Simulating Human Imprecision in Temporal Statements of Intelligent Virtual Agents

Susanne Schmidt
Universität Hamburg
Germany
susanne.schmidt@uni-hamburg.de

Sven Zimmermann
Universität Hamburg
Germany
info.sven.zimmermann@gmail.com

Celeste Mason
Universität Hamburg
Germany
celeste.mason@uni-hamburg.de

Frank Steinicke
Universität Hamburg
Germany
frank.steinicke@uni-hamburg.de

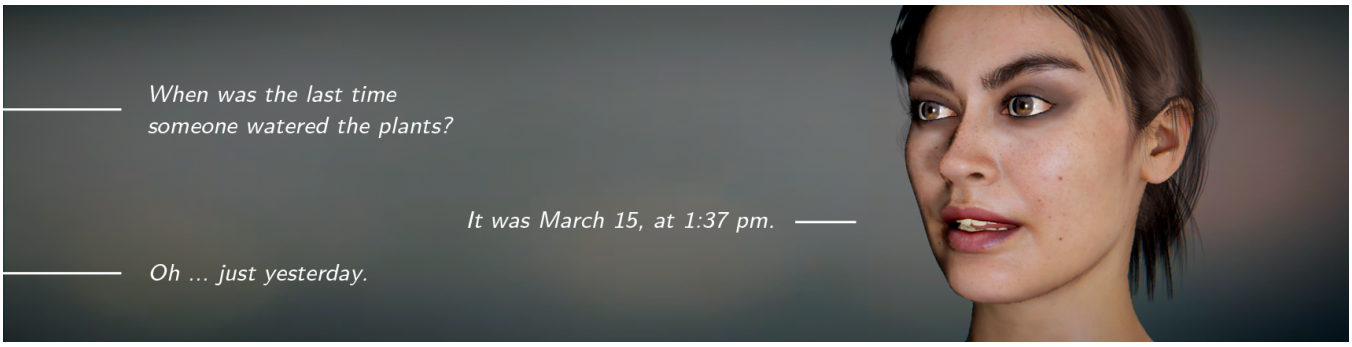

Figure 1: Illustration of an intelligent virtual agent providing overly precise temporal information to a user.

ABSTRACT

Research on intelligent virtual agents (IVAs) often concerns the implementation of human-like behavior by integrating artificial intelligence algorithms. Thus far, few studies focused on mimicry of cognitive imperfections inherent to humans in IVAs. Neglecting to implement such imperfect behavior in IVAs might result in less believable or engaging human-agent interactions. In this paper, we simulate human imprecision in conversational IVAs' temporal statements. We conducted a survey to identify temporal statement patterns, transferred them to a conversational IVA, and conducted a user study evaluating the effects of time precision on perceived anthropomorphism and usefulness. Statistical analyses reveal significant interaction between time precision and agents' use of memory aids, indicating that (i) imprecise agents are perceived as more human-like than precise agents when responding immediately, and (ii) unnaturally high levels of temporal precision can be compensated for by memory aid use. Further findings underscore the value of continued inquiry into cultural variations.

CHI '22, April 30–May 6, 2022, New Orleans, LA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517625

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *Natural language interfaces*; • **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

intelligent virtual agents, conversational agents, time perception

ACM Reference Format:
Susanne Schmidt, Sven Zimmermann, Celeste Mason, and Frank Steinicke. 2022. Simulating Human Imprecision in Temporal Statements of Intelligent Virtual Agents. In *CHI '22: ACM CHI Conference on Human Factors in Computing Systems, April 30–May 6, 2022, New Orleans, LA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3491102.3517625

1 INTRODUCTION

Intelligent virtual agents (IVAs), such as Amazon's Alexa and Apple's Siri, are becoming part of our everyday life. The convergence of research in the fields of machine learning and the Internet of things, among others, alongside significant improvements in voice communication technologies provide an unprecedented opportunity for users to control their connected homes, appliances, and myriad aspects of their lives through natural voice commands [48]. By utilizing immersive technology such as virtual reality (VR) in these scenarios, interactive conversational IVAs can appear in different forms, ranging from voice-only to embodied audio-visual representations, for example, with anthropomorphic bodies [57, 68]. IVAs displaying sufficiently advanced degrees of anthropomorphism

have been shown to increase trust and social presence during interactions with humans [14]. When designing (embodied) IVAs, researchers and developers often strive for the ideal of "human perfection", i.e., a realistic appearance, intelligent behavior, and interactions indistinguishable from those with real humans. This includes, for example, the perfection of simulated awareness towards conversation partners as well as multi-channel stimuli within the environment; the perfection of human speech, free of artifacts, capable of supporting semantics with appropriate intonation; or the perfection of intelligent dialog, making sense of complex linguistic patterns and delivering correct, logical responses [49]. While this approach is scientifically challenging, it neglects, however, that humans are not perfect. A "perfect" IVA would not pass the Turing test, as it would always be distinguishable from a human conversation partner [41]. Humans are also characterized by various "imperfections" that are not inherent to IVAs by design.[1] For instance, an IVA that is capable of providing accurate and precise details on past events without confounding effects such as a decay of memory traces over time would not appear as a believable human (where a definition of believability with regard to IVAs would be possessing qualities that allow its acceptance as real [2]). This is because human perception and cognitive processing of events underlie a number of biases, and recall of events gets less accurate with the passage of time [21]. Similar imperfect human traits can be noted for the other before-mentioned examples, including distractibility to irrelevant sensory inputs [38], speech disfluency expressed by filler words such as *'um'* and *'you know'* [40], and illogical (yet predictable) reasoning [41]. Neglecting these human imperfections in IVAs could reduce the perceived naturalness and believablity of IVAs. On the other hand, it is precisely because of their partially superior properties that IVAs are particularly suitable for assisting people with a wide range of tasks, such as message processing or time management. Therefore, creators of IVAs must continually attempt to optimize the balance of superior abilities, which provide the optimal performance expected of a computer, versus the simulation of human frailties, which may provide a more believable and engaging experience for users in a variety of contexts.

In this paper, we investigate both positive and negative effects of incorporating human-like "imperfect" response behavior into IVAs. While there are various aspects of human imperfection that could be transferred to IVAs, we are specifically interested in the integration of imprecise temporal references to past events. IVA assistance with time-related issues can be valuable for a variety of scenarios. IVAs can notify users about missed events (e.g., *'your mom called around 2 hours ago'*), provide information about other users (e.g., *'Dan already left the office shortly after noon'*), or support the tracking of progress, such as medicine consumption or nursing (*'last time you took your pills was this morning'*). While some degree of imprecision is common to all of these sample statements, overly precise responses could reinforce the impression of communicating with an AI rather than a human, as illustrated in Figure 1. This is because in human-to-human communication, temporal imprecision is typical for several reasons, including our limited memory capacity [21, 37]. There is also an ongoing discussion in the research community as to the existence of a time measurement mechanism innate to humans, as well as the level of precision with which we may measure time [6]. Humans are not only limited by these factors, but also consciously employ them when delivering temporal information with an appropriate level of detail and relevance targeted to the conversation partners' needs in a given context. While it is not often trivial to estimate how much each of these factors contributes to the resulting imprecise temporal statement, we attempt to focus on patterns of temporal imprecision caused by imperfect memory rather than other factors, such as social norms.

We consider the following research questions with a focus on the representation of time:

- How should IVAs refer to past events to appear natural?
- Do users prefer IVAs with a high level of anthropomorphism or (potentially unnatural) precision?
- How can we achieve a reasonable compromise between superior precision and human likeness?

The research questions are addressed in a three-stage process, involving (i) an assessment of temporal patterns used in human speech, (ii) a subsequent comparison between four different cultures to examine the generalizability of the results, and (iii) a final evaluation of the identified temporal patterns implemented into an IVA.

The remainder of this paper is structured as follows. Section 2 provides a theoretical overview of how humans memorize and express the time of past events, followed by a summary of related research projects with a particular focus on IVAs. In Section 3, we present a survey that we conducted to identify human patterns for expressing time. To gain insights into whether the obtained patterns are generalizable to different cultures, Section 4 presents a follow-up survey that was conducted in three additional countries. The resulting patterns were incorporated into a conversational embodied IVA with voice in-/output and evaluated in a user study, as described in Section 5. The paper is concluded by Section 6, which summarizes the main findings and discusses potential future research in the field of imprecise IVAs.

2 RELATED WORK

In this section, we will consider the concept of time both from a cognitive science and a linguistic perspective. For both research fields, we will summarize related work that attempted to practically implement those theoretical aspects into virtual agents.

2.1 Temporal Memory of Events

2.1.1 Memory and Human Development. As a person matures, from infancy through to adulthood, their conception of time develops, along with their ability to express measures of time [20]. Understanding of singular events in time, and ability to measure their duration, evolve along with the ability to determine their *'locations'* and the *'distance'* between them. Recognition of the order of serial events, as well as how they might be situated within a given reference frame (throughout a day, week, year), can aid in their memory, and provide a means to improve recall [37]. Comprehension of the passing of time develops along with each person's experience.

[1] We put imperfections in quotes, as psychologists and philosophers argued that this imperfection of humans represents their actual perfection (e.g., [8, 51]).

2.1.2 Limitations of Human Memory. Distinguishing the timing of events can depend on numerous factors, due to the instability and natural limitations of a person's memory. Theories regarding the way in which people perceive, store, and recall memories of past events make use of spatial metaphors to describe these methods within various contexts [20, 21, 37]. The *'distance'* (or relative time) of an event from another point in time may be regarded as nearer or farther based on its strength for an individual. So, the person's ability to store and recall the memory of the event, how old or recent they might perceive it, highly depends on their familiarity with that occurrence. Similarly, perception and retention of the *'location'* (or absolute time reference) of an event within a larger period of time might depend on how long it has been since the event occurred. The serial order (or relative time of occurrence) of events also plays a role in development of methods with which a person may consistently gauge the timing of an event, whether it may be precise for recent or important events, or more vaguely defined for distant or minor occurrences. Over time we forget finer details and the accuracy of one's perception of the timing of events decreases, with the median dating error increasing by as much as one day for each week in reference to a defined retention interval, as reported by Thompson [62].

2.1.3 Memory Aids. To compensate for the previously discussed limitations, humans leverage a variety of memory aids, both internal and external ones [35]. While the former include cognitive strategies such as mental rehearsal and the loci method [35], the latter is defined as any device or mechanism external to the user that aims at facilitating the memory. Examples include shopping lists, reminder notes, (electronic) calendars, and alarm clocks [36]. In a study by Intons-Peterson and Fournier [35], participants overall expressed a preference of external aids over internal ones due to their higher dependability, ease of use, and accuracy. Based on these beneficial properties, we hypothesize that cues provided by an embodied IVA that visibly uses such an external memory aid might be perceived as more reliable and increase the agent's believability.

2.1.4 Modeling Memory in Virtual Agents. Gomes et al. [23] proposed a method to provide agents with an episodic memory and, therefore, enable coherent behavior over long time periods. Stimuli that are currently perceived by the agent are correlated with stored stimuli of past events ("episodic memories"). Using a probability density function, stored events with a high correlation to the current situation are more likely to be selected and re-experienced by the agent. A preliminary evaluation indicated a positive influence of the memory model on agent believability. In an extended version of their work, the authors adapted their model to focus on a location-based association of memories [24].

Another approach to achieve long-time coherence and believability of IVAs is to equip them with a personal back story, or so-called autobiographic memories [5, 15]. For example, Ho et al. [34] proposed a storage structure that involves both details of the stored event (e.g., the time, location, involved people) and the event's consequences on inter-personal relationships with the IVA. How memories of past interactions and associated emotional responses can influence long-term relationships between an IVA and users was also investigated by Kasap et al. [39].

Complementing the retrieval of past events, Richards & Bransky [52] conducted a study to investigate IVAs who forget information over time. Their results suggest that partial recall and total loss of recall are preferred over incorrect recall, as they feel more natural and believable to the users. Nevertheless, agents with perfect recall yielded the overall highest scores for the users' enjoyment, level of trust, and believability of the agent.

Research regarding aspects of human memory of temporal events is broad in scope, but we concentrate here on the aspects directly related to the subject of this study. While there is a substantial amount of research devoted to retrospective and prospective memory (past as well as future events) [31, 69, 70], we focus on the reporting of individuals' memories of past events in this paper.

2.2 Temporal Representations and Expressions in Linguistics

2.2.1 Classifications of Temporal Adverbials. In English, there is a diverse set of temporal adverbials that indicate when a past event referenced by the speaker occurred [12, 60]. Such adverbials can take the form of noun phrases (e.g., *'this morning'*, *'on Friday'*, *'2 weeks ago'*), adverbs (e.g., *'yesterday'*), adjectives (e.g., *'earlier'*), or adverbial clauses (e.g., *'when I came home'*), among others [33]. Due to this diversity, constructing an all-purpose systematic classification is difficult and of limited use [33]. Nevertheless, efforts have been made to develop standards for the annotation of temporal expressions in texts. One of them is TimeML, which is an XML-based markup language that captures events, temporal expressions, as well as the relation between them [55]. Each temporal expression is assigned to a type *Date*, *Time*, *Duration* or *Frequency*, whereas for our purpose only the first two types are relevant, since we focus on temporal locations instead of time intervals and recurring events. If the expression itself or its context confidently determines the calendar date or time of day, it is directly mapped to an absolute ISO-compliant value, using rule-based or data-driven (i.e., machine learning) approaches [1]. For instance, assuming that today is Friday, July 12, 2002, *'Last Friday's party didn't start till twelve o'clock midnight'* will be mapped to the value *'2002-07-05T24:00'* [55]. In the TimeML standard, temporal expressions that are vague or have imprecise boundaries are annotated using specific tokens (e.g., *'afternoon'* is mapped to *'TAF'*), placeholders (e.g., *'63'* is mapped to *'XX63'*), or rules of interpretation (e.g., *'last night'* is always mapped to the date of yesterday). The annotation guidelines define every temporal statement as imprecise, if neither a particular time of day nor a calendar date can be induced from the context. This includes expressions with indeterminate precision, such as *'last week'*, as well as seasons, weekends, and parts of day [19]. Aside from this understanding of temporal imprecision, a number of alternative definitions can be found in the literature. For example, Zhou et al. [72] categorize relative time expressions such as *'yesterday'* as well as temporal distances such as *'2 days ago'* as what they call "fuzzy time expressions." In contrast, the imprecision definition of Tissot et al. [65] includes modified values (e.g., *'approximately 10 minutes ago'*) and ranges of values (e.g., *'8 to 10 years ago'*). In our paper, we follow a notion of imprecision similar to the one used in TimeML, with the addition that imprecision increases with a

coarser time granularity (e.g., *'12 hours ago'* is considered more precise than *'yesterday evening'*, which in turn is more precise than just *'yesterday'*). We also note that the imprecision of a response is highly related to the question, as, for example, *'yesterday'* can be considered a sufficiently precise response to the question *'When was your last day off?'* but an imprecise response to *'When did you have lunch?'*. Finally, we classify the examples given by Tissot et al. as signals of uncertainty rather than imprecision.

2.2.2 Absolute and Deictic Temporal Expressions. As noted in Section 2.1.1, the ability to accurately gauge the passing of time develops alongside our development of speech, consequently, the accuracy with which we describe temporal events is refined as a person matures. Even so, the limitations of a person's memory do not extend indefinitely, and so we often use different terms to refer to events depending on how far in the past they may be (along with other contextual conditions) [37]. For more recent events, it is common to rely on absolute expressions, whereas for events further in the past, one would more typically rely on relative expressions of occurrence to the current point in time or other "landmark" events [50]. Use of temporal deixis, pointing terms used to indicate this relative positioning are common (this, that, now, then, etc.), along with adverbs of frequency (sometimes, often, rarely), whether definite or indefinite [42].

2.2.3 Cultural Variations in Time-related Expression. The degree to which time expression for an individual may vary depends on numerous factors, regional and cultural differences being an important component. Sircova et al. [59] compared use of time expressions by individuals from 24 countries in Europe, Asia, Africa, and the Americas, based on their orientations to the past, present, and future with the use of the Zimbardo Time Perspective Inventory (ZTPI). The ZTPI measures five temporal orientations, covering attitudes and emotions or various valances. The work of Graham [29] focuses on how individuals perceive their time to be valued from a consumer research perspective, categorizing groups based on three types - Linear-Separable, Circular-Traditional, and Procedural-Traditional - which they derive from historical cultural contexts. These findings, while noting that they themselves clearly exhibit racial/cultural biases, indicate that cultures in which individuals do not stress the value of experiences in the future (perhaps because they do not feel a sense of agency that would empower them to change it), or do not value delineations between past, present and future, may be more likely to focus on optimizing the utility of the present. Rojas et al. [53] developed a questionnaire to assess attitudes of individuals in UK, Saudi Arabia, Thailand and Chile towards not only past, present and future orientations, but also time (duration) pressure and planning. The prevalence of schedules that promote multi-tasking versus focusing on one task at a time, known as polychronic time and monochronic time respectively, are also key to insights in variations of individuals' attitudes towards time reporting. As an example, from the polychronic perspective, punctuality would not be valued as much as in a monochronic time-focused culture. Usunier et al. [66] present a comparison of time expression usage from France, West Germany, Brazil, Mauritania, and South Korea, using a questionnaire focused on the economicity of time and polychronism vs monochronism. While the questionnaires discussed here cover interesting concepts to base future inquiries upon, our surveys address matters focused only on our present purposes, namely, defining the degree to which individuals consider precision in the reporting of events useful to them when requesting help from a person (or an IVA).

2.2.4 Modeling Temporal Vagueness in Virtual Agents. Burkert et al. [9] focused on the representation of time-related information in the context of non-playable characters (NPCs) in video games, but their findings may be applicable to other forms of IVAs as well. In a survey, they found that users prefer vague temporal patterns such as *'after lunch'* over exact time specifications when asking time-related questions to the virtual characters. The authors concluded (but not validated) that users would expect a similar vagueness in the responses of the virtual characters. Results of the survey were incorporated into a time module for NPCs, which mimics different psychological phenomena, including the above-mentioned usage of vague temporal patterns as well as partially forgetting information over time.

In line with these results, Rong et al. [54] reported a tendency of users to employ imprecise language when setting up reminders with the aid of virtual assistants, such as Apple's Siri or Microsoft's Cortana. The conducted survey indicated that users prefer different temporal granularity for different types of tasks, with minute-level precision being considered rather inappropriate in the majority of cases. However, for short-term reminders (e.g., in the context of cooking or cleaning tasks), Graus et al. [30] found that delays between reminder creation and notification are distributed with distinct spikes around five-minute intervals. Though both papers focus on prospective events, some of their results may also be applicable to past events, therefore warranting further comparative investigations.

In contrast to the existing research projects presented so far, we are particularly interested in (i) the referencing of past events and (ii) the perception of time-related responses given by IVAs rather than time-related queries made by the user. With these objectives in mind, we will collect, cluster, and quantify responses to time-related questions, both under consideration of the existing classifications presented in Section 2.2.1 and the theoretical framework of retrospective memory as shown in Section 2.1.2. We believe that this process will allow us to gather natural utterances that represent the targeted human-agent interactions better than existing datasets (such as the corpus of emails analyzed in [54] or the Cortana reminder logs used in [30]). It will also allow us to identify temporal patterns that may be specific to the described scenario and the previously introduced notion of imprecision. For these patterns, we expect differences in comparison to those reported by Graus et al. [30] and Rong et al. [54] since the main reasons for using imprecise references to past and future events presumably diverge (for the latter, a low level of commitment, low task priority, and external dependencies of the tasks in question were identified [54], all of which are not directly transferable to past events). Finally, we want to leverage the potential of the embodiment of IVAs by complementing verbal temporal responses with nonverbal behaviors such as checking an external memory aid. The aim is to create a balance between superior abilities of IVAs and the simulation

of human frailties, which may provide a more comfortable and engaging experience for users.

3 STUDY 1: COMMON PATTERNS IN TEMPORAL STATEMENTS

The transfer of human-like traits to IVAs requires first and foremost a deep understanding of what is perceived as human-like. Therefore, a preliminary survey was conducted to identify common patterns in how humans refer to the time of past events in common parlance. Participants were asked to answer a total of 18 questions that addressed events with different anticipated times of occurrence. The responses were analyzed by applying a clustering strategy as described below.

3.1 Participants

The survey was distributed to 63 participants, 27 of which were female and 36 male (aged from 18 to 44, $M = 29.65$). Participants were recruited from a local university, a survey exchange website [61], and a volunteering IT company. They did not receive any monetary reimbursement for their participation. None of the participants was an English native speaker, however, being proficient in English was explicitly stated as a requirement for participation in the study.

3.2 Materials

The survey was created using Typeform, which supports participation via any web browser. It contained 18 questions that cover a large time span from events in the most recent past (e.g., *'When did you start filling out this questionnaire?'*) to events that may have occurred multiple years ago (e.g., *'When was the last time you sent out an application?'*). Since we were particularly interested in expressions for indicating recent events, 10 out of 18 questions referred to activities that are common during a regular day (e.g., drinking, checking the mobile phone, having a break). The remaining questions were included to gain a better understanding of whether and how the temporal distance to past events not only affects the dating error (as described in Section 2.1.2) but also the way humans are referring to the events. An overview of the 18 posed questions can be found in Table 1 of the supplementary material.

The study was purposely not conducted under controlled lab conditions with a pre-defined time schedule. Therefore, the same questions led to various responses as the actual time of events can be individually different. Both benefits and disadvantages of this design will be discussed in Section 3.6.

3.3 Methods

Participants received a link to access the questionnaire on the above-mentioned survey exchange platform or via email. They could start the survey at any time without direct contact to the experimenter. The questionnaire was prefaced by a general introduction that explained the procedure without disclosing the specific focus on expressions of time. Participants were explicitly asked to phrase their responses in the same manner as they would in a natural verbal conversation. Lastly, participants were informed that they can skip questions in case they do not remember a particular event or do not want to provide information about it. The introduction was followed by 18 time-related questions, which were presented in the same order for all participants. Responses were received via free text input fields to allow all types of data, including whole sentences, phrases, and dates. The survey was concluded by some demographic questions. On average, it took participants around 10 minutes to complete the survey.

3.4 Data Analysis

All responses of two participants had to be excluded from the analysis because they consisted of meaningless character sequences (e.g., "ee"). The remaining 1098 responses were analyzed using a mixed-methods content analysis approach, with hybrid (observation- and theory-based) coding of the data [3]. Each survey response was transferred onto a digital sticky note on the digital whiteboard platform Miro [46]. After becoming familiar with the data, the first author performed a first iteration, in which only utterances with the same wording but possibly different numerical values were grouped together (e.g., *'2 minutes ago'* and *'30 minutes ago'*). Subsequently, the resulting clusters were spatially rearranged in such a way that their proximity roughly represented a combination of lexical similarity (measured by means of the Levenshtein distance at word level) and semantic relatedness (validated with WordNet [45] sister terms) [22]. For example, the clusters *'x minutes ago'* and *'x hours ago'* were placed next to each other, as were *'x minutes ago'* and *'approximately x minutes ago'*. In contrast, *'yesterday'* was located further away, but close to *'yesterday morning'* as well as *'today'*. Based on the resulting map, the first and third authors discussed possible core variables as well as categories, taking into account both observed patterns (inductive coding) and literature-based concepts as presented in Section 2.2 (deductive coding). Based on the agreed-upon categories, clusters were either retained, merged (e.g., *'x minutes ago'* and *'approximately x minutes ago'* because *'approximately'* was identified as a modifier expressing uncertainty without constituting a new precision category), or separated (e.g., *'x minutes ago'* into utterances with integers and floating-point numerical values). Finally, the number of utterances in each category was counted. The resulting categorization and the determined frequencies are described in the following section.

3.5 Results

The applied content analysis resulted in two core variables *time format* and *time granularity*. While *time format* involves the linguistic representation of dates and times, *time granularity* refers to the smallest unit that is used in a temporal statement, for example, a minute, a day, or a year.

Considering time formats, the two concepts of *temporal locations* and *temporal distances* emerged (cf. Section 2.1). While temporal locations directly point towards the moment an event happened (e.g., *'8 a.m.'*), temporal distances indicate the difference between the current time and the time of the event (e.g., *'two hours ago'*). For temporal locations, we found expressions of time that were numerical (such as *hh:mm*) or conceptual (such as weekdays, calendar months etc.). For example, *'07/04/21'*, *'Yesterday'*, and *'Last Sunday'* can all refer to the same temporal location but are either numerical or conceptual. In contrast to temporal locations, temporal differences showed a consistent form that included a numeral along with a time adverbial such as *ago* or *back*. For each category

Table 1: Different combinations of time format (table rows) and time granularity (table columns) that were identified in the survey responses. For each combination (table cell), both the number of occurrences and an example response are listed.

		Seconds	Minutes	Hours	Days	Weeks	Months	Years
Numerical Location	**...with implicit deictic center**	-	75 *9:45 a.m.*	24 *at 4 p.m.*	-	-	-	2 *2020*
	...with explicit deictic center	-	43 *Last night at 2:00*	32 *Saturday 8 a.m.*	8 *4th January*	-	-	-
Conceptual Location	**...with implicit deictic center**	-	-	2 *In the last hour*	110 *Monday / Today*	28 *Last week*	33 *This September*	15 *Last year*
	...with explicit deictic center	-	-	-	4 *Last week Friday*	-	33 *July 2020*	-
Distance	**...with implicit deictic center**	4 *30 seconds ago*	191 *5 minutes ago*	93 *4 hours ago*	29 *2 days ago*	37 *1 week ago*	58 *1 month ago*	44 *2 years ago*
	...with explicit deictic center	-	4 *Today, 2 mins ago*	3 *Today, 3 hrs ago*	-	-	-	-

of time format, we observed responses that are expressing time (i) relative to the present moment (i.e., the center of deixis is implicit), or (ii) relative to a specified reference point (i.e., the center of deixis is explicit).

For the second core variable, *time granularity*, we found temporal statements in the range of seconds to years. Depending on the time format, time granularity has a differing level of expressiveness. For temporal locations, the level of granularity directly determines the size of a temporal interval that includes the referred event (cf. [71]). For example, all temporal locations in the category *Days* (i.e., numerical, such as *'07/04/21'*, and conceptual, such as *'Yesterday'*) point towards a time interval with a length of 24 hours. In contrast, granularity in temporal distances only indicates the used unit but does not allow for concluding a specific precision of the time indication. For example, the expression *'2 days ago'* could refer to an event that happened exactly 2 days ago, but could also be used to refer to the entire day before yesterday.

Based on this conceptualization, all utterances were assigned to one level of each of the two core variables. From the overall 1098 responses, 94 were excluded because they were empty, did not contain an (unequivocally assignable) reference to a point in time, or referenced the present moment instead of past events. Responses with two alternative indications of time (e.g., *'3 hours ago, at 3 p.m.'*) were counted twice. Table 1 summarizes the number of occurrences for each main category. Besides these main categories, we also observed multiple intermediate levels of time granularity. These included fractions of a minute (e.g., *'Half a minute ago'*, $N = 2$), fractions of an hour (e.g., *'Half an hour ago'*, $N = 12$), and times of the day (e.g., *'This morning'*, $N = 73$) as well as specific parts of a week (e.g., *'Last weekend'*, $N = 2$), month (e.g., *'Late April'*, $N = 6$), and year (e.g., *'In summer'*, $N = 11$). Furthermore, participants used personal landmarks (e.g., *'After waking up'*, $N = 27$) as well as general landmarks (e.g., *'Before the pandemic'*, $N = 10$). Finally, 8 responses were indefinite, such as *'recently'* or *'a while ago'*.

While each category on its own would be suitable for further analyses, we decided to focus on the pattern which was most prevalent in the collected data. With a total of 482 utterances, *temporal distances* were used in 44% of all responses. All responses of the category were further clustered by the time values that were mentioned. Expressions referring to the same point in time such as *'60 minutes'* and *'1 hour'* were grouped. Figure 2 illustrates the resulting distribution. The five most frequent time values were 5 minutes ($N = 44$), 10 minutes ($N = 33$), 30 minutes ($N = 28$), 1 hour ($N = 43$), and 2 hours ($N = 23$).

3.6 Discussion

We decided in favor of an uncontrolled survey, without predetermined verifiable events. This decision was made since we were particularly interested in the phrasing of responses rather than their correctness. A laboratory setting may have influenced the perception and, consequently, the expression of time for controlled events. For example, it lacks personal temporal landmarks that events can be associated with. Furthermore, a controlled long-term experiment would have been necessary to assess remote events. By using multiple questions and increasing the sample size, it is reasonable to assume that a large time span was covered by the responses. Despite the uncontrolled nature of the survey, multiple conclusions can be drawn, which are of particular value for simulating human behavior in IVAs. First, the data suggests a human preference of rounded time references over precise ones. We observed a tendency

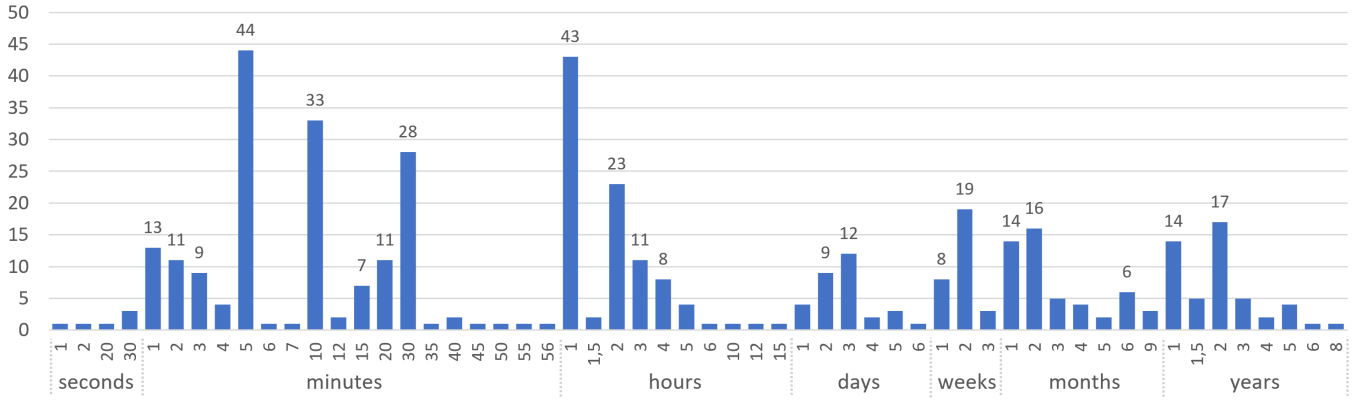

Figure 2: All time values that were observed in temporal distance statements. The y axis shows the number of occurrences for each time value.

towards specific reference points for rounding, including 1, 2, 3, 5, 10, 15, 20, 30, 60, and 120 minutes. Second, it was observed that intervals between reference points become larger as distance to an event increases. For events that (subjectively) happened within the previous 60 minutes, we identified eight common reference points. In contrast, the time period between one and four hours in the past was covered by only four reference points at the top of every hour.

3.7 Limitations

Besides the positive aspects in comparison to a controlled laboratory study, the chosen survey design is also associated with certain limitations. As the actual time of events cannot be verified, correct responses at the reference points (e.g., *'30 minutes'*) are indistinguishable from rounded values at the same point. However, as is can be assumed that the probability of an event occurring 30 minutes or 27 minutes in the past is approximately the same, the inferences drawn from the histogram in Figure 2 are still valid since only modes of the distribution were considered. As a second consequence of the actual event time being unknown, the data does not allow one to draw a conclusion regarding the rules of rounding up and down. From the literature, it is known that the magnitude of temporal displacement of recalled events is affected by a number of factors, including the distance of the event (cf. telescoping bias [63]) as well as its subjective memorability [62]. In the following, we will omit such complex cognitive processes and apply a simple rounding strategy to time references. Also, it is not yet clear to which extent each of the factors described in Section 2.1 is contributing to the observed preference of rounded time expressions. We expected further insights on this question from Study 3, which used the previously described observations to implement different time referencing behaviors into an IVA. Finally, one limitation of the current survey was that while fluency in English was one of the inclusion criteria, none of the participants were native speakers. Although a study conducted by Crawford [11] showed no differences between L1 and (German) L2 students in the usage of two exemplary temporal adverbials, it is not clear whether this finding can be generalized to other temporal patterns. In a follow-up survey, we will examine whether the patterns we have identified hold for other cultures, including U.S. citizens whose first language is English. In the future, such a validation could also be performed in Germany to reveal possible differences between the usage of temporal patterns in the native versus a foreign language.

4 STUDY 2: CULTURAL VARIATIONS IN TEMPORAL STATEMENTS

Since Study 1 was conducted in a single country, conclusions in terms of the time referencing behavior are likely biased towards one specific culture. As the literature suggests cross-cultural differences in the perception of time (see Section 2.2.3), additional considerations are required to evaluate the generalizability of the previously presented results. In order to gain first insights into potential cultural variations, we expanded our initial survey assessing temporal patterns to three more countries.

4.1 Participants

We distributed the survey described in Section 3.2 to inhabitants of the United States (8 female, 6 male; mean age of 30.71 years), Japan (14 male; mean age of 25.50 years), as well as southwestern India (9 female, 5 male; mean age of 36.71 years). These countries were selected because they differ with respect to their orientation towards past, present, and future, their time horizon, as well as their tendency towards polychronicity [32]. As the literature suggests that time perception is strongly shaped by sociocultural influences at a young age [20], participation in the survey required being born in the respective country and having lived there for at least 5 years during childhood. For Japan and India, the survey was translated to Japanese and Kannada, respectively, and delivered together with the English translation. Participants were encouraged to answer questions in the language they feel most comfortable with. In India, one participant used Kannada for their responses while the remaining ones answered in English, which is one of the country's two official languages. In Japan, all participants answered in Japanese. Participants' backgrounds covered a wide societal range, including students, university staff, pensioners, and homemakers.

4.2 Data Analysis

As we intended to validate our findings from Study 1, we used the identified categories as reported in Section 3.5 and conducted a content analysis with deductive coding in Study 2. Before coding all 756 utterances, Japanese and Kannada responses were translated using both DeepL [13] and Google Translate [26], and in case of divergence, a native speaker was consulted for resolution. After coding each utterance, we again recorded frequencies for all categories and color-coded them, as shown in the supplemental materials, to make peaks in the distribution clearly visible and to provide a common basis for comparisons across cultures.

4.3 Results

Since all temporal statements could be unambiguously assigned to one of the categories established in Study 1, the two core variables *time format* and *time granularity* with according levels could be confirmed. In the following, we will discuss emerging trends that we observed in the data.

4.3.1 Time Format and Granularity. The number of occurrences for each category and every country is summarized in Tables 2 to 6 in the supplementary materials. As in the initial survey, there is a preference towards the usage of temporal distances over numerical and conceptual temporal locations for the U.S. (62.0% of responses) and Japan (54.7% of responses). In contrast, the most prevalent time format in Indian responses are conceptual temporal locations, used in 47.6% of the cases. In strong correlation to this preference of conceptual time statements is the high number of vague time references even for recent events. In particular, the statements *'today'* and *'yesterday'* were used in 11.1% and 15.4% of Indian responses, respectively (compared to average values of 0.3% and 5.1% across the other three countries).

4.3.2 Reference Points for Temporal Distances. The reference points for temporal distances identified in Study 1 were confirmed in the follow-up surveys for each of the three selected countries. As for the German sample, the five most frequent time values, reported as a sum over the American, Japanese, and Indian samples, were 5 minutes ($N = 17$), 10 minutes ($N = 15$), 30 minutes ($N = 17$), 1 hour ($N = 35$), and 2 days ($N = 16$). Likewise, the intervals between reference points got larger for increasing temporal distances to the occurred event. As a minor observation, Japan was the only one of the considered cultures using more than 24 hours to refer to past events (e.g., *'50 hours ago'*). However, due to the small number of occurrences ($N = 4$), more comprehensive investigations are required to determine if this can be considered a general trend in the population.

4.3.3 Uncertainty. Besides the trend towards vague statements for recent events, Indian responses showed another difference in comparison to the other cultures. In the initial survey conducted in Germany, 10.5% of the categorized statements expressed some degree of uncertainty by using adverbs like *'about'* and *'around'*, question marks at the end of the statement (representing rising intonation in spoken language), or other indicators, such as *'I guess'*, *'probably'*, *'maybe'*, and *'or so'*. Such indicators were found in a similarly large portion of the American (14.9%) and Japanese (9.8%) responses. In contrast, only 2.9% of the Indian responses included expressions that indicate an uncertainty of the speaker.

4.4 Discussion

While the general categorization established in Section 3 could be confirmed, we observed variations of temporal patterns particularly in the survey responses from India. A preference towards conceptual temporal locations, the tendency to use vague statements for recent events, as well as the rare use of indicators for uncertainty reflect the same underlying principles of time perception in Indian culture. Given the preliminary nature and scale of this inquiry, it would be difficult to attribute these findings to any particular factors, however we might conclude that the tendency toward polychronicity in addition to India's unique traditional, religious and cultural contexts may underlie the divergent results we have observed. Although the literature on time perception also reports differences between Japan, the U.S., and Germany [32, 43, 64], these variations appear to have a less severe impact on the patterns used in temporal statements.

4.5 Limitations

Because Study 2 included only three selected countries and a comparatively small sample of 14 participants per country, it should be considered as a first exploration rather than a statistically reliable evaluation. Nevertheless, multiple trends emerged that indicate cross-cultural differences and, therefore, warrant additional investigations to guarantee a high usability of IVAs for a diverse user group. Differences between states or regions within one country may also be revealed through further refinement of survey locales.

Finally, we did not ask respondents in either survey whether they used a diary or an app to retrieve information. Therefore, addition of questions to gain better understanding of the basis of individual responses may be warranted, as some participants may value accuracy to the degree that they would check the timing of events using memory aids. Such behavior may also vary across cultures and thus could provide valuable insights into locally different expectations for the use of memory aids by an IVA.

5 STUDY 3: TRANSFERRING TEMPORAL PATTERNS TO VIRTUAL AGENTS

The preliminary surveys revealed tendencies of humans to use imprecise temporal locations and to round temporal distances to specific reference values that depend on the elapsed time since the occurrence of the referred event. In Study 3, we aimed at transferring this behavior to an IVA and testing the resulting effects on perceived anthropomorphism and usefulness of the IVA. Anthropomorphism was chosen as a measure because it has been shown in previous studies to be a crucial trait contributing to the believability of an IVA [14]. The following hypotheses were investigated:

(H1) IVAs are perceived as more anthropomorphic when they use imprecise instead of precise indications of time.
(H2) IVAs are perceived as more useful when they use precise instead of imprecise indications of time.
(H3) When providing precise indications of time, IVAs that use a memory aid are perceived as more anthropomorphic than

those responding immediately.

(H1) is directly derived from observations of human communication behavior in the survey, while (H2) is based on the assumption that precise responses possess a higher information content compared to imprecise ones. Acceptance of both (H1) and (H2) would face developers with the decision to either create believable but less useful IVAs or to maintain precision while sacrificing believability. To resolve this dilemma, we suggest adopting another human strategy for coping with limited memory; the usage of a memory aid such as a calendar. Since using memory aids for time management is common in real-world scenarios (cf. Section 2.1.3), we hypothesize that looking up an exact date and time in a digital calendar would justify a high level of precision without reducing the perceived anthropomorphism of the IVA (H3).

5.1 Participants

We invited 24 participants, 18 male and 6 female (aged from 18 to 32, $M = 22.91$). Based on an alpha level of .05, 24 participants yield a power of .74 at the effect size Cohen's $f = .25$ [16]. None of the participants was an English native speaker, however, being proficient in English was mentioned as a requirement for participation in the study. For technical reasons, all participants were required to use Windows as their operating system. They did not receive any financial reimbursement for their participation.

5.2 Materials

To validate our hypotheses, we created a virtual environment that allowed users to talk to an IVA and, with her help, answer a total of 12 questions directed at them. The environment was modeled after a typical office setting, in which the IVA acts as a virtual assistant who created notes on past events (e.g., a missed phone call for a colleague, a sick leave notification, or the arrival of external guests) using digital tools, such as an electronic calendar or a note-taking application. According to van den Hooff [67], electronic calendars are one of the primary tools for facilitating organizational coordination, as they not only support planning of future activities but also serve as an archive for tracking the history of the users' work life. The digital notes were displayed on a laptop for reasons of both plausibility and practicality, as the memory aid must fit thematically into the setting while being clearly recognizable by study participants. The virtual environment was implemented in Unity and was coupled with several Google services to provide speech recognition and dialog generation. For conducting the study, the system was set up on a virtually hosted Windows machine on an Azure server that study participants could connect with.

5.2.1 Virtual Environment. The virtual agent was set up in the game engine Unity as described in the article by Schmidt et al. [56]. Besides a highly detailed, 3D scanned female head model, it featured realistic eye movements (including focusing on points of interest, saccades, and blinking) as well as lip syncing.

To enable the IVA to understand and react to user requests, several Google services were utilized. First, the user's voice was recorded using a builtin or external PC microphone and sent to Google's speech-to-text service [27] to create a transcript. Since the level of background noise was low during the study, and only one speaker was recorded at a time, the transcript was accurate for the majority of utterances. The transcript was then processed by another service called Google Dialogflow [25]. Dialogflow uses a machine learning approach to recognize a user's intent from text input and produces a matching output. Our agent entity created in Dialogflow accepted any form of greeting before automatically transitioning to the main flow. The latter was able to process 12 predefined intents that matched the topics participants had to discuss with the IVA during the study. For each of the intents, we defined a set of 3 to 7 training phrases, depending on the intent's complexity. Due to the limited number of potential matches as well as the known topic order, these low numbers of training examples were sufficient to detect the matching intent, as was confirmed in previous tests. For each intent, we defined two versions of the agent's response, which contained either precise or imprecise indications of time (see Table 7 in the supplementary material for both training questions and agent responses). The textual response was sent to Google's text-to-speech service [28] to synthesize an audio file which can be bound to an audio source within the Unity scene. Since the audio data transfer produces an additional delay, which varies between study participants and, in addition, could be perceived as unnatural, the audio files generated by the text-to-speech service were stored in a local cache before the first experimental session.

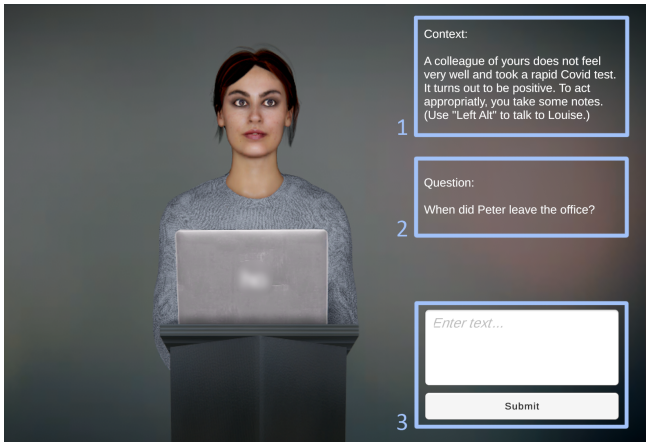

Figure 3: Virtual environment that was used in Study 3, including an IVA [17] and UI elements for (1) the given context, (2) a question directed to the user, and (3) the user's response.

Besides the IVA, UI elements for presenting the textual scenario descriptions and questions as well as for receiving the study participants' responses were displayed. The virtual environment including the IVA and UI elements is illustrated in Figure 3.

5.2.2 Hardware Setup. Due to COVID-19 regulations over the project period, we opted for a remote user study. The study was conducted on a virtual desktop machine running Windows 10 Professional. Participants of the study had to connect to the virtual machine, which was hosted on a Microsoft Azure server [44]. The virtual machine of type *Standard NV8as_v4* was equipped with an 8-core CPU, 28 GB of RAM, and a GPU with 4 GB dedicated memory. With this setup, the Unity application was running at 60 frames

per second. The application was displayed on each participant's private computer monitor.

5.3 Methods

For the main study, we followed a within-subject design with two independent variables and two levels each:

- **Time precision**
  - *Imprecise*: In its responses, the agent uses conceptual temporal locations of the granularity *Days* and *Times of Day* as well as temporal distances of the granularity *Minutes*, *Hours*, *Days*, *Weeks*, and *Months* that are rounded to the reference values determined in Study 1 (cf. Figure 2).
  - *Precise*: In its responses, the agent uses temporal distances of the granularity *Minutes*, *Hours*, and *Days* without rounding to reference values.
- **Memory aid**
  - *No*: The agent responds to questions immediately, with a natural delay of one second.
  - *Yes*: The agent communicates having notes on the referred event and looks at a virtual laptop for 3 seconds before looking back at the participant and providing a response.

For example, a temporal distance of 34 days in the precise agent response was rounded to the nearest reference point of 1 month in the corresponding imprecise response. All pairs of precise and imprecise statements are listed in Table 7 of the supplementary material.

The order of conditions was counterbalanced among participants, with a 100% coverage of order effects due to the number of conditions and participants.

Each participant was invited to a separate video call with the experimenter who provided an introduction to technical aspects and the procedure of the study. Participants were informed about the study's general purpose to evaluate the naturalness of the agent's behavior, without mentioning the focus on time or the particular hypotheses. Each participant filled in a consent form before proceeding with the study.

After the briefing, participants accessed the remote desktop environment and started the main procedure by filling out a demographic questionnaire. Afterward, they switched to the Unity environment with the IVA. For familiarization with the system, each condition started with greeting the IVA and noting down her response. This initial test run was followed by 12 inquiry-response cycles. In the beginning of each cycle, a context scenario as well as a question were displayed in textual form on the UI (see Table 7 in the supplementary material). To gather information for answering the question, participants had to talk to the IVA. Participants were allowed to pose multiple questions to the IVA, for example, to clear up a misunderstanding. For formulating their final response, participants were instructed to use numerical temporal locations (e.g., *'May 10 at 8 a.m.'*) instead of the conceptual temporal locations and temporal distances the agent used (e.g., *'1 hour ago'*). By processing the information retrieved from the agent, we anticipated that participants gain a better impression of how useful the IVA would be in a real-world scenario. The participant's final response had to be typed in and submitted via the system's UI to end the current cycle and start the next one. Each of the 12 cycles included a new question, and every four questions the context scenario changed. There was no option to jump back to a previous cycle or to skip a question.

Table 2: Means and standard deviations for anthropomorphism and usefulness scores.

	Anthropomorphism		Usefulness	
	M	**SD**	**M**	**SD**
Imprecise w/o memory aid	4.183	1.288	5.042	1.681
Precise w/o memory aid	3.158	1.247	5.250	1.675
Imprecise w/ memory aid	4.125	1.030	6.083	0.929
Precise w/ memory aid	3.992	1.141	5.792	1.769

After submitting answers to each of the 12 questions, participants were forwarded to the next questionnaire, which included a usefulness item as well as a scale to assess the IVA's anthropomorphism [4]. The questionnaire then guided participants back to the Unity environment, where they initiated the next condition. This procedure was repeated until participants experienced all four versions of the IVA.

In a final questionnaire, participants had to report their preference towards one of the four IVAs in the context of four different scenarios. They were also encouraged to provide open feedback.

During the whole study, both the participant and the experimenter were present in a video call to ensure support in case of any technical difficulties. However, after the introductory phase the camera and microphone of the experimenter as well as the camera of the participant were turned off. In total, the study took around one hour.

5.4 Results

Multiple questionnaires were used to collect both quantitative and qualitative data on how IVAs with different approaches to referring to past events are perceived by users. In the following section, we will present the results.

5.4.1 Perceived Anthropomorphism. The perceived anthropomorphism of IVAs was measured using the corresponding scale of the Godspeed questionnaire [4]. It originally consists of five items with levels ranging from 1 (e.g., *machinelike*, *unconscious*) to 5 (e.g., *humanlike*, *conscious*). Past studies indicated that the granularity of the scale might be too low to assess subtle differences between virtual agents [58]. Therefore, we increased the number of levels per item to 7. Descriptive results of the measurement are illustrated in Figure 4a.

Since mean ratings were computed by averaging five separate items, they can be considered to be interval data and, therefore,

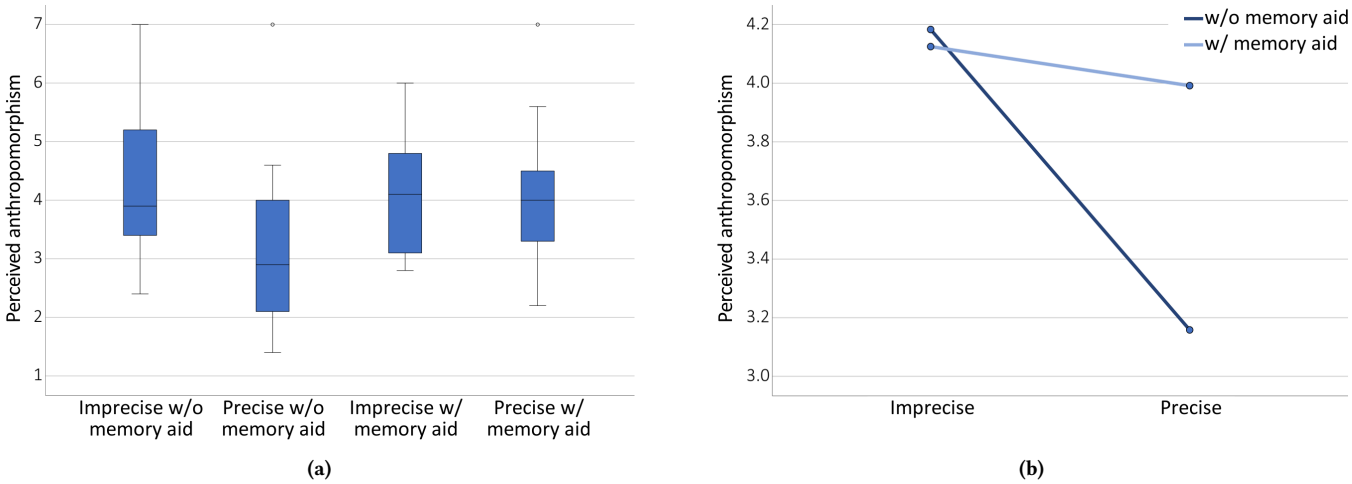

Figure 4: Scores of perceived anthropomorphism of the agent, presented (a) as a box plot, and (b) as an interaction plot. (Please note the different ranges of values for better interpretability.)

analyzed parametrically [10, 47]. We evaluated the mean anthropomorphism ratings using a two-way repeated measures ANOVA. Inspection of both histogram and Q-Q plot showed a small positive skewness (.747) and kurtosis (.389) of the residuals, however, the ANOVA has been shown to be robust against such mild deviations from the normal distribution [47].

The ANOVA revealed a significant interaction between *time precision* and *memory aid* ($F(1, 23) = 5.331$, $\mathbf{p = 0.030}$, $\eta_p^2 = 0.188$). We also found significant main effects of *time precision* ($F(1, 23) = 8.050$, $\mathbf{p = 0.006}$, $\eta_p^2 = 0.281$) and *memory aid* ($F(1, 23) = 4.498$, $\mathbf{p = 0.045}$, $\eta_p^2 = 0.164$) on perceived anthropomorphism. As main effects only have limited conclusiveness in the presence of a significant interaction, we performed a follow-up analysis of simple main effects. Sidak-adjusted comparisons indicated that for precise responses, study participants rated the anthropomorphism of IVAs using a memory aid .833 points higher than of those responding immediately ($F(1, 92) = 5.978$, $\mathbf{p = .016}$), supporting hypothesis (H3). In contrast, for imprecise responses the use of a memory aid did not yield a significant difference ($F(1, 92) = .029, p = .864$). Conversely, for IVAs that did not use a memory aid before responding, rounding of time expressions led to 1.025 points higher anthropomorphism scores compared to precise time expressions ($F(1, 92) = 9.044$, $\mathbf{p = .003}$), supporting hypothesis (H1). If a memory aid was used, ratings of agents with imprecise and precise responses did not significantly differ ($F(1, 92) = .153, p = .697$).

5.4.2 Perceived Usefulness. In contrast to the anthropomorphism score, usefulness was measured using a single Likert item and, therefore, cannot be assumed to be interval data [10]. Hence, we used a non-parametric mixed effects ordinal logistic regression to assess the ability of time precision and memory aid to predict agent usefulness. Preliminary analyses were performed to ensure that the assumptions of multicollinearity and proportional odds were not violated. Both the Pearson chi-square test and the deviance test suggested good model fit.

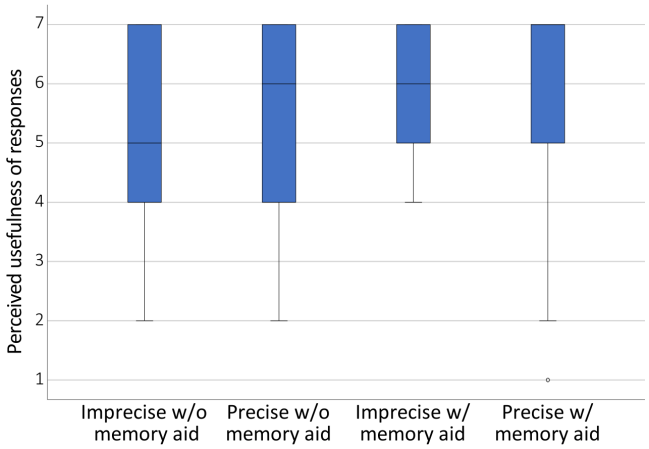

Figure 5: Scores of perceived usefulness of the agent's responses, presented as a box plot.

We found a statistically significant effect of memory aid on agent usefulness ($\chi^2(1) = 7.434$, $\mathbf{p = .006}$). The odds of agents with memory aid achieving a higher usefulness score were 2.860 times that of those who did not use a memory aid. Time precision was not a significant predictor of agent usefulness ($p = .714$), and consequently, hypothesis (H2) could not be confirmed. The data also does not support a significant interaction between time precision and memory aid ($p = .773$). The mean values and standard deviations of both usefulness and anthropomorphism scores are listed in Table 2.

5.4.3 Contextual Preference. After experiencing all four versions of the agent, participants were asked which agent they would prefer in different settings. Table 3 summarizes the responses; both the absolute counts and percentages. When IVAs act as co-workers, the majority (46%) of participants prefer an immediate response with imprecise expressions of time. For virtual receptionists and

Table 3: Absolute and relative number of participants that prefer the respective version of IVAs in different settings.

	Co-worker	Receptionist	Assistant	Friend
Imprecise w/o memory aid	11 (46%)	7 (29%)	5 (21%)	7 (29%)
Precise w/o memory aid	3 (13%)	2 (8 %)	5 (21%)	2 (8 %)
Imprecise w/ memory aid	7 (29%)	8 (33%)	6 (25%)	9 (38%)
Precise w/ memory aid	3 (13%)	7 (29%)	8 (33%)	6 (25%)

friends, 33% and 38% of the participants, respectively, would favor imprecise values that were looked up using a memory aid such as a calendar. Finally, from a virtual assistant, 33% of the participants expected precise responses using a memory aid.

5.5 Discussion

5.5.1 Perceived Anthropomorphism. In Study 3, we found positive evidence for the hypotheses (H1) and (H3). Data analysis revealed a spreading interaction between the factors *time precision* and *memory aid*, as illustrated in Figure 4b. As supported by a subsequent analysis of simple main effects, perceived anthropomorphism is indeed higher for imprecise time references than for precise ones, if no additional memory aid is used (H1). Furthermore, if precise indications of time are made, participants rated IVAs with a memory aid higher than those who responded immediately (H3). Therefore, if providing a natural conversation is the main intention of an IVA, rounding of time values should be preferred. However, if the context of an IVA necessitates high precision, a memory aid could be used to avoid a significant reduction of the agent's anthropomorphism. It has to be mentioned, though, that the latter option should be used only occasionally. In the post-questionnaire, one participant reported that the frequent use of a memory aid and the consequent delay of the agent's response was perceived as increasingly annoying throughout the conversation.

5.5.2 Perceived Usefulness. The second hypothesis (H2) could not be confirmed, since mean values for perceived usefulness of the agent are almost the same for precise and imprecise time references. One could argue that study participants could not recognize rounded values as such. However, each block of questions contained multiple vague time references such as *'yesterday morning'* and most rounded time references was preceded by an adverb such as *'around'* or *'roughly'*. Indications of a different reason for the missing effect of precision on usefulness can be found in the open feedback provided by the study participants. In the post-questionnaire, two participants expressed that they would have preferred an absolute time format for the agent's responses. While no explanation for these requests was provided, they could be related to the task of the study. At the end of each of the overall 48 cycles, participants were required to convert the agent's responses from conceptual temporal locations or temporal distances to numerical temporal locations. As this step is more difficult for precise values (e.g., *current time - 57 minutes*) than for rounded values (e.g., *current time - 1 hour*), users may have perceived the precise version of the agent as less useful to fulfill the task. While this aspect could be adapted in a future study, it also provides interesting insights into the usage of IVAs in a real scenario. While precise time values contain more information, they may result in a higher cognitive load for users. Therefore, it has to be carefully considered whether the current context of the IVA and the type of provided information require a precise response. In that case, the agent might switch to an absolute numerical time format. Furthermore, since the amount of cognitive load required for processing conceptual temporal locations and temporal distances varies, follow-up studies could consider these different types of imprecise statements separately.

While no difference between usefulness of precise and imprecise agents was present, IVAs that were checking their notes were rated as significantly more useful than those who answered immediately. This is interesting, since the objective information content was the same for both cases. With regard to analogous situations with human communication partners, it can be hypothesized that looking up a requested piece of information can convey a higher impression of reliability than recalling it from memory (cf. Section 2.1.3). In addition, immediately provided, overly precise time references for events in the remote past (e.g., *'34 days ago'*) may be perceived as fabricated answers since it is rather uncommon to receive correct time references in such a format in real conversations (according to Thompson [62] the dating error would be around 5 days for such a time interval). Before implementing the suggested behavior in a real IVA, further studies with different response delays are advisable to determine the optimal trade-off between naturalness and usefulness of the agent. As the artificial delay adds up with repeated inquiries, it should be chosen to be as small as possible while preserving the positive effects on perceived anthropomorphism.

5.5.3 Contextual Preference. Despite the significant effects in subjective ratings of agent anthropomorphism, we found mixed preferences regarding the assessed response behaviors of agents. These differences could reflect individual weightings of the agent's qualities, such as its human likeness and usefulness. These weights are not only differing between users but also between contexts. For example, most participants (75%, 62%, and 67%, respectively) would prefer more natural, imprecise time expressions when the IVA appears as a co-worker, receptionist, or friend. In contrast, a slim majority of participants (54%) would expect a virtual assistant to provide precise responses, and a majority of them would even tolerate higher response times due to the utilization of a memory aid such a calendar. Reasons for this tendency could be previous experiences with real assistants, who usually organize appointments in a structured way [18], as well as the desire of users to get the most valuable support when it comes to their personal time management.

5.6 Limitations

Study 3 was designed as a first attempt to gauge the potential of transferring human time referencing behavior to a conversational

IVA based on a specific scenario and, as such, leaves room for further investigations into specific sub-aspects. The study simulates a typical office setting where the IVA acts as an assistant to the user. This scenario implies some factors that may affect the perception of the temporal patterns used. First, Section 5.5.3 indicated that precise language might be more appropriate in such formal scenarios than in informal situations, such as conversations between friends. Therefore, the perceived believability of the IVA might be influenced by the context of the human-agent interaction. In addition, perceived traits of the agent itself, for example, how formal it appears in terms of clothing and used vocabulary, could also influence its perception by users. The context, as well as the (perceived) role of the IVA, might also make the use of memory aids other than a digital calendar on a laptop (e.g., a smartphone or a physical calendar) more plausible.

Besides these possible variations of the setting, follow-up studies could further investigate which underlying principles account for the effectiveness of the memory aid. The current study design prohibits a definitive conclusion about which aspect of the memory aid conditions caused the observed changes in the users' perception of the IVA – the fact that the IVA was reading the provided information from an external device or the response delay, which in itself could indicate a recall operation of the IVA. However, basic literature on the interpretation of paralinguistic features by communication partners suggests that the speaker's response latency and the listener's feeling of another's knowing are negatively correlated, meaning that answers preceded by long pauses sound less likely to be correct [7]. Based on this theoretical framework, an increased response latency alone would have reduced the estimated likelihood that the IVA knew the correct answer, which contradicts the higher usefulness ratings for these conditions. A tailored user study could provide empirical evidence that this theory of human communication is also applicable to IVAs.

In terms of agent representation, Study 3 was based on an embodied IVA. However, since variations in temporal statements were only expressed via speech, they could be applied to an audio-based IVA (such as Amazon's Alexa or Apple's Siri) as well. In this case, other non-visual approaches to simulating a memory aid would need to be explored.

6 CONCLUSIONS

In this paper, we compared different approaches for IVAs to refer to the time of past events. The basis was provided by a survey ($N = 63$), which aimed at revealing common patterns that humans use in temporal statements. Through the analysis of more than 1000 responses, we could confirm previous results on the preference of humans towards vague time statements and, in addition, identify specific reference points for the rounding of temporal distances. The identified patterns were subsequently transferred to a conversational virtual agent with voice in- and output in order to assess their effectiveness with regard to the agent's believability. A user study with 24 participants was conducted to evaluate two different time precision levels as well as the usage of memory aids. Statistical analyses indicate a significant positive effect of imprecise time statements on perceived anthropomorphism of the agent. For precise statements, anthropomorphism scores were significantly increased when the agent was noticeably utilizing a memory aid before providing time-related responses. Based on these results, we suggest implementing IVAs with a human-inspired way of expressing time, even though it is at the cost of reduced information content. If precision is of major importance in a specific context and, therefore, should be kept high, the use of a memory aid such as a calendar can counteract a loss in the agent's believability.

While the intention of this work was to gain some insights into the importance of time in human-agent communication, there are several potential follow-up questions to pursue in future research projects. In the presented user studies, we focused on a general comparison between precise and imprecise references to past events. These investigations could be extended to consider more nuances of temporal precision as well as possible correlations to factors other than the used temporal granularity. With respect to the latter, research on prospective events has already shown that the use of temporal patterns may depend, among other things, on the timing of the conversation, its context, as well as implicit knowledge about the event itself [30, 54], which could similarly apply to past events. Furthermore, the preliminary survey revealed some patterns of temporal statements that were not incorporated into our categorization but still could be valuable in certain scenarios. First, we observed the rare use of adverbs of indefinite time such as *'recently'* or *'a while ago'*. While these may be perceived as unnatural or even impolite when used as direct responses to a time-related question, it might be reasonable to add them to the agent's vocabulary for dialogs that are not primarily focused on time. Second, the referencing of temporal landmarks, either in the form of global events (e.g., New Year's Eve) or the user's personal events (e.g., parental leave), could increase the naturalness of human-agent conversations even further. Finally, an extension of the survey to three additional countries revealed cross-cultural similarities in temporal patterns, for example, regarding reference points for rounding temporal distances, but also indicated cross-cultural differences in terms of the vagueness of statements and the usage of indicators for uncertainty. Therefore, more user-centered research should be carried out to guarantee a high suitability of IVAs for different people in various contexts.

ACKNOWLEDGMENTS

This work was partially funded by the German Federal Ministry of Education and Research (BMBF) and the German Federal Ministry for Economic Affairs and Climate Action (BMWK).

REFERENCES

[1] David Ahn, Sisay Fissaha Adafre, and Maarten de Rijke. 2005. Extracting Temporal Information from Open Domain Text: A Comparative Exploration. *Journal of Digital Information Management* 3, 1 (2005), 14–20.
[2] Jan Allbeck and Norman I Badler. 2001. Consistent Communication with Control. In *Proceedings of the Workshop on Non-Verbal and Verbal Communicative Acts to Achieve Contextual Embodied Agents at the Conference on Autonomous Agents (AGENTS)*. ACM, New York, NY, US, 1–6.
[3] Andreas Armborst. 2017. Thematic Proximity in Content Analysis. *Sage Open* 7, 2 (2017), 1–11.
[4] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.
[5] Timothy Bickmore, Daniel Schulman, and Langxuan Yin. 2009. Engagement vs. Deceit: Virtual Humans with Human Autobiographies. In *Proceedings of*

the International Conference on Intelligent Virtual Agents (IVA). Springer, Berlin, Heidelberg, Germany, 6–19.
[6] Richard A Block. 2003. Psychological Timing Without a Timer: The Roles of Attention and Memory. In *Time and Mind II: Information Processing Perspectives*. Hogrefe & Huber Publishers, Göttingen, Germany, 41–59.
[7] Susan E Brennan and Maurice Williams. 1995. The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners About the Metacognitive States of Speakers. *Journal of Memory and Language* 34, 3 (1995), 383–398.
[8] Stephen Buetow and Katharine Wallis. 2019. The Beauty in Perfect Imperfection. *Journal of Medical Humanities* 40, 3 (2019), 389–394.
[9] Ondřej Burkert, Cyril Brom, Rudolf Kadlec, and Jiří Lukavskỳ. 2010. Timing in Episodic Memory: Virtual Characters in Action. In *Remembering Who We Are – Human Memory for Artificial Agents Symposium*. De Montfort University, Leicester, UK, 1–8.
[10] James Carifio and Rocco Perla. 2008. Resolving the 50-Year Debate Around Using and Misusing Likert Scales. *Medical Education* 42, 12 (2008), 1150–1152.
[11] W Crawford. 2008. Place and Time Adverbials in Native and Non-Native English Student Writing. In *Corpora and Discourse: The Challenges of Different Settings*. John Benjamins Publishing Company, Amsterdam, Netherlands, 267–289.
[12] David Crystal. 1966. Specification and English Tenses. *Journal of Linguistics* 2, 1 (1966), 1–34.
[13] DeepL. 2021. DeepL Translate: The World's Most Accurate Translator. https://www.deepl.com/en/translator. Accessed: 2021-12-29.
[14] Virginie Demeure, Radosław Niewiadomski, and Catherine Pelachaud. 2011. HHow Is Believability of a Virtual Agent Related to Warmth, Competence, Personification, and Embodiment? *Presence* 20, 5 (2011), 431–448.
[15] Joao Dias, Wan Ching Ho, Thurid Vogt, Nathalie Beeckman, Ana Paiva, and Elisabeth André. 2007. I Know What I Did Last Summer: Autobiographic Memory in Synthetic Characters. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. Springer, Berlin, Heidelberg, Germany, 606–617.
[16] Alexander Eiselmayer, Chat Wacharamanotham, Michel Beaudouin-Lafon, and Wendy E. Mackay. 2019. Touchstone2: An Interactive Environment for Exploring Trade-Offs in HCI Experiment Design. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, US, 1–11.
[17] Eisko. 2018. Animatable Digital Double of Louise by Eisko©. www.eisko.com
[18] Thomas Erickson, Catalina M. Danis, Wendy A. Kellogg, and Mary E. Helander. 2008. Assistance: The Work Practices of Human Administrative Assistants and Their Implications for IT and Organizations. In *Proceedings of the Conference on Computer Supported Cooperative Work (CSCW)*. ACM, New York, NY, US, 609–618.
[19] Lisa Ferro. 2001. *TIDES Instruction Manual for the Annotation of Temporal Expressions*. Technical Report. The MITRE Corporation.
[20] W. Friedman. 2005. Developmental and Cognitive Perspectives on Humans' Sense of the Times of Past and Future Events. *Learning and Motivation* 36 (2005), 145–158.
[21] William J Friedman. 2004. Time in Autobiographical Memory. *Social Cognition* 22, 5: Special issue (2004), 591–605.
[22] Wael H Gomaa and Aly A Fahmy. 2013. A Survey of Text Similarity Approaches. *international Journal of Computer Applications* 68, 13 (2013), 13–18.
[23] P Gomes, Ana Paiva, and Carlos Martinho. 2010. Episodic Memory Retrieval Through Re-appraisal of Past Events. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA) - Supplementary Materials*. Springer, Philadelphia, US, 26–27.
[24] Paulo F Gomes, Carlos Martinho, and Ana Paiva. 2011. I've Been Here Before! Location and Appraisal in Memory Retrieval. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS) - Volume 3*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, US, 1039–1046.
[25] Google. 2021. Dialogflow | Google Cloud. cloud.google.com/dialogflow/
[26] Google. 2021. Google Translate. https://translate.google.com/. Accessed: 2021-12-29.
[27] Google. 2021. Speech-to-Text: Automatic Speech Recognition | Google Cloud. cloud.google.com/speech-to-text/
[28] Google. 2021. Text-to-Speech: Lifelike Speech Synthesis | Google Cloud. cloud.google.com/text-to-speech/
[29] Robert J Graham. 1981. The Role of Perception of Time in Consumer Research. *Journal of Consumer Research* 7, 4 (1981), 335–342.
[30] David Graus, Paul N Bennett, Ryen W White, and Eric Horvitz. 2016. Analyzing and Predicting Task Reminders. In *Proceedings of the Conference on User Modeling Adaptation and Personalization (UMAP)*. ACM, New York, NY, US, 7–15.
[31] Simon Grondin. 2010. Timing and Time Perception: A Review of Recent Behavioral and Neuroscience Findings and Theoretical Directions. *Attention, Perception, & Psychophysics* 72, 3 (2010), 561–582.
[32] Charles Hampden-Turner, Fons Trompenaars, and Charles Hampden-Turner. 2020. *Riding the Waves of Culture: Understanding Diversity in Global Business*. Nicholas Brealey Publishing, Boston, MA, US.
[33] Martin Haspelmath. 1997. *From Space to Time*. Lincom Europa, München, Germany, Newcastle, UK.
[34] Wan Ching Ho and Kerstin Dautenhahn. 2008. Towards a Narrative Mind: The Creation of Coherent Life Stories for Believable Virtual Agents. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*. Springer, Berlin, Heidelberg, Germany, 59–72.
[35] Margaret J Intons-Peterson and JoAnne Fournier. 1986. External and Internal Memory Aids: When and How Often Do We Use Them? *Journal of Experimental Psychology: General* 115, 3 (1986), 267.
[36] Margaret Jean Intons-Peterson and George L Newsome. 1992. External Memory Aids: Effects and Effectiveness. In *Memory Improvement*. Springer, New York, NY, US, 101–121.
[37] Steve MJ Janssen, Antonio G Chessa, and Jaap MJ Murre. 2006. Memory for Time: How People Date Events. *Memory & Cognition* 34, 1 (2006), 138–147.
[38] Ryota Kanai, Mia Yuan Dong, Bahador Bahrami, and Geraint Rees. 2011. Distractibility in Daily Life Is Reflected in the Structure and Function of Human Parietal Cortex. *Journal of Neuroscience* 31, 18 (2011), 6620–6626.
[39] Zerrin Kasap, Maher Ben Moussa, Parag Chaudhuri, and Nadia Magnenat-Thalmann. 2009. Making Them Remember—Emotional Virtual Characters With Memory. *Computer Graphics and Applications* 29, 2 (2009), 20–29.
[40] Charlyn M Laserna, Yi-Tai Seih, and James W Pennebaker. 2014. Um... Who Like Says You Know: Filler Word Use as a Function of Age, Gender, and Personality. *Journal of Language and Social Psychology* 33, 3 (2014), 328–338.
[41] Douglas B Lenat. 2009. Building a Machine Smart Enough to Pass the Turing Test. In *Parsing the Turing Test*. Springer, Dordrecht, Netherlands, 261–282.
[42] Willem JM Levelt, Graham Richardson, and Wido La Heij. 1985. Pointing and Voicing in Deictic Expressions. *Journal of Memory and Language* 24, 2 (1985), 133–164.
[43] Robert V Levine and Ara Norenzayan. 1999. The Pace of Life in 31 Countries. *Journal of Cross-Cultural Psychology* 30, 2 (1999), 178–205.
[44] Microsoft. 2021. Windows Virtual Desktop | Remote Desktop | Microsoft Azure. azure.microsoft.com/en-us/services/virtual-desktop/
[45] George A Miller. 1995. WordNet: A Lexical Database for English. *Commun. ACM* 38, 11 (1995), 39–41.
[46] Miro. 2021. The Visual Collaboration Platform for Every Team: Miro. https://miro.com/. Accessed: 2021-12-29.
[47] Geoff Norman. 2010. Likert Scales, Levels of Measurement and the "Laws" of Statistics. *Advances in Health Sciences Education* 15, 5 (2010), 625–632.
[48] Nahal Norouzi, Gerd Bruder, Brandon Belna, Stefanie Mutter, Damla Turgut, and Greg Welch. 2019. A Systematic Review of the Convergence of Augmented Reality, Intelligent Virtual Agents, and the Internet of Things. In *Artificial Intelligence in IoT*. Springer, Cham, 1–24.
[49] Nahal Norouzi, Kangsoo Kim, Jason Hochreiter, Myungho Lee, Salam Daher, Gerd Bruder, and Greg Welch. 2018. A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In *Proceedings of the International Conference on Intelligent Virtual Agents (IVA)*. ACM, New York, NY, US, 17–22.
[50] Rafael E Núñez, Benjamin A Motz, and Ursina Teuscher. 2006. Time After Time: The Psychological Reality of the Ego-and Time-Reference-Point Distinction in Metaphorical Construals of Time. *Metaphor and Symbol* 21, 3 (2006), 133–146.
[51] Asher R Pacht. 1984. Reflections on Perfection. *American Psychologist* 39, 4 (1984), 386.
[52] Deborah Richards and Karla Bransky. 2014. Forgetmenot: What and How Users Expect Intelligent Virtual Agents to Recall and Forget Personal Conversational Content. *International Journal of Human-Computer Studies* 72, 5 (2014), 460–476.
[53] José I Rojas-Méndez, Gary Davies, Omer Omer, Paitoon Chetthamrongchai, and Canan Madran. 2002. A Time Attitude Scale for Cross Cultural Research. *Journal of Global Marketing* 15, 3-4 (2002), 117–147.
[54] Xin Rong, Adam Fourney, Robin N Brewer, Meredith Ringel Morris, and Paul N Bennett. 2017. Managing Uncertainty in Time Expressions for Virtual Assistants. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, US, 568–579.
[55] Roser Saurı, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML Annotation Guidelines Version 1.2.1.
[56] Susanne Schmidt, Oscar Ariza, and Frank Steinicke. 2020. Intelligent Blended Agents: Reality–Virtuality Interaction with Artificially Intelligent Embodied Virtual Humans. *Multimodal Technologies and Interaction* 4, 4 (2020), 85.
[57] Susanne Schmidt, Gerd Bruder, and Frank Steinicke. 2018. Effects of Embodiment on Generic and Content-Specific Intelligent Virtual Agents as Exhibition Guides. In *Proceedings of the International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (ICAT-EGVE)*. Eurographics Association, Geneve, Switzerland, 13–20.
[58] Susanne Schmidt, Oscar Javier Ariza Nunez, and Frank Steinicke. 2019. Blended Agents: Manipulation of Physical Objects within Mixed Reality Environments and Beyond. In *Proceedings of the Symposium on Spatial User Interaction (SUI)*. ACM, New York, NY, US, 1–10.

[59] Anna Sircova, Fons JR Van De Vijver, Evgeny Osin, Taciano L Milfont, Nicolas Fieulaine, Altinay Kislali-Erginbilgic, Philip G Zimbardo, et al. 2015. Time Perspective Profiles of Cultures. In *Time Perspective Theory; Review, Research and Application*. Springer, Cham, Switzerland, 169–187.

[60] Carlota S Smith. 1978. The Syntax and Interpretation of Temporal Expressions in English. *Linguistics and Philosophy* 2, 1 (1978), 43–99.

[61] SurveyCircle. 2021. The Largest Community for Online Research. www.surveycircle.com/

[62] Charles P Thompson. 1982. Memory for Unique Personal Events: The Roommate Study. *Memory & Cognition* 10, 4 (1982), 324–332.

[63] Charles P Thompson, John J Skowronski, and D John Lee. 1988. Telescoping in Dating Naturally Occurring Events. *Memory & Cognition* 16, 5 (1988), 461–468.

[64] Catherine Tinsley. 1998. Models of Conflict Resolution in Japanese, German, and American Cultures. *Journal of Applied Psychology* 83, 2 (1998), 316.

[65] Hegler Tissot, Marcos Didonet Del Fabro, Leon Derczynski, and Angus Roberts. 2019. Normalisation of Imprecise Temporal Expressions Extracted From Text. *Knowledge and Information Systems* 61, 3 (2019), 1361–1394.

[66] Jean-Claude G Usunier. 1991. Business Time Perceptions and National Cultures: A Comparative Survey. *MIR: Management International Review* 31, 3 (1991), 197–217.

[67] Bart Van Den Hooff. 2004. Electronic Coordination and Collective Action: Use and Effects of Electronic Calendaring and Scheduling. *Information & Management* 42, 1 (2004), 103–114.

[68] Isaac Wang, Jesse Smith, and Jaime Ruiz. 2019. Exploring Virtual Agents for Augmented Reality. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. ACM, New York, NY, US, 1–12.

[69] Dan Zakay and Richard A Block. 2004. Prospective and Retrospective Duration Judgments: An Executive-Control Perspective. *Acta Neurobiologiae Experimentalis* 64, 3 (2004), 319–328.

[70] Gal Zauberman, B Kyu Kim, Selin A Malkoc, and James R Bettman. 2009. Discounting Time and Time Discounting: Subjective Time Perception and Intertemporal Preferences. *Journal of Marketing Research* 46, 4 (2009), 543–556.

[71] Qing Zhou and Richard Fikes. 2002. A Reusable Time Ontology. In *Proceeding of the AAAI Workshop on Ontologies for the Semantic Web*. The AAAI Press, Menlo Park, CA, US, 35–40.

[72] XiaoJia Zhou, HaoMin Li, XuDong Lu, and HuiLong Duan. 2011. Temporal Expression Recognition and Temporal Relationship Extraction from Chinese Narrative Medical Records. In *Proceedings of the International Conference on Bioinformatics and Biomedical Engineering (ICBBE)*. IEEE, New York, NY, US, 1–4.