajari, Eric P. S. Baumer, Jess Hohenstein, Malte F. Jung, and Dominic DiFranzo. 2023. Methodological Middle Spaces: Addressing the Need for Methodological Innovation to Achieve Simultaneous Realism, Control, and Scalability in Experimental Studies of AI-Mediated Communication. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 73 (April 2023), 28 pages. https://doi.org/10.1145/3579506

Authors' addresses: Zhila Aghajari, zha219@lehigh.edu, Lehigh University, 322 Building C, 113 Research Drive, Bethlehem, PA, USA; Eric P. S. Baumer, ericpsb@lehigh.edu, Lehigh University, 235 Building C, 113 Research Drive, Bethlehem, PA, USA; Jess Hohenstein, jch378@cornell.edu, Cornell University, 236 Gates Hall, Ithaca, NY, USA; Malte F. Jung, mfj28@cornell.edu, Cornell University, 206 Gates Hall, Ithaca, NY, USA; Dominic DiFranzo, djd219@lehigh.edu, Lehigh University, 328 Building C, 113 Research Drive, Bethlehem, PA, USA.

2573-0142/2023/4-ART73 $15.00
https://doi.org/10.1145/3579506

1 INTRODUCTION

Technologies mediate human communication. From modern multi-party telepresence, to text messaging, to television and radio, to even the technology of writing, rigorous research has produced important insights about how these various technologies shape and constrain different communicative interactions [31, 32, 62, 81, 85]. Hancock et al. [30] suggest the ways that AI (artificial intelligence) mediates human communication deserve similar rigorous research attention.

However, the AI in AI-MC (AI-Mediated Communication) acts as a source of methodological challenges to conducting controlled experimental studies. It is difficult to conduct experiments that simultaneously present participants with a believable experience (i.e., experimental realism) [21], allow researchers sufficient experimental control required to design experimental manipulations [11, 70], and can be scaled to enable large online social experiments [2]. Most prior experimental methods that are used to study AI-MC are able to address one or two of these challenges, but not all of them simultaneously.

For example, some past studies use vignette and screenshot methods where participants are told about or shown static images of an AI-MC system and are asked about their attitudes towards these technologies [e.g., 16, 91]. While these approaches allow for full experimental control and can be run at large scale, they lack experimental realism, in that they do not provide a believable proxy for people's interactions using an actual AI-MC system. Another common method is to use already-existing applications to conduct experimental research [e.g., 35, 36, 53], facilitating experimental realism. However, the lack of control over the design of these platforms limits researchers' ability to design and control the experimental conditions. A "Wizard of Oz" approach [40, 41], on the other hand, can provide both experimental control and experimental realism. However, its requirement of a human confederate severely limits scalability.

While they can produce valuable insights, the inherent limitations of the methods noted above constrain the scope of possibilities for rigorous experimental AI-MC research. Put differently, producing a rigorous, thorough understanding of AI-MC requires the development of new methodologies that provide *simultaneous realism, scalability, and control* in experimental studies of AI-MC.

Instead of perfecting one or two criteria of experimental research and sacrificing the third criterion (such as in the aforementioned approaches of experimental research in AI-MC), we believe the key to achieving all the required criteria of experimental research is abandoning the pursuit of perfecting any single criterion. Relaxing one criterion only slightly can open up space to achieve the other required criteria. This argument is particularly relevant to the focus of AI-MC research, where researchers aim to examine the *impacts* of AI-MC on individuals' interactions. For this goal, there is no need to replicate a real AI-MC system and thereby gain perfect "mundane" realism (i.e., perfect replication of real world settings) [6, 21, 54]. Rather, a believable proxy of these systems can be sufficient to encourage participants to take the experiments seriously and be engaged with its activities [63, 84]. Similar approaches, this paper argues, can also be applied to relaxing the strict criteria of experimental control and of scalability, as well. Doing so increases the validity of AI-MC studies by simultaneously: (1) ensuring that human participants engage in actual communication among one another with the belief that their communication is mediated by an AI system; (2) granting the researcher high levels of experimental control by allowing them to make choices about how the simulated system functions; and (3) enabling studies to be run at arbitrarily large scales. To describe this approach, we define the term *methodological middle spaces* to describe techniques that combine aspects of different existing methods in order to balance between these multiple but necessary criteria for experimental studies of AI-MC.

This paper provides a conceptual grounding for exploring methodological middle spaces for AI-MC studies. To do so, Section 2 reviews prior research on AI-MC and Section 3 articulates

the unique methodological challenges this phenomenon poses to experimental research. Next, Section 4 presents the approach of exploring methodological middle spaces as a way to address these challenges and advance experimental research in AI-MC. To demonstrate the utility of this approach, the paper describes Moshi, a platform for experimental studies that applies the concept of methodological middle spaces and thereby supports balancing among all three required criteria for conducting text-based AI-MC studies (Section 5.1). Through a series of studies conducted using Moshi, the paper describes how experimental design for methodological middle spaces manifests in practice (Section 6), highlighting how this approach helps to achieve all three criteria for experimental studies of text-based AI-MC. It demonstrates how Moshi helps answer research questions that would have been challenging, perhaps even impossible, to explore using existing methods. Next, the paper discusses the utility of exploring methodological middle spaces beyond text-based AI-MC to advance experimental studies in other forms of AI-MC (Section 7). The limitations of employing the concept of methodological middle spaces to design experimental AI-MC studies are also discussed. Thus, this paper does not offer a methodological contribution, in terms of providing a new method. Instead, it offers a conceptual contribution to how to design experimental studies for AI-MC research.

In summary, the contributions of this paper are twofold. The primary contribution of the paper is presenting the concept of exploring methodological middle spaces, and illustrating how this approach can result in achieving simultaneous realism, scalability, and control in experimental studies of AI-MC. The secondary contribution of the paper is offering an application of the methodological middle spaces concept and illustrating how this application allows researchers to explore concepts in text-based AI-MC.

2 BACKGROUND ON PRIOR AI-MEDIATED COMMUNICATION RESEARCH

2.1 AI-Mediated Communication

The field of computer-mediated communication (CMC) has explored how digital communication has transformed interpersonal communication in terms of trust, language choice, and the culture of discussion [31, 32, 80, 81, 85]. Advances in AI have lead to its increased integration into communication platforms. Hancock et al. [30] suggest this integration of AI into interpersonal communication has the potential to once again transform online communication. The inclusion of AI into interpersonal communication, where AI is used to modify, augment, or even generate communication, is referred to as AI-Mediated Communication (AI-MC).

One of the most prevalent forms of AI involvement in online interactions is its integration into text-based communication. We can see examples of text-based AI-MC ranging from relatively low to high amounts of intervention, such as auto-correct, text suggestions, grammar corrections, and auto-completion [1, 15, 93], as well as smart replies and auto-responses in instant messaging and e-mails [39]. AI can even go beyond text suggestions and generate content on behalf of a sender in synchronous communication [e.g., 75].

Since these systems can enhance people's communication skills (e.g., by improving their writing style or saving time in text production), they are already widely-used at scale. For instance, AI-generated smart replies in Gmail constitute 12% of sent messages, representing about 6.7 billion emails each day [69]. While previous research suggests that the involvement of AI is affecting text-based conversations [35], we know little about its precise impacts. The next section provides an overview of some of the most pressing research questions in text-based AI-MC, as well as the challenges and limitations of existing methods.

2.2 Research Questions in Text-Based AI-MC

AI-MC raises a diverse range of research questions which focus on how the involvement of AI into human-human communication changes people's communications and relationships [30]. Similar to CMC before it, the mediation in AI-MC is not neutral, but affects people and their communication in different ways [4, 30, 35, 38].

A line of research questions concerns how individuals perceive AI-mediation communication, and how their perceptions of the role of AI might influence their interpersonal perceptions. For example, Hohenstein and Jung [36] examine the impacts of smart replies on interpersonal perceptions and suggest that the presence of AI-generated smart replies increases perceived trust between human communicators. In another case, Jakesch et al. [38] show that mixing AI- and human-written profiles can lead to mistrust in accounts whose profiles were labeled as or suspected to be written by AI.

AI-MC can also affect the way people describe themselves and disclose their information [30]. Previous research suggests that the selective presentation provided in computer-mediated tools can influence the way individuals perceive their real-world selves, which can result in an identity shift [26]. This phenomenon could be even more nuanced in AI-MC, wherein AI can leverage personal information such as conversational history and contact information to offer different personalized self-presentations depending on who a person is interacting with and what platform they are using.

A growing amount of work is investigating the impacts of AI-MC on people's language choices [34, 35, 53]. For example, examining the impacts of "smart replies" in text messaging, Hohenstein and Jung [35] show that positive language of smart replies may result in using more positive language in a conversation [35]. In another case, Mieczkowski et al. [53] examine the impacts of smart replies and suggest that AI-generated language has the potential to undermine some dimensions of interpersonal perception, such as social attraction.

AI-MC also has the potential to impact relationship maintenance. The use of AI agents for things like automated birthday wishes and automated scheduling [75] can undermine perceptions of effort in relationships. Decreased perceptions of effort can in turn negatively influence individuals' relationships given that people consider the amount of effort their partners dedicate to their relationship when evaluating the quality of their relationships [81].

Only recently have the impacts of AI-MC in group dynamics been acknowledged [22, 42]. For instance, Kim et al. [42] show the efficacy of AI to facilitate group discussion via encouraging reluctant participants to contribute. In another case, Duan et al. [22] show in a group discussion between non-native and native English speakers, AI can intervene to ask clarification questions and effectively help nonnative English speakers follow the conversation and contribute to the discussion.

While researchers have already started to explore some of the myriad potential effects of AI mediating our communication, there is a great body of research questions that needs to be explored. However, AI's involvement and its associated complexities present specific methodological challenges to exploring AI-MC. The next section discusses some of these challenges. While this paper focuses on text-based AI-mediated communication, these methodological challenges apply to the broader area of AI-MC as well.

3 METHODOLOGICAL CHALLENGES IN EXPERIMENTAL AI-MC RESEARCH

As noted above, experimental studies of AI-MC introduce a variety of unique methodological challenges. These challenges can be described as arising from the need to satisfy three simultaneous requirements: experimental *realism*, such that participants feel and believe that they are engaged

in a realistic setting; experimental *control*, where the only variables that differ across conditions are the variables that researchers are experimentally manipulating; and the *scalability* of experimental protocols to be conducted with large numbers of participants. This section illustrates how most prior experimental methods used to study AI-MC are able to achieve only one or two of these criteria but not all three simultaneously.

3.1 Challenges for Experimental Realism

While experimental research must be conducted in settings that have a degree of realism [49], it is important to distinguish between different types of realism. Mundane realism involves "the extent to which events occurring in the research setting are likely to occur in the normal course of the subjects' lives, that is, in the 'real world'" [6, p. 485]. Often, researchers "assume that in order to *generalize* to 'real life,' the laboratory setting should *resemble* the real-life one as much as possible. [...] This assumption is false" [emphasis original 54, p. 385]. Arguably of greater importance is experimental realism [21], which occurs "if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has impact on them" [6, p. 485]. For example, for studies involving smart replies [e.g., 34], it is likely unnecessary for a laboratory environment to recreate the exact setting (physical, social, etc.) in which a smart reply would be suggested. That is, experimental AI-MC studies need not create mundane realism. Instead, it is more important that participants take the activity seriously (rather than focusing on the fact that they are participating in an experimental study [63]) and that they be involved (i.e., mentally engaged with and attentive to the study activities [84]).

Achieving each of these two aspects of experimental realism poses its own unique methodological challenges for AI-MC research. For example, researchers often investigate phenomena surrounding AI-MC using screenshots or vignettes. In such studies, researchers illustrate the AI mediator to participants either by describing how it works or by presenting one or more static images of it to the participants [16, 38, 91]).

This approach has been successfully used in prior CMC research [24, 85]. Static versions of communication media such as social media feeds [23, 29], online news articles [46], product reviews [89], profile pages [38, 76], and others can be replicated in ways that enable participants to treat them in much the same way they would treat the actual interface—skimming, scrolling, reading, etc. Screenshots can even be used for the few AI-MC applications that result in static content, such as profile generation [38]. However, the interactivity and dynamism of many AI applications are hard, perhaps even impossible, to replicate with such static screenshots. Whether consciously or subconsciously, most research study participants are acutely aware that they are participating in a research study [63]. Thus, while participants may participate in these kinds of studies in good faith, materials such as verbal descriptions, static screenshots, vignettes, and related techniques likely make it even more difficult for participants to suspend disbelief and to immerse themselves in the experience, i.e., to forget that they are part of an experiment.

Furthermore, such methods also prevent participants from becoming involved with the study materials. While participants may become mentally engaged with such materials, actual interaction is particularly important when conducting AI-MC studies for at least three reasons. First, researchers want to know not how participants imagine that they might interact via an AI-MC system but rather how they actually do interact. In some domains (e.g., privacy), there are significant differences among beliefs, attitudes, intentions, and actual behaviors [7, 59, 72]. While it may be interesting to study beliefs, attitudes, or intentions about AI-MC, static screenshot studies limit experimenters' ability to study actual behaviors. Second, static screenshots may exacerbate demand characteristics [63]. Focusing on these study materials, rather than on engaging in an actual interaction, may make it easier for participants to reason about the purpose of a study, even if only unconsciously. Third,

asking participants to imagine themselves interacting via an AI-MC system may introduce other biases. For instance, different individuals have widely varying perspectives on AI [17]. Such pre-existing perceptions can greatly differ from, and perhaps even overwhelm, the description of the AI mediator presented by researchers. Thus, studies of AI-MC based on screenshots may tell us more about each individual participant's views on AI than they do about how AI might mediate human communication.

To be clear, some prior methods for studying AI-MC do provide experimental realism. However, these methods often compromise on other requirements, as described below.

3.2 Challenges for Experimental Control

Experimental research must be conducted in a controlled setting where the only variables that differ across conditions are the variables that researchers are experimentally manipulating [11, 70]. The controlled setting allows researchers to explore the effect of experimental manipulation on the outcome while controlling for any extraneous variables that may have an effect on the outcomes. This setting ensures that the observed effects are based on solely the experimental manipulation and enhances internal validity of the findings [45]. Lack of control over the experimental setting, however, can introduce confounding variables that might influence the results. As a result, the researchers cannot ensure that the results are due to the experimental treatment.

However, ensuring experimental control poses methodological challenges for AI-MC research. For example, a common approach to experimental research in AI-MC is the use of the already-existing messaging platforms [35, 53]. Using these platforms eliminates the substantial up-front investment required to develop and launch AI-mediated messaging systems and offers a high degree of experimental realism. Already-existing messaging platforms have also been successfully used in prior CMC research [28, 74]. These already-existing platforms are particularly helpful to conduct virtual experiments as CMC is virtual by nature. Also, the virtual experiment can reduce the likelihood of experimenter bias and the experimenter error that might occur in the laboratory setting.

At the same time, the manifold nature of AI components in AI-MC requires a high degree of control that is not offered in the already existing applications. The lack of control over these platforms makes it challenging to investigate the manifold AI-MC concepts. That is, researchers do not have control over the underlying AI in the existing AI-MC systems. Therefore, it is not possible to modify features of the underlying AI and investigate how various aspects of the AI component influence people's interactions via the AI-MC system. For example, to investigate when and how people attribute the agency to AI in AI-MC, and how the attribution influences their interactions, it is essential to be able to modify the level at which the AI becomes involved in AI-MC.

Additionally, AI-MC platforms are designed to achieve a certain outcome for their parent companies, often a business outcome. Designs to enhance that outcome may align with, contradict, or be completely unrelated to research questions about how different interface features and presentation modalities influence outcome variables. Furthermore, it is difficult to control for the spurious effect of extraneous variables. For example, in a study that explores the effects of AI-MC on trust, participants' prior mindset about the parent companies of the applications that are used in the experiment can influence the results [65].

Furthermore, researchers do not have control over the dynamics of ever-changing commercial platforms. This lack of control over potential changes in commercial AI-MC systems makes replication of prior work challenging. This limitation prevents researchers from building upon prior work to explore the concepts of AI-MC. For example, in a recent experiment, Hohenstein and Jung [35] demonstrate that the use of Google Allo, a since-decommissioned platform that combines an AI assistant with instant messaging to create an AI-mediated messaging application,

can influence what and how people communicate. The researcher noticed that suggested responses with a positive sentiment were noticeably more common than suggested responses with a negative sentiment. They were interested to investigate whether this excess of positive compared to negative suggested responses was having a priming effect on conversational dynamics. However, conducting a follow up experiment on this work was virtually impossible given that the platform was closed by the parent company (i.e., Google) immediately after their first experiment.

Some methods for studying AI-MC do provide experimental control. However, such methods sacrifice either experimental realism (described above) or scalability (described below).

3.3 Challenges for Scalability

Scalability is another methodological challenge facing experimental AI-MC research. The social and interpersonal effects of an AI-MC system can be subtle and rendered hidden in a micro view [52]. Such effects, while still real and impactful, may only be seen by in a more macro view of a population [2]. This requires using methodologies that can quickly, easily and cheaply scale to thousands of participants. However, most methods for controlled laboratory experiments cannot scale in this way due both to physical and practical limitations.

For example, Hohenstein and Jung [35] brought participants into a physical laboratory space where participants could interact with an actual AI-MC system while the researchers recorded all conversational and screen interaction data. Doing so can capture large volumes of rich, detailed data about how the AI mediated participants' communication. At the same time, this approach requires the researcher to run each group of participants one at a time. Such methods generate volumes of rich data, but they also pose huge logistical challenges in terms of participant recruitment, scheduling, prepping and cleaning the physical experimental laboratory room, and staffing researchers to oversee and run each experiment. Furthermore, the commercial platform used by Hohenstein and Jung [35] required them to collect data via screen recordings then manually transcribe all text and smart replies for each participant. Running such studies with larger numbers of participants becomes highly impractical, if not entirely impossible.

These studies also risk their participant populations being that of convenience rather than representative as a whole. Laboratory studies that take place at universities often have an over representation of college aged participants, and they often under represent other populations. Increasing the scale of a study (i.e., the number of participants) also provides more opportunities to increase the representativeness of the participant sample.

The "Wizard of Oz" (WoZ) [40, 41] technique can also be used in AI-MC based research. This approach uses a human confederate, "the wizard," to control or act in place of an AI system, "Oz" [40, 41]. Doing so allows researchers to conduct usability experiments with systems that do not yet exist. WoZ also combines a higher degree of experimental control compared to a commercial system, because the human wizard has full control over the AI's functioning, with a high degree of experimental realism, because the participant believes that they are interacting via an actual functioning AI system. Indeed, WoZ has previously been used in some AI-MC studies [37, 57].

However, WoZ studies face scalability issues in terms of the human confederate, i.e., the "wizard." These types of studies are limited by the number of human confederates they have and the scheduling complexity in managing and connecting the confederates to participants. Thus, in many ways, WoZ studies face similar challenges. Furthermore, some tasks of an AI-MC system lie beyond the scope of of what a human can reasonably do under synchronous time constraints, such as providing labels describing a large text corpus, or providing recommendations based on a series of participant behaviors. Although some work has developed tools for making WoZ studies easier to run [e.g., 43] or for combining WoZ with other techniques [e.g., 20, 67], there remain basic constraints on what a human wizard can accomplish while maintaining experimental realism.

4 EXPLORING METHODOLOGICAL MIDDLE SPACES

The above review of methodological challenges in experimental AI-MC research demonstrates prior methods are able to address one or two of the required experimental criteria, but not all of them simultaneously. Therefore, there is a need for methodological innovation to achieve simultaneous experimental realism, experimental control, and scalability in studies of AI-MC. As a means to address these challenges, we suggest the concept of methodological middle spaces, as presented in the remainder of this section.

The notion of methodological middle spaces suggests that the key to achieving all the criteria of experimental research simultaneously is abandoning the pursuit of perfecting any single criterion. Indeed, complete achievement of any one criterion often requires significant sacrifices in terms of other criteria. Instead, we suggest slightly relaxing each of the criteria to obtain a balance among them. Put differently, the concept of methodological middle spaces suggests that exploring and innovating methods that make small concessions on each of these criteria can help us avoid sacrificing any criterion entirely.

The remainder of this section illustrates how the concept of methodological middle spaces might be applied to modify existing methods in ways that can simultaneously achieve multiple yet necessary experimental research criteria. For each method, this section identifies the specific criterion (or criteria) that the existing methods strives to meet perfectly. It explains why the perfection of that criterion (or criteria) acts as a bottleneck to achieving the other needed criteria in experimental research. Next, it discusses ways to modify these existing methods via relaxing the perfected criterion (or criteria) to a degree that provides the space for achieving the previously sacrificed criteria. Furthermore, it shows how the suggested way to relax the perfected criterion (or criteria) and thereby seek a balance between all the criteria of experimental research is relevant to the goals of AI-MC studies.

Studies using screenshots or vignettes, as discussed in Section 3, provide perfect experimental control, allowing researchers to manipulate and control the exact setting of what the subjects experience during the experiment. However, this perfect experimental control requires almost entirely sacrificing experimental realism, in that they do not provide interactivity and dynamics of AI-MC applications. The concept of methodological middle spaces suggests that by relaxing experimental control and adding some degree of dynamism to the screenshot studies, researchers can achieve the interactivity of AI-MC systems (i.e., experimental realism) while still maintaining the degree of experimental control required for the certain research questions. For instance, a screenshot of a social media news feed or chat conversation could easily be animated to give the impression of live, semi-synchronous interaction, thereby helping immerse participants in the experience [6, 21]. The degree to which researchers should trade off (e.g., whether the selection of content to show responds in any way to a participant's actions) depends on the research questions and the phenomena of interest that researchers are looking to examine.

Alternatively, the use of commercial AI-MC tools provides perfect "mundane" realism (i.e., perfect replication of real world settings) [6, 21, 54]. However, this perfect "mundane" realism comes at the cost of completely sacrificing experimental control. That is, the researchers have to use them as they are and cannot manipulate any aspects of these AI-MC tools. The methodological middle spaces concept suggests that relaxing perfect "mundane" realism to the degree required for experimental realism (i.e., providing a realistic setting to encourage participants to take the study activities seriously and be mentally engaged with them) [63, 84] is still sufficient for the goal of AI-MC. As described in Section 3.1, experimental research does not require to *resemble* the real-life (i.e., "mundane" realism) [54]. Rather, it is more important that the experiment is involving to the subjects, so that they take the experiment seriously (i.e., experimental realism) [21].

Pursuing experimental realism, rather than mundane realism, opens space for researchers to study interactions with possible systems that differ, either slightly or perhaps dramatically, from existing AI-MC tools. This compromise can get back the sacrificed experimental control in commercial AI-MC tools, allowing researchers to manipulate specific details about how the system functions and design different experimental conditions. Indeed, this balance is particularly relevant to AI-MC research. In AI-MC research, the goal is not to develop and advance AI-MC tools *per se*. Instead, the focus of AI-MC research is on examining the effects of AI-MC systems on people's interactions and relationships. To pursue this goal, there is no need to replicate a real AI-MC system (i.e., "mundane" realism). Rather, it is sufficient to provide participants with a realistic enough prototype of such an AI-MC system to encourage the participants to take the activity seriously and actively engage in the experiment (i.e., experimental realism) [6, 21].

In addition, a "Wizard of Oz" approach can be modified following the methodological middle spaces to achieve all the experimental criteria simultaneously. As explained in Section 3, the WoZ approach provides near-perfect experimental control, since the human confederate acting as the Wizard controls every aspect of the system's functioning. It also provides near-perfect experimental realism, since from the participant's perspective they are interacting with a fully functioning system. However, these human confederates are the the bottleneck that prevent achieving scalability. The approach of methodological middle spaces suggests that relaxing the experimental control and experimental realism via modifying the wizard's role can provide the ground for achieving scalability, while still maintaining experimental control and experimental realism to the degree required for experimental AI-MC research. As in the above examples, these changes could be made in varying degrees. For instance, semi-automated tools similar to Suede [43] or Quasi [67] could be extended in ways that enable a single human wizard to manage multiple simultaneous studies. More advanced tools might enable asynchronous monitoring, requiring the wizard to check in only intermittently. Going further, instead of a human confederate, the wizard could be replaced by an automated bot. Indeed, some work has already been done on using bots as confederates [e.g., 42, 44]. The use of a bot as a wizard is unlikely to offer the same level of experimental control and experimental realism that a human confederate can. However, a wizard bot may be able to achieve these two criteria to the degree required for some, perhaps many, AI-MC experiments, while also providing the groundwork for achieving scalability. Again, the exact trade-off among experimental control, experimental realism, and scalability must be determined by the research question and the required degree of experimental control.

This description of methodological middle spaces is summarized in Figure 1. It shows how most prior methods used to study AI-MC achieve two criteria for experimental research nearly perfectly while sacrificing the third criterion almost entirely. As argued above, relaxing these criteria only slightly can help recover significant amounts of the sacrificed criterion. Since doing so in practice may be non-trivial, the following section describes a platform designed to support experiments that use methodological middle spaces. The subsequent section moves from abstract descriptions to concrete examples of three different studies that used methodological middle spaces to conduct AI-MC experiments.

5 DEMONSTRATING AN APPLICATION OF METHODOLOGICAL MIDDLE SPACES

The above section introduces our notion of methodological middle spaces. This section offers a description of one platform we developed, Moshi, that inhabits a particular methodological middle space. This platform is not a single point but rather offers a space of possible balances and tradeoffs among different experimental criteria. Moshi is an AI-MC text-messaging platform designed for conducting controlled experiments. Moshi takes its design inspiration from commercial text-based AI-MC platforms. We named our platform Moshi, as "Moshi Moshi" means hello in Japanese when

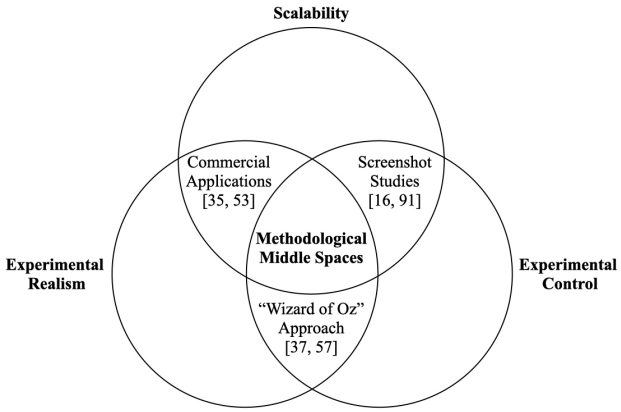

Fig. 1. Prior experimental methods to study AI-MC (e.g., screenshot studies [16, 91], commercial applications [35, 53], and the "Wizard of Oz" approach [37, 57]) are able to achieve at most two criteria of experimental research. The methodological middle spaces that this paper proposes exploring can achieve all these experimental criteria simultaneously.

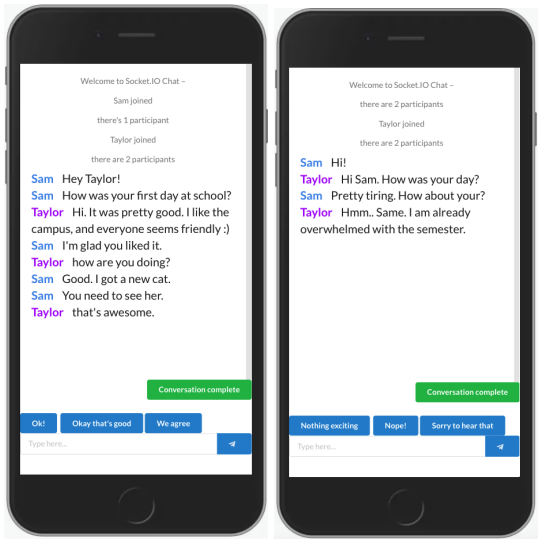

Fig. 2. The figure shows how Moshi enables to intentionally manipulate the content of smart replies that participants observe, to investigate the effects of the smart replies' linguistic characteristics on users' language (Discussed in Section 6.2). The figure on the left shows a version with positive sentiment smart replies, and the figure on the right shows a version with negative sentiment smart replies. To be clear, this is just one kind of manipulation that Moshi enables researchers to do.

answering the phone. This section illustrates how Moshi addresses the methodological challenges of conducting text-based experimental AI-MC research.

5.1 The Moshi Platform System Design

Moshi is a web-based research platform that allows researchers to engage online participants in text-based, real-time interpersonal communication. Moshi runs on all major modern browsers (i.e., Google Chrome 60, Mozilla Firefox 54, Microsoft Edge 14 and Apple Safari 10) and does not require participants to install anything. The interface is reactive to device type and resizes itself to work well on desktops, tablets, and mobile devices (e.g., Android and iOS).

In addition to the standard text box to send messages, participants can also receive smart replies that they can click or tap to send automatically, such as shown in Figure 2. Similar to existing chat apps, users can also scroll to see the history of the conversation at any point.

Moshi provides a modular scaffold for experiments on text based AI-MC, allowing other researchers to modify it and explore their own research questions. The following example illustrates different aspects of Moshi that can be defined and modified to examine different research questions.

Automated machine translation of text is one example of AI-MC that could be studied using Moshi. For example, two participants are told that they are communicating with someone that does not speak their native language (e.g., an English speaker writing to someone who speaks Chinese). As they write to their partner in English, the system translates this into Chinese and sends this translated message. When their partner writes back to them in Chinese, the system translates this into English. Using Moshi, we could design a study where each partner is actually writing English, and the Chinese translation they see is not real or accurate. They are in fact writing back and forth to each other in English and are lead to believe that not only are their messages being translated to and from Chinese, but that their partner is a Chinese speaker. Such a design would enable researchers to explore how people interact with each other when mediated by language translation systems and how the design of the system and their perceptions of it affects their interactions with others.

The remainder of this section uses this machine translation example to illustrate the various features of Moshi. We describe how each feature helps enable one or more of the three required criteria for experimental research (realism, control, and scalability) without sacrificing the others.

5.1.1 Context and Task. The context in which AI-MC is used and the kind of goals it is used to accomplish matter. Hancock et al. [30] discussed that the contexts of use matters in the way people accept AI-MC. "For example, AI-mediation is widely accepted when used to improve clarity, like auto-correct or machine translation between languages in text-based communication" [92]. Moshi allows researchers to control and change the *contexts* in which, as well as the *tasks* for which, AI-MC is used. Doing so enables researchers to investigate where and why people accept AI-MC differently in these different contexts.

For example, in our machine translation example, the context for use of AI-MC is clarity of communication. That is, the system is translating the written chat to help participants communicate more clearly. But what if the AI system was presented to the participants not just as a way to communicate better, but to present themselves in a better light? What if the platform informed users that the translation would not just be a direct translation, but would rewrite the messages to make the participant seem more like a native speaker of the language, or appear more intelligent and well read? This new context is not about communicating more clearly, but about presenting the participant as someone different than they really are, enabling testing of different hypotheses about how such systems might impact interpersonal communication. Note that the Moshi system itself need not change to test these different hypotheses, only the experimental set up and instructions given to participants.

The task is another attribute that can be controlled and changed in Moshi. Participants can be asked to use the chat for completing an explicit task assigned to them, e.g., negotiating the price of

a set of goods, solving a word puzzle together, or playing an economic game such as the prisoner's dilemma. The Moshi platform can help instruct, guide, show, and enable this task as well. For example, Moshi could be extended to provide an actual word puzzle interface that the participants could use as they chat with one another. Thus, Moshi's provisions for selecting task and context provide researchers a high degree of experimental control over the setting in which participants communicate.

5.1.2 Aesthetics. Moshi was designed to resemble other modern chat applications, like Facebook Messenger, Google Chat, and Apple's iMessage. However, it does not look identical to any of these as to avoid any prior preconceptions or mental models participants might have about these applications or the companies that created them [35, 53]. That said, the *aesthetics* of Moshi can be easily modified to look identical to any of these applications, or any chatting application if that is useful for the study design. The look and feel of the chatting application found in Moshi is completely up to the study designers. Currently, this level of experimental control is only available using screenshot studies. Since Moshi enables participants to actually engage in communicating with one another, it provides such control without sacrificing experimental realism.

5.1.3 Mediation. AI can take various actions to mediate human communication, such as auto-correction [1], smart replies suggestion [27], text-generation on behalf of a sender [38], encouraging group members to contribute in a group discussion [42].

In the language translation study example above, the *mediation* action is the auto language translation. Moshi allows for the presentation and control of this mediation. In this example, the language translation is not a real system, but a random Chinese text generator. Since the participants do not know Chinese themselves, they are lead to believe the AI system is accurately translating their messages. Due to this "fake" language translation system, the researchers have full control of this system and can focus their design on exploring how language translations systems might influence participants conversations. and their perceptions of their partners.

Of course Moshi is not limited to this example. It can provide believable AI mediations in many different contexts. For instance, Moshi could be used to conduct experiments about the impacts of existing auto-correct feature [1], or auto-complete feature in text-messaging and investigate whether each of these mediation actions has any impacts on language production. For example, prior work suggests people presented with smart replies shortcuts that were skewed positively were more likely to write more positive reviews online [4]. Using Moshi, researchers can test the link between the sentiment of text suggestion and language production in the contexts of real-time text-based conversation. (Section 6.2 illustrates how Moshi enabled us to explore this research question).

The experimental control that Moshi provides allows the investigation of more complex mediation actions than modern systems can actually do. For example, it allows researchers to examine the impacts of influencing messages for specific interpersonal outcomes such as appearing more trustworthy [38] or achieving the most desirable tone [15].

Furthermore, in addition to one-on-one discussion, Moshi can be used to investigate AI's involvement in complex group conversations, which is a topic of active exploration [42, 42]. For instance, AI can also be used to facilitate multiparty collaboration between native speakers of a language and nonnative speakers. In one case, Kim et al. [42] developed a chatbot that ask clarification questions to help nonnative English speakers follow the conversation and contribute to the discussion. Moshi can provide the experimental control required to investigate how such AI mediation might impact group dynamics.

Additionally, the mediation actions that AI can take in AI-MC is not confined to an action that takes place in each step of the conversation (e.g., suggesting text, text-generation, auto-correct,

suggest and modify language tone, etc.). The AI system can be designed to analyze inputs such as human-authored messages, communication history, and personal information, and take mediation actions such as providing feedback about how the conversation is progressing, and suggest ways to improve the conversation. For example, the AI can monitor the conversation and provide how each participant is contributing to the discussion, or encourage reluctant participants to contribute [42]. Moshi could be used to explore the impacts of such interventions. For example, an experiment using Moshi could explore the efficacy of AI's intervention in motivating more contributions in group discussions and how such interventions can impact subsequent behaviors of the receiver of this intervention.

With more control over the mediation actions of a seemingly real AI system, Moshi enables the exploration of interventions that may impact language production, interpersonal perception, collaboration dynamics, and task performance (both in one-on-one conversations and group discussions), while at the same time facilitating both experimental realism and scalability.

5.1.4 Timing. Moshi enables researchers to control and investigate the *timing* mechanism of the AI-mediation. For instance, in the example of automated machine translation, the researchers can control whether or not the translation action takes place simultaneously. The researchers can design the simulation to pose certain delays in the translation process and investigate how the delay might impact participants conversations.

Additionally, the AI-mediation action can take place either at every step of the conversation, upon sending or receiving messages, or it can be triggered at certain times of the conversation. For example, an AI agent that aims to mediate toxic conversations might intervene only when an instance of toxic language occurs [94]. Similarly, in the case of AI agent that aims to facilitate conversations between native and non-native English speaker via asking clarification questions, the AI agent might only intervene when it detects an expression that requires more clarification. However, in cases such as suggesting text, or auto-correct, the AI takes action at each step of the conversation. Moshi allows the researchers to control when the AI should take action.

5.1.5 Content. Moshi allows researchers to manipulate the *content* that an AI-MC system interjects into human communication. Section 5.1.3 above describes the various kinds of actions that an AI might take in mediating human communication. For some of those actions, the AI system presents content to the human user.

In the machine translation example explained above, Moshi allows researchers to control the content that the machine translation is generating. In this example, where the machine translation is not an actual translator, the content is generated based on some random Chinese words. The researchers have control over the length of the displayed Chinese words. Also, Moshi allows the researchers to manipulate different aspects of the text that the participants see, and investigate how it can impact their interpersonal communication.

In the case of Moshi, that content consists of smart replies. With a commercial production system (such as Allo), controlling the content of those replies would be difficult, perhaps even impossible. With Moshi, on the other hand, researchers have full control over the exact content in those smart replies.

This control provides two important benefits. First, it ensures experimental control (see Section 3.2). That is, researchers can be certain that each study participant is exposed to the exact same smart replies under the exact same situation. Second, alternatively, this control can provide researchers with the ability to test hypotheses about the content of smart replies by experimentally manipulating that content. For instance, a researcher may have questions about the influence of sentiment (positive, neutral, or negative) in smart replies. Does the sentiment in each smart reply influence whether the participant uses that smart reply? Does that effect vary with the relative

intensity of positive or negative sentiment? Might the aggregate sentiment across all shown smart replies influence how long a participant spends selecting a smart reply, thus indicating differing levels of cognitive effort? Does the relative proportion of each sentiment across the shown replies (e.g., one positive and two negative vs. two positive and one negative) influence the sentiment valence of participants' subsequent statements in the chat? These offer but a few examples all related to sentiment. Similar research questions could be examined around, e.g., the use of pronouns, politeness, passive vs. active verb constructions, or a host of other attributes. By allowing researchers to specify these detailed attributes of smart replies, Moshi provides a high degree of experimental control while still maintaining both experimental realism and scalability.

Furthermore, this approach could easily extend to other kinds of content. For example, very specific changes to profiles—that participants believe have been automatically generated but that are actually fabricated according to researcher-specified templates—can provide researchers the opportunity to test hypotheses related to identity, self-presentation, or even attitudes toward AI systems. Carefully crafted alterations to the output of computer translation, such as the level of politeness or the circumstances under which the system admits its inability to provide a meaningful translation, might be used to test hypotheses about how AI might mediate cross-cultural collaboration. Researchers could even manipulate the corrections provided by an auto-correct system to determine subconscious influences on a user's writing or other behavior. in any of these cases, the researcher could also elect to keep the content provided by the AI system consistent across participants. This level of control thus enables researchers to change the single independent variable they wish to manipulate while simultaneously leaving all other attributes and behaviors of the AI system constant.

5.1.6 Symmetry. Prior work has noted the design possibilities raised by explicitly manipulating the symmetry or asymmetry of media spaces [82]. By allowing for the intentional manipulation of asymmetry, Moshi expands the space of possible hypotheses that can be tested.

Our running example of machine translation as a mediating AI technology illustrates some of these possibilities. For example, the initial set up described above had two participants, both of whom speak English as their first language, communicating by being told that their messages are being translated into Chinese. Given the constructed nature of this environment, there is little reason that both participants need to be subject to the same manipulations. At the most simple, participant Alice could see their messages "translated" into Chinese and see Chinese messages from participant Bob "translated" into English, while Bob simply sees all messages in English. Doing so would enable researchers to test hypotheses about how the involvement of AI might affect perceptions about their interactions. For example, how two different participants perceive the same conversation based on whether or not each participant believes that machine translation is mediating that interaction.

Moshi allows researchers to apply this intentional manipulation of asymmetry to any of the dimensions described above. For example, the content of auto-replies could differ, such as providing one participant with replies of a more positive emotional valence. For another example, the AI mediation could have different timing for each participant, either in terms of speed, or in terms of when during an interaction the mediation takes place (e.g., constantly vs. periodically).

5.1.7 Disclosure. Another concern with AI-MC systems is whether and when the existence of AI-mediation should be disclosed [30, 38]. Moshi allows for the intentional manipulation of disclosure, and enables to test different hypotheses around efficacy of disclosure of AI's involvement in different contexts.

In the machine translation example, the mediation (i.e., translation from English to Chinese and vice versa) is disclosed to both of the participants. While both the participants speak English as

their first language, they are being told their messages are being translated into Chinese. However, the disclosure of the involvement of AI might influence the receiver to interpret the messages as not completely genuine as the result of AI's involvement [38]. Moshi allows to test this hypothesis. Participant Alice could be told about the involvement of AI and that their messages "translated" into Chinese and see Chinese messages from participant Bob "translated" into English, while Bob being told that he sees all messages in English. This setting allows to test whether or not two participant who use the same system perceive their conversation differently based on the information about the existence of the machine translation. In another case where the AI-MC occurs using an actual auto translation system (e.g., Google Translate), disclosure of the AI might help the receiver interpret the error-ridden texts as a result of imperfect AI and not the writers' intent [30].

In addition to investigating different dimensions for the disclosure of the use of an AI-MC system, Moshi allows researchers to investigate whether, when and how to disclosure more specific information about the AI's objective function. For example, in the use of AI as a moderator for group discussion, some members in the group discussion could be told the AI is a fair system and makes fair decisions with the probability of 95%, while the other members of the group simply do not receive this information. Doing so would allow researchers to test whether two participants who use the same system perceive the fairness of the system differently based on the information they receive about the AI's fairness. Researchers can also investigate how the participants perceptions of the AI's fairness might influence their language production as well as their contribution to the group discussion.

6 EXPERIMENTAL DESIGN FOR METHODOLOGICAL MIDDLE SPACES: THREE EXAMPLES

This section provides specific examples of how the concept of methodological middle spaces has been applied to experimental studies of AI-MC. It uses a strategy from prior conceptual work [e.g., 14, 64], wherein an abstract concept is clarified via its application to a number of concrete examples. Similarly, we illustrate how the concept of methodological middle spaces can be used to inform the design of specific experimental methods. To do so, we describe three studies, all of which used the Moshi platform described above to design experiments in text-based AI-MC to address the challenge of achieving simultaneously experimental realism, experimental control, and scalability (Discussed in Section 3).

Each example begins with describing a research question in AI-MC that was either difficult or impossible to examine using existing approaches. Since this section does not focus on the findings and implications of these studies, it omits detailed descriptions of the data analysis and results (for those details, see [anonymized]). Instead, it focuses on the methods, outlining the specific methodological challenges that researchers would face using prior approaches to investigate that research question. In particular, it shows how existing methods would perfectly achieve one or even two experimental criteria while entirely sacrificing the other(s). Next, the example demonstrates how slightly relaxing one or two criteria, as suggested by the concept of methodological middle spaces and implemented via the Moshi platform, provides significant benefits in terms of addressing the previously sacrificed criterion. Thus, this section shows how applying methodological spaces opens up the possibility of AI-MC studies that simultaneously achieve all three criteria of experimental research.

6.1 Example 1: Examining the Effects of Commercially-Available Smart Replies on Language and Interpersonal Perceptions

6.1.1 Motivation. Given that smart replies are already in use in various email and chat applications (e.g., Gmail, LinkedIn), it is important to understand the effects of real, commercially-available

smart replies on language production and interpersonal perceptions [30, 35]. One recent study gives insights that AI-mediated conversations are linguistically different than those without AI mediation [35]. However, using prior approaches researchers were not able to conduct a controlled experiment to investigate how and why this difference occurs.

In this example, we show how applying the concept of methodological middle spaces enables investigating the link between the Google Reply algorithm [27] and participants' language and interpersonal perceptions. Based on initial insights from prior work [35], we hypothesized that the sentiment of smart replies is related to the sentiment of conversations (i.e., observing more positive smart replies is related to more positive sentiment in conversation). We also hypothesized that, since Google smart replies are heavily positive [35], the use of Google smart replies improves interpersonal perceptions, in terms of participants' perceptions around their partner's cooperation as well as their sense of affiliation towards them.

6.1.2 Experimental challenges using existing approaches. Investigating this research question using prior experimental approaches would only achieve at most two of the necessary criteria of experimental research. For example, a screenshot study would achieve high levels of control and scalability, but it would also entirely sacrifice experimental realism. Since such studies do not provide the interactivity of a text-messaging application, they cannot engage participants in an actual conversation that is mediated via smart replies. As a result, it would be impossible to collect any conversational data and investigate the link between smart replies generated using the Google Reply algorithm and users' language production. A Wizard of Oz approach would enable the researchers to simultaneously address the challenges of realism and of experimental control. However, as described in Section 3, the WoZ approach would face the challenge of scalability due to the use of a human confederate as the wizard. For example, Hohenstein and Jung [35] ran their study with 72 participants (36 dyads), which still did not provide enough data to achieve statistically powerful results. Since each dyad had up to one hour to complete the task, running even this small study with a confederate playing the role of the AI would have taken weeks worth of researcher time.

Finally, the use of commercial AI-MC applications, on the other hand, could address the challenge of experimental realism. For example, Hohenstein and Jung [35] used Google Allo [25] to explore this research question. However, such commercial AI-MC tools do not provide experimental control required to investigate the effect of Google smart replies in a controlled setting [11, 70] where the only variables that differ across conditions are smart replies. For example, Hohenstein and Jung [35] had to use two different applications (i.e., whatsapp and Google Allo) for their control and their smart replies conditions, respectively. The use of different applications to investigate this research question can introduce extraneous variables that may have an effect on the outcomes. In addition, using these commercial applications the researchers were not able to collect data systematically. Instead, they had to conduct an in-lab study, where dyads attended sessions in-person and used the commercial messaging application (i.e., Google Allo). The researchers were required to collect data via screen recording and to manually transcribe each conversation, including smart replies, from screen recording of those sessions. With this procedure, gathering 36 conversations (i.e., 72 participants) took hundreds of hours over months of time. Additionally, participation was restricted to students at a specific university, limiting researchers' abilities to collect a diverse sample of participants. Lack of a systematic data collection limited the researchers' ability to achieve statistically powerful, replicable results.

6.1.3 Finding Methodological Middle Spaces. Moshi relaxes the perfect "mundane" realism (i.e., perfect replication of real world settings) in the use of commercial AI-MC applications to the degree required for experimental realism (i.e., providing a realistic setting to encourage participants

to take the study activities seriously and be mentally engaged with them) [6, 21, 54]. That is, instead of resembling any real AI-MC application, Moshi provides a prototype of such system. While this prototype of AI-MC applications is slightly different from the real AI-MC tools, it still provides experimental realism required to engage participants in an actual conversation. By giving mundane realism and instead pursuing experimental realism, Moshi provides the ground to achieve experimental control and scalability. That experimental control allowed us to design three conditions (i.e., both participants have smart replies, only one participant has smart replies, neither participants has smart replies) where the only variables that differ across conditions are the presence of smart replies. In the smart replies condition we integrated the existing Google smart reply-generating API [27] into Moshi. Using the same application across all the experimental conditions, we were able to change only the variable of interest (i.e., smart replies) across conditions, while keeping the other settings constant to avoid introducing any external variables between conditions.

In terms of scalability, unlike the use of commercial applications, Moshi allows researchers to collect data systematically. That is, instead of collecting the data via screen recording and manually transcribing each conversation, including smart replies, from screen recordings of those sessions, Moshi records all the required data systematically. In this example, Moshi enabled us to collect the details of participants' interaction necessary for linguistic analysis: conversation ID, message ID, time, message text, whether the message is a smart reply, and the smart replies shown at each step. We were able to collect this detailed data in real time, without any need to transcribe or clean any data after each study session. In addition, Moshi allowed us to collect messaging conversations through Mechanical Turk with 219 participant dyads (N=438), eliminating the need to bring participants into a laboratory and allowing for a more diverse participant sample. Achieving this fine-grained data collection at scale was not possible using the previously-employed procedure using Google Allo.

6.2 Example 2: Examining the Effects of Sentiment of Smart Reply on User Language

6.2.1 Motivation. Hohenstein and Jung [35] suggest the excess of positive language in Google smart replies API could cause the sender and receiver to also use more positive language in subsequent messages. Our previous study also reveals a potential link between sentiment of smart replies in Google Allo and the sentiment of subsequent conversations. However, those studies only compared the presence and absence of smart replies, rather than directly manipulating the sentiment of those replies to examine impacts on the conversation. Motivated by the insights from prior work and our finding in the previous study, we hypothesize that the sentiment of smart replies that the participants observe, even if they are not used during their conversation, have an impact on the sentiment of the participants' conversations.

6.2.2 Experimental challenges using existing approaches. To investigate this research question, the main challenge was to design a **realistic**, *controlled* experiment. We needed experimental control to manipulate the presence or absence of smart replies. In the presence of smart replies, we needed further to be able to control the sentiment of the smart replies that the participants observed during the experiment so that they have either a positive or negative sentiment. In addition, we needed experimental realism to provide the participants with a realistic setting where they could engage in an actual conversation. However, the existing approaches would not achieve both these requirements simultaneously. For example, while commercial AI-MC systems would allow us to achieve experimental realism, they would not allow us to have the control required to manipulate the sentiment of smart replies. A screenshot study, on the other hand, would allow for experimental control required to manipulate the sentiment of smart replies in each experimental condition, but

this approach would not allow for the experimental realism required to engage the participants in an actual discussion where they could engage in the experiment and use the AI-MC system. In addition, a screenshot study would not allow us to collect any conversational data and investigate the link between the sentiment of smart replies and the sentiment of subsequent conversational utterances. Lastly, the WoZ approach would provide experimental control to manipulate smart replies, and it would provide experimental realism to engage participants in an actual conversation. However, this approach faces the challenge of scalability as it requires a numerous hours of work from one or more human confederates.

6.2.3 Finding Methodological Middle Spaces. Moshi leverages the experimental control of screenshot study to achieve experimental control required to manipulate the smart replies. However, it relaxes the degree of experimental control offered in screenshot studies to also achieve **experimental realism**. To do so, it integrates static smart replies designed for the purpose of this study into an interactive text-messaging application. This application still allows to manipulate the smart replies to ensure participants observe either only positive sentiment or negative sentiment smart replies in each experimental condition. However, the dynamics of this application introduces some degree of nondeterministic behavior. In that, unlike a screenshot study wherein the researcher can control the exact wording of smart replies at each stage of the conversation,this application chooses the smart replies in a nondeterministic fashion from a pool of smart replies. Specifically, the smart replies are pulled randomly from an input json file[1] without being too repetitive (i.e., all three utterances shown in each instance were different, and the same utterance was not shown in immediately subsequent instances).

This relaxed experimental control still maintains the degree of experimental control required for this research question. That is, Moshi still allows to control the mediation action (described in subsection 5.1.3) that is the suggestion of smart replies with certain sentiment. The provided experimental control still allows to conduct a controlled between-subjects experiments with three conditions: positive and negative sentiment smart replies and a no smart reply as the control condition. An instance of Moshi with positive and negative sentiment smart replies, respectively, is shown in Figure 2. As a result of this compromise of perfect experimental control, Moshi provides the spaces to achieve experimental realism to the degree required for this research. In particular, Moshi enabled us to provide participants with a realistic setting to engage them in an actual mediated conversation. It should be noted that the mundane realism is not achieved as the suggested smart reply might not necessarily be relevant to the users' conversation (unlike in commercial applications where smart replies are generated based on the users' messages). However, as discussed in section 4, achieving the experimental realism is sufficient for the purpose of this research study.

This example shows how compromising experimental control, Moshi enabled to achieve simultaneous experimental control and experimental realism. That is, similar to the first study, Moshi enabled to achieve experimental realism (i.e., the participants engaged in an actual AI mediated conversation). More importantly, this trade off still maintained the experimental control to the degree required for this study. That is, Moshi still enabled to control the sentiment of smart replies, allowing to allowing to conduct a between-subject study wherein participants only observe either positive or negative sentiment smart replies depending on their experimental condition. In addition, Moshi enabled us to collect the details of interactions, including all necessary details for linguistic analysis (i.e., conversation ID, message ID, user name, time, message text, whether the message is a smart reply, and smart replies shown). By enabling fine-grained data collection at large scale,

[1]The json files were generated from previous work [35] where crowdworkers rated the sentiment of commercial smart replies, and the files included only those that were rated as having definitive positive or negative sentiment.

Moshi achieved scalability required to achieve statistically powerful, replicable results. The results and findings of this study are reported in [anonymized].

6.3 Example 3: Examining the Effects of "We" Smart Replies on Teams

6.3.1 Motivation. Prior work suggests that pronoun usage is a valid marker of how individuals think about themselves and their relationships, with first-person plural pronoun (e.g., "we") usage representing the degree of relational focus and cooperative communication [68]. Motivated by these insights from prior work, we were interested to investigate whether providing smart replies that were more likely to include first-person pronoun biased (e.g., "We can do it!", "We love it") could alter conversational linguistics on a team discussion, as well as team perceptions. Specifically, we aimed to examine whether first-person plural pronoun smart replies would lead to an increased use of first-person plural pronouns, and whether that can increase feelings of affiliation and improve interpersonal perceptions.

6.3.2 Experimental challenges using existing approaches. To explore this research question, similar to the second example, we need to design a *realistic, controlled* experiment that can also achieve **scalability**. We needed experimental control to control the presence or absent of the smart replies. In the smart replies condition, we needed to have experimental control to manipulate the content of the smart replies to ensure they are chosen from first-person plural pronoun smart replies. In addition, we needed experimental realism to provide a realistic setting to allow them engage in an actual conversation. Moreover, we also needed to conduct the study at scale so that we could generate statistically powerful, replicable results.

The experimental challenges that are explained in the previous two studies held for this study, as well. That is, prior approaches would not achieve all of these requirements simultaneously. Using an already existing commercial application would provide experimental realism, but would not provide experimental control to manipulate the content of the smart replies required to investigate the effect of the aforementioned biased smart replies. A screenshot study, on the other hand, would provide experimental control to manipulate the smart replies the participants observe during the study. However, this method lacks experimental realism as it would not allow participants to engage in an actual conversations. Lastly, while the WoZ approach would allow for both experimental control and experimental realism, it would face the challenge of scalability as this approach requires significant effort from confederates. Briefly, none of the existing methods would provide the three important aspects of experimental research (i.e., experimental realism, experimental control, and scalability) simultaneously.

6.3.3 Finding Methodological Middle Spaces. Moshi relaxes the experimental control and the experimental realism in the WoZ approach to enable scalability. To do so, it replaces the human confederate of the WoZ approach, which is the bottleneck to achieving scalability, with an automated bot. This bot randomly chooses smart replies from a predefined set of smart replies[2], and displays them to the participants during their conversation. While the smart replies are controlled and chosen from a predefined category of smart replies (i.e., first-person pronoun biased smart replies), the level of control over this selection conducted by the bot is relatively less compared to that of a human confederate. In particular, the use of an automatic bot as the wizard introduces some degree of nondeterministic behavior, in that the bot chooses smart replies randomly from a pool of smart replies. The researchers cannot ensure the exact smart replies displayed to the participants at each stage before the experiment runs. However, the use of bot as the wizard still provides the

[2]Given that there is no smart replies generator to produce only first-person pronoun biased (e.g., "We can do it!", "We love it"") smart replies, we manually generated the smart replies.

experimental control required for this research question. Specifically, given that the goal of the research is to understand the effects of first-person pronoun biased smart replies, it is sufficient to ensure that the smart replies are first-person pronoun biased. A perfect experimental control to control the exact wording of the smart replies the participants observe at each stage of their conversation would not be necessary. As a result of this compromise, Moshi was able to overcome the bottleneck of human confederate in the the WoZ approach and addressed the challenge of scalability.

It needs to be noted that by replacing the human confederate with an automatic bot, Moshi also relaxes experimental realism of the WoZ approach. In particular, the process of generating smart replies employed using Moshi does not accurately reflect the actual process by which smart replies are generated. That is, the smart replies are drawn from a fixed corpus created from existing smart reply transcripts to ensure they are first-person plural smart replies. Since these smart replies might not be relevant to the participants' ongoing conversations, this setting can decrease experimental realism. However, while this setting does not provide mundane realism (i.e., perfect replication of real world smart replies) [54], it still engages the participants in an actual conversation required to achieve experimental realism [6, 21]. Through compromising experimental realism, however, Moshi maintained the experimental control necessary to integrate the researcher's designed smart responses into the chat application.

Following this design, the Moshi application enabled us to design a two condition (i.e., first-person plural pronoun smart replies, no smart replies) between-subjects online study with 101 triads of participants (N=303). This way, we were able to generate initial insight into questions of whether an intelligent agent can be used to enhance team members' relational focus. The study procedure was the same as in our previous experimental study, and after completing a group conversation, we used survey instruments to measure the impact of AI mediation on interpersonal perceptions (e.g., Inclusion of Other in the Self (IOS), [5], IAS-R [88]). Linguistic analyses of the conversations examining the frequency of the "we" pronoun [68] were considered an index of relational focus [47].

This example shows how relaxing the experimental control and experimental realism of the WoZ approach via automatizing the role of wizard, Moshi provides spaces to achieve experimental control and experimental realism, and scalability simultaneously. This balance of experimental criteria enabled us to examine the effects of biased smart replies (i.e., smart replies featuring first-person plural pronouns from researchers-supplied list) on language choice, and on interpersonal perceptions in group discussions.

6.4 Summary

The above three studies combine to illustrate how the concept of methodological middle spaces was able to guide modifications of prior approaches to achieve all the required criteria of experimental research simultaneously. In the first example, we relax the experimental realism of commercial AI-MC tools to provide greater experimental control and to ease scalability. The second example demonstrates how relaxing the experimental control of a screenshot study can provide the interactivity and dynamics of AI-MC applications, thus recovering significant amounts of the previously sacrificed experimental realism. In the third example, relaxing the experimental realism and experimental control of the WoZ approach via automating the role of wizard greatly eases scalability. Thus, these three examples collectively demonstrate how the concept of methodological middle spaces can be applied to adapt a variety of different methods in ways that can recover previous sacrifices in any of the three main criteria for experimental studies.

7 DISCUSSION

As pointed out above, most of the currently dominant methods for experimental studies of AI-MC cannot simultaneously achieve all the key criteria of experimental research (i.e., experimental realism [6, 21, 54], experimental control [11, 45, 70], and scalability [2]). As a means to address these challenges, this paper presents the concept of methodological middle spaces (Discussed in Section 4). This concept is illustrated in the design of Moshi, a platform for conducting text-based AI-MC experiments. Exploring methodological middle spaces to design Moshi enabled us as researchers to design specific experimental set-ups and test hypotheses that would be challenging, and perhaps impossible, to test using existing methods.

Rather than promoting one particular tool, this paper advocates for the broader goal of exploring the methodological middle spaces. Put differently, rather than using Moshi itself, we encourage future researchers to consider what other kinds of methodological middle spaces for experimental research might be suggested by Moshi and its design?[3]

In some ways, the core idea behind methodological middle spaces is not terribly novel. One can find prior work wherein the design of controlled social experiments trades off one of the three criterion in favor of others [e.g., 16, 19, 22, 42, 87, 91].

Instead, the novelty of this paper's contribution comes from its utility to support future work in two ways. First, methodological middle spaces provides a conceptual vocabulary to reason about these trade-offs. By explicating how aspects of an experimental design that were intended to ensure one criterion (e.g., control) can impact other criteria (e.g., realism), researchers can reason about these trade-offs in a more conscious, reflective manner. Second, methodological middle spaces explicitly emphasizes the notion of balance. Rather than perfect attainment of any one criterion, methodological middle spaces encourages slightly relaxing these criteria to obtain a balance among them. Slight concessions on one criterion (e.g., control) can significantly increase other criteria (e.g., realism). Thus, this concept can help researchers to describe the rationale for how their experimental design choices balance among different competing criteria (control, realism, and scalability), as well as to justify why those choices are appropriate to the research question(s) of interest.

The remainder of this discussion suggests some of the possibilities provided by the concept of methodological middle spaces for AI-MC research. These suggestions illustrate how future researchers might extend the concept of methodological middle spaces and innovations from this paper: first to other text-based AI-MC; then to other, more diverse forms of AI-MC. Rather than a fully prescriptive dictum, this section instead gestures towards the possibilities that future researchers should pursue.

7.1 Methodological Middle Spaces for Other text-based AI-MC

We can envision the utility of methodological middle spaces to explore many more concepts in text-based AI-MC. As one example, the second study above (Section 6.2) shows that smart replies can be designed to encourage more positive sentiment in dyadic conversations. Methodological middle spaces, either as implemented in Moshi or more generally, can be leveraged to build upon this finding and explore the link between sentiment of smart replies and the overall sentiment of the conversation in group discussions. For instance, researchers could test the importance of symmetry in smart reply sentiment, i.e., whether or not smart replies boost sentiment of a conversation even

[3]The fields of human-computer interaction, computer-supported cooperative work, social computing, etc. already encompass a diverse ecology of methodological paradigms [61]. To clarify, the point here is not about hybrid approaches that cross these different paradigms (e.g., combining statistical machine learning and grounded theory method [8, 55]). Instead, the point is to ask how we might meaningfully combine existing approaches within the single methodological umbrella of experimental studies to help examine previously untestable hypotheses.

if they are only provided to some members in a group discussion. Our first study (Section 6.1) suggested that this effect occurs in dyads, but does it hold for group conversations, and of what size? Is there a critical mass in terms of the number of members in the group discussion who need to have the more positive smart replies to influence the sentiment of overall conversation?

These ideas could be taken further and applied to other questions in other text-based AI-MC. How might the presence of auto-correct influence perceptions of professionalism or politeness [18, 33]? Could the language of automatically generated profiles be manipulated to encourage viewers to draw specific conclusions about the user [38]? How might alterations to the outputs of automatic summarization systems shape perceptions both about what topics are prevalent and about how a community discusses those topics [9, 66]? Such studies would be difficult to conduct, if not impossible, using existing methods. However, the methodological middle spaces presented in this paper gives researchers the conceptual grounding to properly relax the strict constraints embodied in current methods while still adhering to the requirements of experimental research.

7.2 Methodological Middle Spaces for Other Forms of AI-MC

The concept of methodological middle spaces is not unique to text-based AI-MC; It can be extended to advance experimental methods to study other forms of AI-MC. For example, one concerning form of AI's involvement in human communication is deepfakes, where AI can be deployed to create a misrepresentation of what a person says or does in audio or video [77, 79]. Exploring the impacts of this technology on human communications using already existing methods would face experimental challenges similar to the ones described in Section 3. Exploring methodological middle spaces, however, can help modify existing methods to address those challenges. That is, following this concept, researchers can modify existing methods to design their experiments, while adhering to the required criteria of experimental research.

For example, to examine the impacts of deepfakes on interpersonal communication, instead of aiming to develop an actual deep fakes platform, researchers can relax the perfect "mundane" realism to the degree required for experimental realism. For instance, researchers can design an experimental video calling platform that purports to create real-time fake image manipulations of the callers. That is, instead of actually creating real deep fakes of the callers' video feeds using advanced AI technology, the platforms could simply suggest that the video that the participants see of their conversational partner is created using an AI manipulation. Without replicating a real deepfakes system, this experimental design still provides experimental realism required to engage the subject in a realistic experiment. By giving mundane realism and instead pursuing experimental realism, other required criteria of experimental research (i.e., experimental control, and scalability) can be provided. In particular, such a platform provides high degree of experimental control required to manipulate specific details about how the system functions. For example, this setting allows researchers to explicitly reveal the system's features, conceal them, manipulate them, or apply them asymmetrically across a wide range of attributes to enable many different types of research studies. In addition, by allowing to remotely collect diverse data at large scale, the platform address the challenge of scalability, required to achieve statistically powerful results.

Exploring the methodological middle spaces can also be particularly useful for experimentally investigating various ethical issues in AI-MC, such as bias, fairness, and transparency [30, 38, 60, 83]. For example, should AI mediation reveal itself, and, if so, what should such a disclosure look like to enhance and not to harm individuals' communication? Relaxing "mundane" realism, researchers can gain experimental control to prototype different AI interactions, e.g., with varying levels of disclosure, to explore these concepts and to advance our understanding of how these systems should be regulated and developed so that interactions are improved and unexpected social consequences are prevented.

In addition, the methodological possibilities provided by exploring methodological middle spaces enable researchers to go beyond the current state of the art in technology, and simulate systems that are not yet technologically possible or economically practical to develop [12, 13, 48, 50]. Exploring the methodological middle spaces to design different fictional scenarios allows researchers to explore the various impacts of future AI-mediated systems on human communications. Recent work has explored speculative methods, such as design fiction, as a means of eliciting participant responses around various not-yet-existing technologies [e.g., 3, 56, 58, 71, 73, 90]. Adapting such methods for experimental research could provide compelling opportunities for studying AI-MC systems that have not yet been implemented.

The methodological middle spaces advocated in this paper could also be useful more broadly to expand possibilities in designing controlled social experiments. For example, prior work implicitly balanced between experimental control of lab experiment, and experimental realism of field experiments in designing controlled social experiments in online spaces [10, 19, 51, 78]. While these designs do not completely replicate real-life social interactions, they still allow for creating believable social media scenarios required to achieve experimental realism. As a result, they open spaces to provide experimental control and scalability in conducting these experiments. Space precludes a full discussion of the applicability and the utility of these methods for other controlled social experiments beyond AI-MC research. However, we encourage future research to explore the methodological possibilities provided by exploring methodological middle spaces.

7.3 Limitations of Methodological Middle Spaces

While the concept advocated here works to expand methodological possibilities in designing AI-MC experiments, it is unlikely to be universally appropriate for every AI-MC experimental study. In some types of studies, it may be impossible to balance the needs of experimental realism, experimental control, and scalability. For example, studies that aim to explore the impacts of video-based AI-mediation on online relationships for a long period of time might require to maintain perfect mundane realism to ensure participants engage in the experiment effectively. In such contexts, the experimental setting may need to almost replicate an actual AI-MC system in order to engage the participants in the experiment for a long period of time. Designing such a setting may require researchers to implement their own AI system in order to run their experiments, which prohibits to follow the concept of methodological middle spaces.

When faced with such challenges, it is worth recalling the distinction between experimental realism and mundane realism [6, 21, 54]. For short interactions, people may be more able and/or willing to carry on meaningful engagements with a seemingly complex but actually fairly simplistic system (e.g., ELIZA [86]). While it is likely impossible to create a simplistic proxy for every conceivable AI system, this paper illustrates how a variety of AI techniques (e.g., smart replies, machine translation, deep fakes.) could likely be simulated to a sufficient degree to achieve experimental realism.

Another concern associated with methodological middle spaces is whether the findings from such studies will generalize to other environments. While these studies can be designed to present a realistic environment, they cannot ensure that participants will interact in the same way outside the study. For example, participants have different incentives (e.g., financial) for their interactions with the studies than would typical users of real AI-MC platforms. Such issues of generalizability, though, are not unique to the methodological methods provided by the advocated concept in this paper. Rather, they are fundamental limitations of experimental methods. In continuing this paper's call for methodological innovation, it would be valuable for future research to explore methodological innovations, beyond those provided by exploring methodological middle spaces, to improve the generalizability of experimental research findings.

8 CONCLUSION

This paper identifies some of the unique methodological challenges in conducting experimental AI-MC research, namely, simultaneously achieving the criteria of experimental realism, experimental control, and scalability. It highlights how most prior experimental methods that are used to study AI-MC focus on perfect attainment of one or two of these criteria, to the detriment of the other(s). To address these challenges, this paper contributes the concept of methodological middle spaces that combine elements of existing methods to seek a balance among all the criteria without sacrificing any of them.

We illustrate the utility of the concept of methodological middle spaces by employing it to design a platform for AI-MC research called Moshi. By drawing on prior approaches (i.e., screenshot studies, Wizard-of-Oz, and use of commercial application) but relaxing strict adherence to any one criterion for experimental research, the methodological middle space inhabited by Moshi achieves a balance among the three different criteria. Through a series of concrete examples, we then show how the concept of methodological middle spaces can be applied to inform the design of specific experimental studies for text-based AI-MC research. These examples demonstrate how exploring methodological middle spaces can expand methodological possibilities to examine research questions that were challenging, perhaps even impossible, to investigate using existing methods.

Thus, the experiments conducted using Moshi illustrate how applying methodological middle spaces can expand methodological possibilities for text-based AI-MC. By applying this concept in other domains, future research may be able to enhance methodological possibilities in ways that help develop a fuller, more rigorous understanding of how AI mediates human communication.

9 ACKNOWLEDGMENTS

We would like to thank the anonymous reviewers for their constructive comments. This material is based in part upon work supported by the US Army Research Lab under Award No. W911NF2120092.

REFERENCES

[1] Grammarly (2018). Free Online Writing Assistant. https://www.grammarly.com/. (Accessed on 10/05/2020).
[2] Omar Al-Ubaydli, John A List, and Dana L Suskind. 2017. What can we learn from experiments? Understanding the threats to the scalability of experimental results. *American Economic Review* 107, 5 (2017), 282–86.
[3] Aloha Hufana Ambe, Margot Brereton, Alessandro Soro, Laurie Buys, and Paul Roe. 2019. The Adventures of Older Authors: Exploring Futures through Co-Design Fictions. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Glasgow, 358. https://doi.org/10.1145/3290605.3300588
[4] Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. 2018. Sentiment Bias in Predictive Text Recommendations Results in Biased Writing.. In *Graphics Interface*. 42–49.
[5] Arthur Aron, Elaine N Aron, and Danny Smollan. 1992. Inclusion of other in the self scale and the structure of interpersonal closeness. *Journal of personality and social psychology* 63, 4 (1992), 596.
[6] Elliot Aronson, Marylinn B. Brewer, and J. Merrill Carlsmith. 1985. Experimentation in Social Psychology. In *Handbook of Social Psychology* (third edition ed.). Random House, New York, 441–486.
[7] Susan B. Barnes. 2006. A Privacy Paradox: Social Networking in the United States. *First Monday* 11, 9 (2006).
[8] Eric P. S. Baumer, David Mimno, Shion Guha, Emily Quan, and Geri K. Gay. 2017. Comparing Grounded Theory and Topic Modeling: Extreme Divergence or Unlikely Convergence? *Journal of the Association for Information Science and Technology (JASIST)* 68, 6 (June 2017), 1397–1410. https://doi.org/10.1002/asi.23786
[9] Michael S. Bernstein, Bongwon Suh, Lichan Hong, Jilin Chen, Sanjay Kairam, and Ed H. Chi. 2010. Eddi: Interactive Topic-based Browsing of Social Status Streams. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST)*. ACM, New York, 303–312. https://doi.org/10.1145/1866029.1866077
[10] Aparajita Bhandari, Marie Ozanne, Natalya N Bazarova, and Dominic DiFranzo. 2021. Do You Care Who Flagged This Post? Effects of Moderator Visibility on Bystander Behavior. *Journal of Computer-Mediated Communication* 26, 5 (2021), 284–300.
[11] Anol Bhattacherjee. 2012. Social science research: Principles, methods, and practices. (2012).

[12] Oloff C Biermann, Ning F Ma, and Dongwook Yoon. 2022. From Tool to Companion: Storywriters Want AI Writers to Respect Their Personal Values and Writing Strategies. In *Designing Interactive Systems Conference*. 1209–1227.
[13] Mark Blythe. 2014. Research through design fiction: narrative in real and imaginary abstracts. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 703–712.
[14] Kirsten Boehner, Rogério DePaula, Paul Dourish, and Phoebe Sengers. 2007. How emotion is made and measured. *International Journal of Human-Computer Studies* 65, 4 (2007), 275–291.
[15] Boomerang. (2018). Write Better Email. https://www.boomeranggmail.com/respondable/. (Accessed on 10/05/2020).
[16] Tommy Bruzzese, Irena Gao, Griffin Dietz, Christina Ding, and Alyssa Romanos. 2020. Effect of Confidence Indicators on Trust in AI-Generated Profiles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–8.
[17] Stephen Cave, Kanta Dihal, and Sarah Dillon. 2020. *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford University Press, Oxford.
[18] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL, Sofia, Bulgaria, 250–259.
[19] Dominic DiFranzo, Samuel Hardman Taylor, Franccesca Kazerooni, Olivia D Wherry, and Natalya N Bazarova. 2018. Upstanding by design: Bystander intervention in cyberbullying. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.
[20] Steven P. Dow, Manish Mehta, Blair MacIntyre, and Michael Mateas. 2010. Eliza Meets the Wizard-of-Oz: Blending Machine and Human Control of Embodied Characters. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Atlanta, GA, 547–556. https://doi.org/10.1145/1753326.1753408
[21] Jamie Druckman. 2022. *Experimental Thinking: A Primer on Social Science Experiments*. Cambridge University Press, Cambridge.
[22] Wen Duan, Naomi Yamashita, Yoshinari Shirai, and Susan R Fussell. 2021. Bridging Fluency Disparity between Native and Nonnative Speakers in Multilingual Multiparty Collaboration Using a Clarification Agent. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–31.
[23] Motahhare Eslami, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Seoul, 153–162. https://doi.org/10.1145/2702123.2702556
[24] Lauren Figueredo and Connie K Varnhagen. 2005. Didn't you run the spell checker? Effects of type of spelling error and use of a spell checker on perceptions of the author. *Reading Psychology* 26, 4-5 (2005), 441–458.
[25] Amit Fulay. 2016. Say hello to google allo: A smarter messaging app. https://blog.google/products/allo/google-allo-smarter-messaging-app/
[26] Amy L Gonzales and Jeffrey T Hancock. 2008. Identity shift in computer-mediated environments. *Media Psychology* 11, 2 (2008), 167–185.
[27] Google. 2020. Smart Reply. https://developers.google.com/ml-kit/language/smart-reply Accessed: 2020-02-25.
[28] Russell Haines and Joan Ellen Cheney Mann. 2011. A new perspective on de-individuation via computer-mediated communication. *European Journal of Information Systems* 20, 2 (2011), 156–167.
[29] Kevin Hamilton, Karrie Karahalios, Christian Sandvig, and Motahhare Eslami. 2014. A Path to Understanding the Effects of Algorithm Awareness. In *Extended Abstracts of the ACM Conference on Human Factors in Computing Systems (CHI EA) (CHI EA)*. ACM, Toronto, ON, 631–642. https://doi.org/10.1145/2559206.2578883
[30] Jeffrey T Hancock, Mor Naaman, and Karen Levy. 2020. AI-Mediated Communication: Definition, Research Agenda, and Ethical Considerations. *Journal of Computer-Mediated Communication* 25, 1 (2020), 89–100.
[31] Susan C Herring. 2002. Computer-mediated communication on the Internet. *Annual review of information science and technology* 36, 1 (2002), 109–168.
[32] Susan C Herring. 2007. Language and the Internet. *Language@ Internet* 4, 1 (2007).
[33] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 52:1–52:14. https://doi.org/10.1145/3134687
[34] Jess Hohenstein, Dominic DiFranzo, Rene F. Kizilcec, Zhila Aghajari, Hannah Mieczkowski, Karen Levy, Mor Naaman, Jeff Hancock, and Malte Jung. 2021. Artificial intelligence in communication impacts language and social relationships. arXiv:2102.05756 [cs.HC]
[35] Jess Hohenstein and Malte Jung. 2018. AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–6.
[36] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[37] Jess Hohenstein, Lindsay E. Larson, Yoyo Tsung-Yu Hou, Alexa M. Harris, Aaron Schecter, Leslie Dechurch, Noshir Contractor, and Malte F. Jung. 2022. Vero: A Method for Remotely Studying Human-AI Collaboration. In *Hawaii International Conference on System Sciences (HICSS)*. Honolulu, HI, 254–263.

[38] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T Hancock, and Mor Naaman. 2019. AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

[39] Anjuli Kannan, Karol Kurach, Sujith Ravi, Tobias Kaufmann, Andrew Tomkins, Balint Miklos, Greg Corrado, Laszlo Lukacs, Marina Ganea, Peter Young, et al. 2016. Smart reply: Automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 955–964.

[40] J. F. Kelley. 1983. An Empirical Methodology for Writing User-Friendly Natural Language Computer Applications. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI)*. ACM, Boston, MA, 193–196. https://doi.org/10.1145/800045.801609

[41] J. F. Kelley. 1984. An Iterative Design Methodology for User-Friendly Natural Language Office Information Applications. *ACM Transactions on Information Systems* 2, 1 (Jan. 1984), 26–41. https://doi.org/10.1145/357417.357420

[42] Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator Chatbot for Deliberative Discussion: Effects of Discussion Structure and Discussant Facilitation. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.

[43] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST) (UIST)*. ACM, San Diego, CA, 1–10. https://doi.org/10.1145/354401.354406

[44] Peter M Krafft, Michael Macy, and Alex" Sandy" Pentland. 2017. Bots as virtual confederates: design and ethics. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 183–190.

[45] Kathy J Kuipers and Stuart J Hysom. 2014. Common problems and solutions in experiments. In *Laboratory experiments in the social sciences*. Elsevier, 145–177.

[46] Hui Li, Nan Chen, Minjuan Zhou, Chenyi He, Jingbo Li, and Yujie Shi. 2019. How People Browse Mobile News Feed? A Study for Mobile News Feed Design. In *International Conference on Human-Computer Interaction*. Springer, 248–265.

[47] Wei Fang Lin, Yi Cheng Lin, Chin Lan Huang, and Lung Hung Chen. 2016. We Can Make It Better: "We" Moderates the Relationship Between a Compromising Style in Interpersonal Conflict and Well-Being. *Journal of Happiness Studies* 17, 1 (2016), 41–57. https://doi.org/10.1007/s10902-014-9582-8

[48] Rhema Linder, Chase Hunter, Jacob McLemore, Senjuti Dutta, Fatema Akbar, Ted Grover, Thomas Breideband, Judith W Borghouts, Yuwen Lu, Gloria Mark, et al. 2022. Characterizing work-life for information work on mars: A design fiction for the new future of work on earth. *Proceedings of the ACM on Human-Computer Interaction* 6, GROUP (2022), 1–27.

[49] Gardner Ed Lindzey and Elliot Ed Aronson. 1968. The handbook of social psychology. (1968).

[50] Yihe Liu, Anushk Mittal, Diyi Yang, and Amy Bruckman. 2022. Will AI Console Me when I Lose my Pet? Understanding Perceptions of AI-Mediated Email Writing. In *CHI Conference on Human Factors in Computing Systems*. 1–13.

[51] Philipp K Masur, Dominic James DiFranzo, and Natalya Natalie Bazarova. 2021. Behavioral Contagion on Social Media: Effects of Social Norms, Design Interventions, and Critical Media Literacy on Self-Disclosure. (2021).

[52] Scott E Maxwell. 2004. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological methods* 9, 2 (2004), 147.

[53] Hannah Mieczkowski, Jeffrey T Hancock, Mor Naaman, Malte Jung, and Jess Hohenstein. 2021. AI-Mediated Communication: Language Use and Interpersonal Effects in a Referential Communication Task. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–14.

[54] Douglas G Mook. 1983. In Defense of External Invalidity. *American Psychologist* 38 (1983), 379–387.

[55] Michael Muller, Shion Guha, Eric P. S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the ACM Conference on Supporting Group Work (GROUP)*. Sanibel Island, FL.

[56] Larissa Vivian Nägele, Merja Ryöppy, and Danielle Wilde. 2018. PDFi: Participatory Design Fiction with Vulnerable Users. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction (NordiCHI '18)*. ACM, Oslo, Norway, 819–831. https://doi.org/10.1145/3240167.3240272

[57] Hoyeon Nam, Hankyung Kim, and Youn-kyung Lim. 2021. User Experience of Agent-Mediated Interactions with Multiple Conversational Agents. In *HCI International 2021 - Posters*, Constantine Stephanidis, Margherita Antona, and Stavroula Ntoa (Eds.). Springer International Publishing, Cham, 472–476.

[58] Renee Noortman, Britta F. Schulte, Paul Marshall, Saskia Bakker, and Anna L. Cox. 2019. HawkEye - Deploying a Design Fiction Probe. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, Glasgow, Scotland Uk, 1–14. https://doi.org/10.1145/3290605.3300652

[59] Patricia A. Norberg, Daniel R. Horne, and David A. Horne. 2007. The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors. *Journal of Consumer Affairs* 41, 1 (2007), 100–126.

[60] Sladjana Nørskov, Malene F Damholdt, John P Ulhøi, Morten B Jensen, Charles Ess, and Johanna Seibt. 2020. Applicant fairness perceptions of a robot-mediated job interview: a video vignette-based experimental survey. *Frontiers in Robotics and AI* 7 (2020), 586263.

[61] Judith S Olson and Wendy A Kellogg. 2014. *Ways of Knowing in HCI.* Vol. 2. Springer.

[62] W. J. Ong. 1982. *Orality and literacy: the technologizing of the word.* London. Methuen.

[63] Martin T Orne. 1962. On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American psychologist* 17, 11 (1962), 776.

[64] Leysia Palen and Paul Dourish. 2003. Unpacking" privacy" for a networked world. In *Proceedings of the SIGCHI conference on Human factors in computing systems.* 129–136.

[65] Shinyoung Park, Akira Harada, and Hiroya Igarashi. 2006. Influences of personal preference on product usability. In *CHI'06 Extended Abstracts on Human Factors in Computing Systems.* 87–92.

[66] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. NewsCube: Delivering Multiple Aspects of News to Mitigate Media Bias. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI) (CHI).* ACM, Boston, 443–452. https://doi.org/10.1145/1518701.1518772

[67] Seema Patel, William Bosley, David Culyba, Sabrina A. Haskell, Andrew Hosmer, TJ Jackson, Shane J. M. Liesegang, Peter Stepniewicz, James Valenti, Salim Zayat, and Brenda Harger. 2006. A Guided Performance Interface for Augmenting Social Experiences with an Interactive Animatronic Character. In *Proceedings of the AAAI Conference on Artificial Intelligence in Interactive Digital Entertainment (AIIDE).* AAAI, Marina del Rey, CA, 72–79.

[68] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.

[69] Kraus Rachel. (2018). Gmail smart replies may be creepy, but they're catching on like wildfire. ((2018)). (Accessed on 10/05/2020).

[70] Hans Radder. 2003. *The philosophy of scientific experimentation.* University of Pittsburgh Pre.

[71] Sumita Sharma, Netta Iivari, Marianne Kinnula, Grace Eden, Alipta Ballav, Rocio Fatas, Ritwik Kar, Deepak Ranjan Padhi, Vahid Sadeghie, Pratiti Sarkar, Riya Sinha, Rucha Tulaskar, and Nikita Valluri. 2021. From Mild to Wild: Reimagining Friendships and Romance in the Time of Pandemic Using Design Fiction. In *Designing Interactive Systems Conference 2021 (DIS '21).* Association for Computing Machinery, New York, NY, USA, 64–77. https://doi.org/10.1145/3461778.3462110

[72] Paschal Sheeran. 2002. Intention—behavior relations: a conceptual and empirical review. *European review of social psychology* 12, 1 (2002), 1–36.

[73] Dilruba Showkat and Eric P. S. Baumer. 2022. "It's Like the Value System in the Loop": Domain Experts' Values Expectations for NLP Automation. In *Proceedings of the ACM Conference on Designing Interactive Systems (DIS) (DIS).* ACM, Virtual Conference, 100–122. https://doi.org/10.1145/3532106.3533483

[74] Choon-Ling Sia, Bernard CY Tan, and Kwok-Kee Wei. 2002. Group polarization and computer-mediated communication: Effects of communication cues, social presence, and anonymity. *Information Systems Research* 13, 1 (2002), 70–90.

[75] Nick Statt. 2018. Google now says controversial AI voice calling system will identify itself to humans. https://www.theverge.com/2018/5/10/17342414/google-duplex-ai-assistant-voice-calling-identify-itself-update. (Accessed on 10/05/2020).

[76] Kristin Stecher and Scott Counts. 2008. Thin Slices of Online Profile Attributes. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM).* AAAI, Seattle, WA, 127–135.

[77] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1–13.

[78] Samuel Hardman Taylor, Dominic DiFranzo, Yoon Hyung Choi, Shruti Sannon, and Natalya N Bazarova. 2019. Accountability and Empathy by Design: Encouraging Bystander Intervention to Cyberbullying on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–26.

[79] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2387–2395.

[80] Crispin Thurlow, Laura Lengel, and Alice Tomic. 2004. *Computer mediated communication.* Sage.

[81] Stephanie Tong, Joseph B Walther, et al. 2011. Relational maintenance and CMC. *Computer-mediated communication in personal relationships* 53, 9 (2011), 1689–1699.

[82] Amy Voida, Stephen Voida, Saul Greenberg, and Helen Ai He. 2008. Asymmetry in Media Spaces. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW).* ACM, San Diego, CA, 313–322. https://doi.org/10.1145/1460563.1460615

[83] T Franklin Waddell. 2018. A robot wrote this? How perceived machine authorship affects news credibility. *Digital journalism* 6, 2 (2018), 236–255.

[84] Lisa Slattery Walker. 2014. Developing your experiment. In *Laboratory experiments in the social sciences*. Elsevier, 127–144.
[85] Joseph B Walther and Malcolm R Parks. 2002. Cues filtered out, cues filtered in. *Handbook of interpersonal communication* 3 (2002), 529–563.
[86] Joseph Weizenbaum. 1966. ELIZA–a Computer Program for the Study of Natural Language Communication between Man and Machine. *Commun. ACM* 9, 1 (Jan. 1966), 36–45. https://doi.org/10.1145/365153.365168
[87] Angelina Widener and Sohye Lim. 2020. Need to belong, privacy concerns and self-disclosure in AI chatbot interaction. *Journal of Digital Contents Society* 21, 12 (2020), 2203–2210.
[88] Jerry S Wiggins, Paul Trapnell, and Norman Phillips. 1988. Psychometric and geometric characteristics of the Revised Interpersonal Adjective Scales (IAS-R). *Multivariate Behavioral Research* 23, 4 (1988), 517–530.
[89] Lotte M Willemsen, Peter C Neijens, and Fred Bronner. 2012. The ironic effect of source identification on the perceived credibility of online product reviewers. *Journal of Computer-Mediated Communication* 18, 1 (2012), 16–31.
[90] Richmond Y. Wong, Deirdre K. Mulligan, Ellen Van Wyk, James Pierce, and John Chuang. 2017. Eliciting Values Reflections by Engaging Privacy Futures Using Design Workbooks. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW (Dec. 2017), 111:1–111:26. https://doi.org/10.1145/3134746
[91] Nancy Viola Wuenderlich and Stefanie Paluch. 2017. A nice and friendly chat with a bot: User perceptions of AI-based service agents. (2017).
[92] Bin Xu, Ge Gao, Susan R Fussell, and Dan Cosley. 2014. Improving machine translation by showing two outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3743–3746.
[93] Elana. Zeide. 2015. Algorithms Can Be Lousy Fortunetellers. https://slate.com/technology/2015/05/crystal-app-algorithmic-fortunetelling-for-employers-and-potential-customers.html. (Accessed on 10/05/2020).
[94] Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345* (2018).

Received January 2022; revised July 2022; accepted November 2022