DOI: 10.14515/monitoring.2021.1.1793


N. Klowait, M. A. Erofeeva

THE RISE OF INTERACTIONAL MULTIMODALITY IN HUMAN-COMPUTER INTERACTION

For citation:
Klowait N., Erofeeva M. A. (2021) The Rise of Interactional Multimodality in Human-Computer Interaction. *Monitoring of Public Opinion: Economic and Social Changes.* No. 1. P. 46–70. https://doi.org/10.14515/monitoring.2021.1.1793.
Правильная ссылка на статью:
Кловайт Н., Ерофеева М. А. Мультимодальный поворот в исследованиях взаимодействия человека и компьютера // Мониторинг общественного мнения: экономические и социальные перемены. 2021. № 1. С. 46—70. https://doi.org/10.14515/monitoring.2021.1.1793. (In Eng.)

DOI: 10.14515/monitoring.2021.1.1793

THE RISE OF INTERACTIONAL MULTIMODALITY IN HUMAN-COMPUTER INTERACTION

Nils KLOWAIT[1,2,3] — Research Fellow at the Center for Innovative Social Research; Senior Research Fellow at the International Center for Contemporary Sociological Theory; Researcher
E-MAIL: nils.klowait@gmail.com
https://orcid.org/0000-0002-7347-099X

Maria A. EROFEEVA[1,2,3] — Cand. Sci. (Soc.), Researcher at the Center for Sociological Research; Senior Research Fellow at the International Center for Contemporary Sociological Theory; Researcher
E-MAIL: erofeeva-ma@universitas.ru
https://orcid.org/0000-0002-0874-5272

[1] The Russian Presidential Academy of National Economy and Public Administration, Moscow, Russia
[2] The Moscow School of Social and Economic Sciences, Moscow, Russia
[3] The Sber Gamification Lab, Moscow, Russia

Abstract. The field of human-computer interaction (HCI) investigates the intersection between the design of devices and user practices. From an early focus on interaction modeling based on psychological experiments, the field has since experienced a shift towards the study of how actual humans interact with autonomous devices. The field became conductive to ethnographic,

МУЛЬТИМОДАЛЬНЫЙ ПОВОРОТ В ИССЛЕДОВАНИЯХ ВЗАИМОДЕЙСТВИЯ ЧЕЛОВЕКА И КОМПЬЮТЕРА

КЛОВАЙТ Нильс — научный сотрудник Центра перспективных социальных исследований, Российская академия народного хозяйства и государственной службы при Президенте РФ, Москва, Россия; старший научный сотрудник Международного центра современной социологической теории, Московская высшая школа социальных и экономических наук, Москва, Россия; эксперт, Лаборатория геймификации Сбера, Москва, Россия
E-MAIL: nils.klowait@gmail.com
https://orcid.org/0000-0002-7347-099X

ЕРОФЕЕВА Мария Александровна — кандидат социологических наук, научный сотрудник Центра социологических исследований, Российская академия народного хозяйства и государственной службы при Президенте РФ, Москва, Россия; старший научный сотрудник Международного центра современной социологической теории, Московская высшая школа социальных и экономических наук, Москва, Россия; эксперт, Лаборатория геймификации Сбера, Москва, Россия
E-MAIL: erofeeva-ma@universitas.ru
https://orcid.org/0000-0002-0874-5272

Аннотация. Предметом исследований взаимодействия человека и компьютера является взаимосвязь между дизайном устройств и пользовательскими практиками. Изначально взаимодействие «человек — компьютер» моделировалось на основании психологических экспериментов, однако со временем в этой области произошел сдвиг в сторону изучения того, как

DOI: 10.14515/monitoring.2021.1.1793

observational and videographic studies of human-device interaction. Conversation-analytic HCI became possible. That said, this new wave of researchers was never truly able to dethrone the psychological common sense of the field. With recent developments in both the technical-sensorial capabilities and outward actuational range of embodied virtual agents, the field of HCI has once again returned to the question of the sequential unfolding of the interaction between users and intelligent agents, and the multimodal interactional repertoire that is deployed throughout. This review will highlight the situational orientation of high-impact research in the field, and relate it to the cotemporaneous development of ethnomethodological and conversation analytic frameworks.

Keywords: human-computer interaction, embodied conversational agents, conversation analysis, multimodality, interactive resources

Acknowledgments. The article was prepared in the framework of a research grant funded by the Ministry of Science and Higher Education of the Russian Federation (grant ID: 075-15-2020-908). The article was prepared in cooperation with the Sber (ex.-Sberbank) Gamification Lab.

люди используют автономные устройства в реальных условиях. Начали проводиться качественные исследования (этнография, включенное наблюдение, видеоанализ), основанные в том числе на методологии конверсационного анализа. Благодаря недавним разработкам в области сенсорных и коммуникативных способностей воплощенных виртуальных агентов (аватаров) исследователи вернулись к вопросу о последовательном развертывании взаимодействия между пользователями и аватарами, а также о доступном мультимодальном интерактивном репертуаре последних. В статье анализируются актуальные направления исследований взаимодействия человека и компьютера в контексте развития конверсационного анализа.

Ключевые слова: взаимодействие человека и компьютера, воплощенные разговорные агенты, конверсационный анализ, мультимодальность, интерактивные ресурсы

Благодарность. Статья подготовлена в рамках гранта, предоставленного Министерством науки и высшего образования Российской Федерации (№ соглашения о предоставлении гранта: 075-15-2020-908). Статья подготовлена в сотрудничестве с Лабораторией геймификации Сбера.

Introduction

The field of human-computer interaction (HCI) is a strongly multidisciplinary endeavor that focuses on questions surrounding the engagement of humans with various kinds of interfaces. HCI has historically been the domain of engineering and computer science. As such, it has always had an eminently practical concern with the design of useable, functional, properly integrated technologies. That said, over its brief history, it has seen a number of theoretical interventions by various disciplines.

DOI: 10.14515/monitoring.2021.1.1793

One of these interventions occurred in the mid-1990s: the discipline of *conversation analysis* then attempted, with only moderate success, to establish itself as a resource for the study of naturally-occurring device use, and as a toolset for the development of more humanlike interactional technologies [see Klowait, 2018b]. This review will investigate both the reasons for the lack of universal adoption of conversation analysis as the go-to conceptual toolset for HCI, as well as the emerging trends that point to an upcoming conversation-analytic renaissance in the field.

Conversation Analysis

In the middle of the 20th century, a new approach to the analysis of human interaction emerged: conversation analysis (CA). Spearheaded by thinkers like H. Sacks, G. Jefferson, and E. A. Schegloff, it expanded its progenitors — E. Goffman and H. Garfinkel — in two substantial ways. On the one hand, it made the profound — yet unsystematic — insights of Goffman commensurable with a methodologically robust framework of multi-generational analytic practice. It moved pointedly away from newspaper vignettes, implicitly known cultural codes, and 'illustrative examples', instead focusing on gathering and analyzing naturalistic (i. e. unstaged, unrehearsed, unscripted) data in the form of audio/video recordings.

CA equally moved away from Garfinkel's early focus on the irreducible indexicality of the single situation, making first attempts to gain generalized systematic insights about typical structures employed by interactants for practical purposes: how hotline operators accomplish the smooth conclusion of a call, how speakers project epistemic authority, how interactants refuse birthday party invitations without major disruptions to conversational flow [for an introduction see Stokoe, 2018]. Regardless of the concrete interaction type (face-to-face dyads, teleconferencing, multi-party lecture environments, etc.) or modality (talk, prosody, non-verbal, etc.), CA's focus remained squarely on the participants' methods of getting things done, and the multitude of resources employed for this purpose.

The resources, meanwhile, turned out to be unexpectedly minute and subtle: for example, we habitually display an orientation to notable silences during our interactions with others. Imagine a two-person conversation, where the first speaker asks, 'Will you come to my birthday party?', followed by a response by the second speaker. If there is even a one-second silence between question and response, the first speaker may already expect a rejection. The notable silence, in other words, is something all parties orient to — we know that everybody knows that a long silence is problematic here. Moreover, there is something more problematic about rejecting a birthday invitation than accepting it, and all parties tend to be aware of this, making interactional elements like silences, coughs, gaze shifts, interruptions, key elements of analytic focus for CA. Imagine a situation where a birthday invitation is followed by an immediate 'no', without flourish or elaboration; the general fact that this hypothetical is extraordinary hints at the implicit normative structuring going on here.

A classic example from early conversion analysis illustrates how speakers appropriate the implicit norms of conversation to achieve their practical goals all the while orienting to these very norms [Sacks, Jefferson, 1992: 3]:

DOI: 10.14515/monitoring.2021.1.1793

(2)[1]
A: This is Mr. Smith may I help you
B: Yes, this is Mr. Brown
(3)
A: This is Mr. Smith may I help you
B: I can't hear you.
A: This is Mr. Smith.
B: Smith.

The above are transcripts from different telephone conversations between a client and an emergency psychiatric hospital; A is the staff member, B is the client. Suppose that (2) and (3) are conversational sequences that frequently appear in the very beginning of these kinds of phone calls. Suppose further that this is the first interaction between A and B. If one were to create an interpretational gloss of (2) and (3), respectively, it would likely be something akin to 'In (2), A and B introduce themselves to each other' and 'In (3), there seems to be a problem with the connection that prevents B from hearing A's introduction'. With such a gloss, nothing truly remarkable seems to occur.

However, if — upon noticing that, in (3), B's half of the introduction does not occur — it is possible to view (3) as an example of an **intentional** interactional device employed by B to avoid a name-based introduction on their part. We can ask ourselves: *apart from its explicit meaning, what does 'I can't hear you' achieve, interactionally?* We could then ask ourselves *how* this sequence is employed and what it says about the nature of conversation in general. Sacks [Sacks, Jefferson, 1992] suggests that (3) is an example of a sequence where B gracefully avoids giving up his name, whilst not being outright confrontational (e. g. 'I don't want to tell you my name', silence, etc.). A repeats their introduction, yet B's previously expected complementary contribution can now be less problematically skipped through the following confirmation sequence.

The field of conversation analysis is incredibly rich, and this brief paragraph cannot do Sacks' analysis justice. For the purposes of this introduction, what matters is that Sacks, through this shift of interpretation, laid the foundation for the analysis of interaction of hitherto unheard-of granularity.

Compared with earlier microsociological theories, CA's shift from the unique to the typical, repeatable, and comparable made a number of advancements possible. First of all, scholars could now specialize on a class of interactional phenomena. For example, following Goffman's [1981] writing on response cries, Heritage [1984; 2016] spent the last three decades investigating 'oh' as a change-of-state token, i. e. a way whereby interactants publicly produce a purportedly internal change of knowledge (e. g. 'Oh, I see'). The relative stability of this analytic focus, in turn, made it possible to attempt a longitudinal investigation of changes in language use. Couper-Kuhlen [2019], for instance, recently started to investigate how 'oh' has changed in its use over time.

Secondly, repeatable interactions invite the possibility of formalization and application. If we, for instance, know how people manage to hang up politely, why not use that insight to provide better customer service? If a particular turn of phrase contributes

[1] Transcript numbering preserved from the original.

DOI: 10.14515/monitoring.2021.1.1793

ammunition for adversarial conversational developments, why not discourage its use in institutional settings? If name-based introductions invite 'I can't hear you'-type exchanges, perhaps these introductions can be dropped for certain types of calls? Thus, the comparative rigor and systematicity of contemporary CA has contributed to a number of voyages into multidisciplinary fields of application.

The following section will introduce the first such intervention into the field of HCI, starting with an account of the assumptions common to the field prior to this intervention.

The Ethnographic Turn in HCI: The Two Waves of Conversational Intervention

It would not be correct to call HCI a traditional disciplinary field since it focused on a specific object from its very inception. However, engineering and computer science were traditionally allied disciplines, to which psychology soon joined. The latter managed to cement itself as a key player in most in-field discussions. While there are many reasons for this, one key reason is arguably the introduction of the *Model Human Processor* as a ready-made metaphor for the study of human-interface interactions [Card, Moran, Newell, 1983].

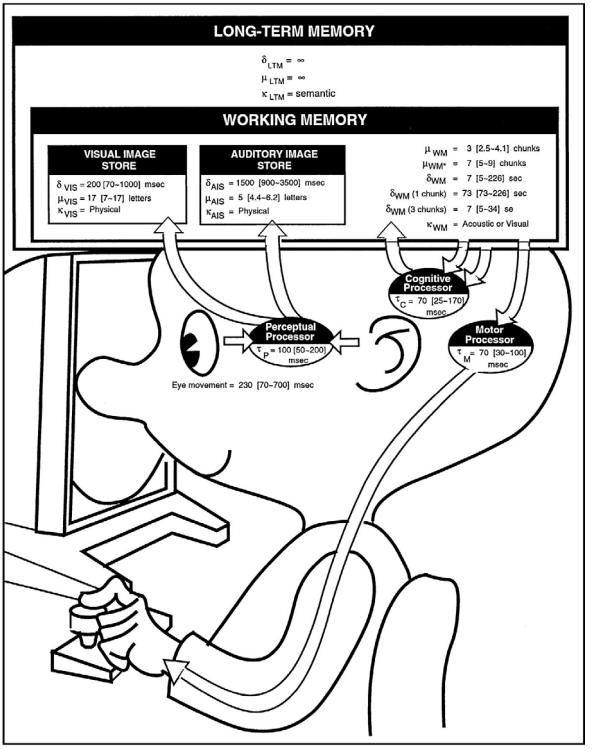

Fig. 1. The Model Human Processor[2]

[2] Source [Card et al., 1983: 26].

DOI: 10.14515/monitoring.2021.1.1793

In the Model Human Processor, one could conceive of human-computer interaction as an integrated computational system, where the human's capacity to (re-)act, process and memorize could be quantitatively assessed in terms of its enmeshedness with a specific interface (see fig. 1). While this general approach may appear somewhat simplistic, its potential applications proved to be quite productive for the field. This is primarily because it, itself, functioned as an interface between the myriad findings of contemporary cognitive psychology and matters of interface design. This intersection could be readily investigated in psychological research laboratories all over the world.

The Ethnographic Turn

Over time, however, a number of researchers of a more sociological persuasion made their forays into this field. Contrary to the psychological take on HCI, sociologists, most notably L. Suchman, argued that there is a key difference between a model of human-interface interaction and the way it unfolded in actuality [Suchman, 1987]. The issue was not so much that the model of human interaction was flawed — and should therefore be refined or replaced with a better model — but that the very idea of modeling human action as a processing sequence did not account for the inherently procedural nature of human action. Adopting insights from ethnomethodology [Garfinkel, 1967], thinkers like Suchman stressed how humans interact through a process of procedural reinterpretation and reassessment of what is going on. These processes of actual in-situ interactions could not be expressed in sequential models; they had to be studied as they unfolded.

This shift in focus contributed to what may be called the ethnographic turn in HCI [for review see Carroll, 2010; Klowait, 2018a]. While laboratory-based studies of human-interface interactions still had their place, they were now more readily expanded by observation-based investigations of the actually unfolding interaction, and the radically contingent systematicities to be found therein.

This ethnographic turn, notably, was not an ethnomethodological turn. Although Suchman [1987] imported some conversation-analytic insights, these insights came at the price of simplification. The distinction between *plans* and *situated actions* allowed future researchers to argue that *we have to see how interaction* ***actually*** *unfolds*, but did not entrench ethnomethodology *specifically* as the methodology of choice.

Sociological Bugfixing

Moreover, Suchman's work, while certainly making a good case for ethnographic investigation of the device's actual-situated-use-practices, is not incompatible with a fairly regular iterative design cycle (see fig. 2).

A device was created based on plans, prototyped in a certain number of expected scenarios, and deployed. Upon evaluating the performance of the device in the field, changes to the design were implemented in the plan for the next cycle of deployment.

In other words, the ethnographic turn could very easily be integrated as a kind-of social bugfixing: sociologists would iron out quirks that emerged at the messy and difficult-to-predict stage of human-device interaction, by observing and cataloguing 'errors' to be fixed at later stages. Thus, the ethnographic turn partially became palatable to a much less radical paradigm of human-computer interaction. While it gave

DOI: 10.14515/monitoring.2021.1.1793

ample job opportunities to aspiring ethnographers of technology, it remained a far cry from a wholesale adoption of conversation-analytic principles. An attempt to go beyond such first forays was attempted in the early 1990s.

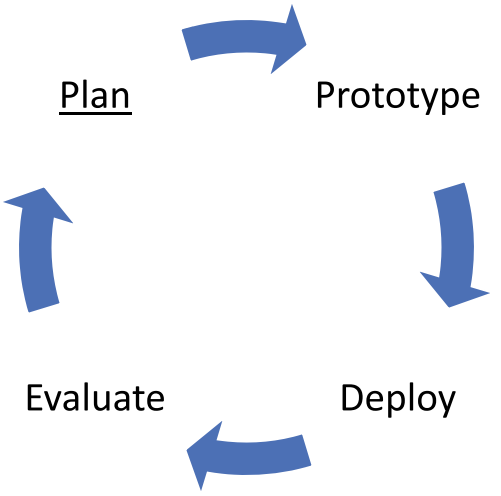

Fig. 2. Iterative design cycle [3]

The Second Wave of Conversation Analysis in HCI

In 1990, Luff et. al published a collection of articles in a volume entitled *Computers and Conversation* [Luff, Gilbert, Frohlich, 1990] detailing their vision for a conversation-analytic school of human-computer interaction [Frohlich, Luff, 1990: 187—188, emphasis ours]:

> *Sacks suggests a programme of analysis of the details of individual sequences of talk, directed towards the discovery of what he calls a technology of conversation; a kind of common machinery and know-how for the manufacture of conversational contributions within a particular culture.* ***The direction of the process here is from phenomena to technology****.*
>
> *Our own programme has been one in which we have attempted to apply Sack's programme of analysis in reverse. That is, we have taken various aspects of the technology of conversation as it has been described in the CA literature and tried to build them into the operation of a computer in such a way as to generate and support orderly sequences of talk.* ***The direction of the process here is from technology to phenomena****.*

In other words: conversation analysis reveals the rules of interaction. These rules can now be integrated into our non-human systems.

[3] For a formal treatment of iterative design, see [Boehm, 1988].

DOI: 10.14515/monitoring.2021.1.1793

If successful, this attempt would establish a firm stronghold of conversation analysts in HCI. After all, the field of CA was still relatively young and would go on to consistently generate novel findings about the structures of social action on the microscale: Sacks' legacy made it possible to see order in the most mundane snippets of recorded interaction, which in turn could generate insights about how to set up more effective computer interfaces. CA was, at least on its promotional leaflets, a powerful new microscope that could reveal a hitherto unavailable actual social reality. It could make a case for its systematicity and compatibility with the kind of systematicities handled by traditional HCI researchers.

Button's Rebuttal

Sacks' [1984] own writing seemed to generate support for the kind of project envisioned by most of the authors of *Computers and Conversation* [Luff et al., 1990]. He argues [Sacks, 1984: 413, emphasis ours]:

> *The gross aim of the work I am doing is to see how finely the details of actual, naturally occurring conversation can be subjected to analysis that will yield the technology of conversation. The idea is to take singular sequences of conversation and tear them apart in such a way as to find rules, techniques, procedures, methods, maxims… **that can be used to generate** the orderly features we find in the conversations we examine.*

Not only is he talking about a technology of conversation, but he explicitly talks about how these technologies can actively generate observed social order. As such, one might get the impression that CA would generally welcome this newfound field where its findings could find practical application. As it turned out, this was far from the case.

At the forefront of the resistance against this kind of understanding of CA was, ironically enough, a chapter of the very volume that proposed this epistemic intervention in the first place [Button, 1990: 84, emphasis ours].

> *Conversation analysis displays: that the rules involved in the organisation of turn-taking for conversation are not part of a mental machinery of rules that stand outside of actual occasions of activity; that they are not algorithms; that they are not sets of instructions; that they are not programs of human thought; that they are not the cause of human action. Rules are oriented to features of action; they are contextual, situated practices of use.*

Button [ibidem] reminded his colleagues that Sacks explicitly moves away from a view that attributes causal efficacy to rules. Much in line with the rich psychological tradition, a causal view of rules would 'mine' CA research for 'social rules' that people 'follow'. Since these rules function as a kind-of set of instructions for people, they may as well have the same function for interactive devices. Against this causal conception of rules, Button highlights the ethnomethodological tradition of viewing rules as a local achievement. In that view, rules are not something that is part of an internal set of instructions of how to interact, but function as locally, mutually, publicly orientable, noticeable, follow-able, break-able and account-able objects in an actual unfolding interaction.

DOI: 10.14515/monitoring.2021.1.1793

Sacks' famous analysis of the first few seconds of talk during suicide prevention calls [Sacks, Jefferson, 1992: 3—11] highlights how ordinary speakers do not reproduce subconscious scripts but rely on a playful, actionable and thoughtful orientation to conversational expectations that can be used to further the distributions of conversational agendas at any given time.

In other words, there is not simply a rule like 'To ordinary speakers of the English language, when a speaker includes their name in their spoken introduction, the other speaker should include their name in the response'. If that were so, we could theoretically formalize sequences of social interaction into definite ritual-like processions steered by a long-yet-finite list of conversational rules. Instead, the rule itself is something that is available to the interactants. Not only that, its mutual availability is, in turn, also mutually available. This moves the relevance of rules from a causal — and therefore easily computerizable and instruct-able — to an instrumental dimension. Rules, therefore, feature as local hinges for the accomplishment of contingent actions.

This rule-shift is important, as it re-classifies the focus of sociology from the question of social order as a function of normative integration to the question of social order as a situated, second-by-second achievement. Computers are therefore principally unable to digest the kinds of rules introduced by CA, at least so long as their engineers implemented these rules at the level of pre-situational instructions. Thus, a conversation-analytic HCI (CA-HCI) is deeply problematic, as it glosses over the fact that CA's programme represents a shift in the analysis and granularity of action. Until computers become sufficiently advanced to be able to rule-orient rather than rule-follow, the machinery of conversation would, as a computer script, only result in a simulacrum of conversation. It can be argued that more advanced AI may actually be capable of moving beyond the rule-following paradigm described as being fundamental to Button-era AI [see Fordham, Gilbert, 1995; Button, Lee, Coulter, Sharrock, 1995; Wooffitt, Fraser, Gilbert, McGlashan, 1997]. The subsequent sections will discuss the technological advancements that have cast doubt on overly simplistic assessments of the capabilities of AI as a basis for understanding human-computer interaction.

Theoretical Shifts in HCI

The Sociological Status Quo: The Media Equation

While CA theorists engaged in a prolonged discussion of the fundamentals of action theory, and its significance for the application of CA to HCI, the dominant sociological contribution to HCI, apart from the purely methodological niche occupied by researchers like Suchman, remained the media equation paradigm. Its persistent foothold in HCI — despite the untimely recent death of one of its founders, C. Nass — can be partially attributed to the ease with which the media equation paradigm blended insights about human interaction and the dominant computational logic of HCI.

Nass, along with Reeves, in their Computers are Social Actors (CASA), media equation, or social response theory [Reeves, Nass, 1996; Nass, Steuer, Tauber, 1994] posited that humans, when using technology, regularly (and unconsciously) attribute humanlike characteristics to non-human material objects. More specifically, the presence of a number of subconscious cues (such as language use, voices, facial features) in a non-human artifact will result in an automatic application of social rules to that

DOI: 10.14515/monitoring.2021.1.1793

object. It means that interfaces imbued with minimal anthropomorphic cues activate social behavioral patterns in users; they unconsciously start to apply stereotypes, display politeness and treat the interface's conduct as though it was a human [Nass, Moon, Green, 1997; Nass, Moon, 2000].

Thus, for example, an embodied conversational avatar with a female face may be met with suspicion when placed in the context of a hardware store. In other cases, users may give more favorable reviews to a product in its presence (provided the abovementioned caveat of minimal cues), adhering to the principle of politeness.

Nass' approach, in short, has a peculiar understanding of what 'social rules' are, and how they are applied: social rules are treated as autonomous, involuntary, oftentimes non-reflexive, responses. These rules may be part of some genetic imprint or they may have been 'installed' by cultural forces over time; in either case, they stand in a causal relation to human action.

In sum, Nass' minimalist theory created a powerfully simple algorithm for designing interactive agents: on the condition of a very small number of anthropomorphic features being present, the machine's behavior could be directly informed by social-psychological insights on human behavior. Best practices for interface design could thus be directly mined from human-focused research, without much consideration for the specific configuration of the agent. The question would become 'what kind of human would you like in place of the avatar?', and the relevant characteristics could then be assembled by adding specific behavioral patterns. This paradigm is still prominently represented across a broad range of research on interface design and robotics [see for example Aeschlimann, Bleiker, Wechner, Gampe, 2020; Cameron et al., 2021; Lawson et al., 2021]. Yet, current theoretical trends in HCI are gradually changing the sociological status quo of the field.

We are currently observing a convergence of two theoretical developments. Firstly, the debates surrounding the second conversation-analytic intervention into HCI are turning to pragmatic — rather than paradigmatic — concerns. At the same time, the field of HCI in general is currently trending towards a need for a greater understanding of the situation of device use. More specifically, HCI is becoming increasingly interested in understanding the interactional resources available to people during the actually-unfolding interaction. The following sections will deal with these shifts in greater detail.

From Paradigmatic to Pragmatic — On Useful Simulacra

Over time, CA-inspired contributions to HCI have shifted away from the radical position of Button [1990]. They have done so by shifting the question from the domain of rule-following vs. rule-orientation to questions of practical accomplishments and sensemaking [Jones, Mitchell, 1994: 528]:

> *While true 'conversation' between human and computer is arguably not possible in principle [e.g. Button, 1990], nevertheless, as a metaphor for human-computer interaction, guidelines for effective human interaction are applicable to the design of intelligent support systems.*

In other words, while a 'conversation' is not possible, *something* is possible. After all, a *something* between human and machine still displays moment-by-moment sen-

DOI: 10.14515/monitoring.2021.1.1793

semaking on the part of the human, and still has an outcome. Wooffitt et al. [1997: 166] make the point more forcefully:

> *The basis of Button's distinction between human-human interaction and human-computer interaction is that, in the former case, rules are embodied in interaction. People orient to rules, whereas computers are determined by them. Thus, computers cannot converse because they cannot register or display any sensitivity to procedures for producing intelligible interaction. Therefore, it makes no sense to talk of interaction between humans and computers. There are good reasons, however, for assuming that this position may be simply incorrect, because in the case of human-computer exchanges, there is always one party that does possess the range of sense-making procedures which, according to Button, demarcate human-human interaction from human-computer interaction: the human participant will still be doing the things that humans do when they interact. That is, the full range of culturally available sense-making procedures will be brought to bear on any occasion, even if the other party to the interaction is a computer.*

In short, the question about the nature of conversational interaction was adjusted to the question of the practical accomplishment of a human-machine encounter. This move is not entirely fair to Button [1990], since his arguments concerned not the impossibility of humans interacting with, say, Wilson the volleyball, but the methodological pitfalls that may come with an overly enthusiastic equation of that type of encounter with a human-human interaction: we cannot make the volleyball do sensemaking by stuffing a CA textbook inside of it.

In practice, however, this move allowed CA-HCI researchers to digest Button's [1990] challenge by investigating (and contributing to the production of) *useful simulacra*.

Moore and Arar [2019], for example, while conceding that 'the "rules" and models of natural conversation provided by CA are not the same kind of rules as those found in a programmed system' and are 'not deterministic, but rather are representations of *resources* that human speakers use in repeated by nondeterministic ways', argue that this 'detailed picture of how human conversation works a speech-exchange system', allowing UX[4] designers to 'create interaction patterns that emulate features of human talk, although certainly with limitations and approximations' [ibid.: 88]. Moore and Arar call these systems 'conversation games', representing a 'distinctive form of interaction, which borrows interaction patterns from natural human conversation but also exhibits its own mechanics' [ibid.: 5].

In sum, by accepting the metaphorical character of 'human machine conversation', contemporary CA researchers sidestep the impossibility of investigating 'human machine conversation, but without quotes' by incorporating Button's methodological cautions without being defeatist about the entire project[5]. Shifting the research to *outcomes* over essences makes it possible to get on with practice-oriented CA-inspired research.

[4] User Experience (UX) design is a design paradigm that focuses on improving the user's experience with a product throughout its entire life cycle.

[5] Another theoretical shift in conversation-analytic approaches to technologies occupies the high-ontology segment of the conceptualization of artefacts: objects as sets of affordances or autonomous interactants. For a discussion of the conceptual and methodological implications of this shift, see [Erofeeva, 2019; Klowait, 2019].

DOI: 10.14515/monitoring.2021.1.1793

As a consequence, HCI—CA cemented its ability to produce meaningful insights to the *in-situ* interaction of humans and computers.

The Situational Turn in HCI

This pragmatic direction of HCI—CA coincided with the field of HCI on the whole. The potential for increased computational performance and representational fidelity contributed to a move towards designing more complex interactional systems. This, in turn, has contributed to a renewed interest in understanding the interactional resources used by humans. This section will discuss the paradigmatic shifts that have occurred within the more practice-and design-oriented spaces of HCI, and link it to the conceptual issues discussed previously. As it turns out, the largely CA-specific discussions that took place more than two decades ago were prescient regarding technological developments that occurred in more recent times. In other words, we may say that the technological capabilities have caught up with more theoretical concerns, and made the latter pragmatically relevant to the field as a whole.

Recently, Realism Maximization Theory (RMT) has emerged, arguing that 'minimal resemblance to human beings is not enough to improve the interaction, but that care must be taken to maximize realism, defined as the virtual character's degree of visual and/or behavioural resemblance to a human being' [Chérif, Lemoine, 2019: 30; Kang, Watt, 2013].

The consequences of RMT, with its inclusion of considerations of behavioural resemblance (rather than the presence of anthropomorphic cues) has contributed to an increased attention to *multimodality*, i. e. the consideration of how human behavior is accomplished with a body-in-space, and how 'different verbal and non-verbal cues such as presence/absence of audio, pitch, prosody, backchannel (BC), turn-taking, body posture/gesture (upper-torso, arms, hands, legs, etc.), facial expression, gaze, etc.' [Norouzi et al., 2018: 19] would need to be considered in interface design.

More abstractly, the shift to behavioral realism marked a shift in emphasis: interactional modalities such as gaze-behavior or gesticulation needed to be investigated in terms of their contribution to human social interaction, which, in turn, would inform human-avatar interaction. For example, the fact that *Clippy* has eyeballs becomes more than just a cue for social scripts; instead, RMT calls for a consideration of what the eyes actually do during interaction: what do they do when the avatar is listening, speaking, thinking, observing movement, and how is that behavior related to prosody, body posture, gesticulation and interactional context?

It is perhaps no coincidence that RMT emerged at a time where the technological capabilities of artificial intelligence increased dramatically. Contemporary conversational agents[6] can be imbued with complex language parsers, sentiment analysis systems, emotion detection, and image and video classifiers; at the same time, they are able to do so whilst rendering a responsive and fully embodied avatar. The central concern of the field — and how to design well-functioning embodied conversational agents (ECAs) — consequently becomes a matter of **congruence**.

[6] In keeping with the relevant research tradition, in what follows these systems will be called ECAs — Embodied Conversational Agents.

DOI: 10.14515/monitoring.2021.1.1793

The Rise of Congruence

Contemporary evidence-based ECA design can be characterized as an attempt to find the correct balance between the avatar's self-presentation and its actual capabilities (see fig. 3).

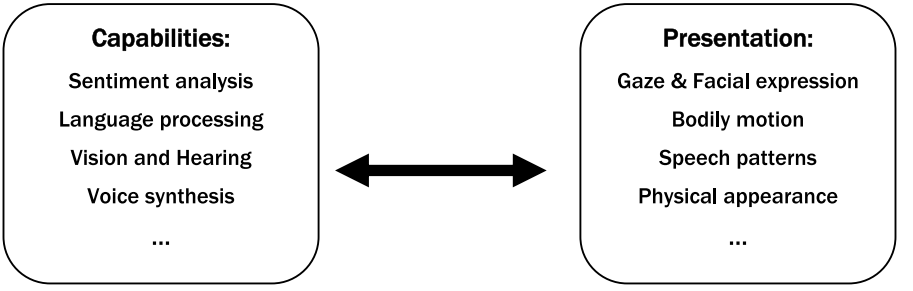

Fig. 3. The tension between capabilities and presentation

Since 'people are more comfortable with a robot clearly identified as non-human than with a human-looking robot that still has imperfections' [Chérif, Lemoine, 2019: 32], the design directive can be bi-directional, where the specific demands for the ECA can dictate the correct trade-off between realism and capability. If, for instance, only minimal behavioural humanlike-ness is required, the avatar may be **scaled back** in realism to adhere to the congruence principle. If, conversely, the avatar is hyper-real, steps must either be taken to imbue it with the correspondingly high level of capabilities, **or** make sure that the limited capabilities of the agent are adequately circumscribed (see Table 1).

Table 1. ***Paradigms mapped to the interaction-realism nexus***

	Low-realism	**High-realism**
Low-interaction capabilities	CASA paradigm	Incongruence
High-interaction capabilities	Limited congruence	RMT paradigm

Evidence-based Insights on Avatar Creation

ECAs, being conversational, are inherently interactive systems. Based on the aforementioned congruence principle, interaction should cohere with social rules, and the modalities present within the expressive range of the ECA should cohere with one another [Krämer, 2008]. As Tan and Liew [2020] argue, 'the mismatch between the seemingly sophisticated embodiment of virtual agents and the agents' lack of functional performance can cause the expectation gap in users, leading to disappointment and frustration with the embodied virtual agents' [ibid.: 1]. Samsung's STAR Labs recently debuted Neon at the International Consumer Electronics Show 2020. The way it was presented is a convenient means of illustrating the issue in practical terms (see fig. 4):

DOI: 10.14515/monitoring.2021.1.1793

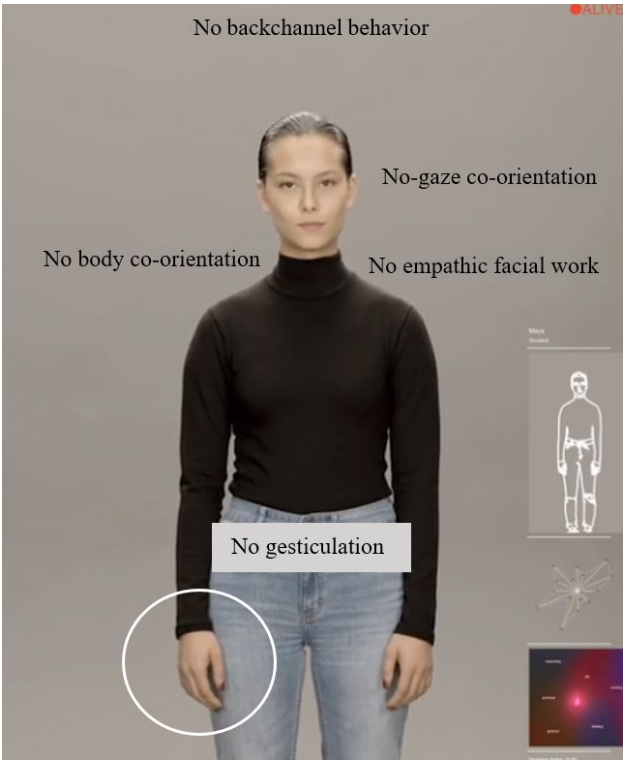

Fig. 4. Neon Avatar[7]

The above graphic is adapted to illustrate the CES demo of the current prototype[8]. It is a stark contrast to the humanlike idle animations that were demonstrated *prior* to the interactive demo: no alignment was demonstrated between human and ECA; no alignment was demonstrated between utterances and body/gestural work, either. While the system is still in development, some of the negative reactions can be explained through this incongruence.

In contrast to this, Soul Machines, arguably one of the industry leaders of realistic ECAs, demonstrates sophisticated backchannel behavior in even its less recent products, such as the Australia and New Zealand Banking Group's digital assistant, Jamie[9]. To reiterate: the focus is not realism at all costs, but the *evidence-based* application of humanlike interactional resources in appropriate contexts.

Verbal Behavior and Beyond

Verbal behavior, unsurprisingly, plays a prominent role in the design of conversational systems. While ECAs can remain 'mute' and talk through a purely text-based system,

[7] Source: URL: www.neon.life (accessed: 03.02.2021); text added by the present authors.

[8] For more details see: URL: https://youtu.be/4IPNOWilDOk (accessed: 03.02.2021).

[9] See: URL: https://youtu.be/eyoBgNY1KA0 (accessed: 03.02.2021).

DOI: 10.14515/monitoring.2021.1.1793

they are typically imbued with both a voice and the corresponding capability to move the lips. This is firstly due to the observation that embodied verbal behaviour — i. e. the match between visible embodied conduct and the audiostream — improves engagement, realism, trust [Bos, Olson, Gergle, Olson, Wright, 2002; Greenspan, Goldberg, Weimer, Basso, 2000], and, curiously, aids comprehension: in their seminal paper, Walker Sproull and Subramani [1994] 'investigated subjects' responses to a synthesized talking face displayed on a computer screen […]. Compared to subjects who answered questions presented via text display on a screen, subjects who answered the same questions spoken by a talking face spent more time, made fewer mistakes, and wrote more comments' [ibid.: 85].

Voice characteristics

Voice — far from being just a means of transmitting information between speakers — is a vehicle for personality, status, identity and relationship. As such, the construction of the ECA's voice is of key importance to the subsequent interaction with the user.

The lowest level of consideration is the question of what kind of voice technology to use. Currently, the choice is made between synthetic text-to-speech (TTS) generators and pre-recorded human voice that is played at the appropriate times. For practical purposes, this means a trade-off between conversational flexibility and expressiveness. Synthesized speech, be it concatenative (i. e. assembled out of fragments of recorded human speech) or parametric (i. e. newly generated through the creation of a linguistic phonetic speech model) has the benefit that it can produce context-aware responses from scratch. The downside, however, is that the speech itself is typically lifeless and robotic (especially in the case of parametric speech synthesis), with locally incoherent prosodic expressiveness. The alternative, using 'canned' pre-recorded (and therefore maximally human-sounding) responses greatly limits interactional flexibility, yet contributes to higher levels of trust when voice source is the only varied factor.

More recently, neural network based systems like Google's WaveNet proposed a 'parametric synthesizer on steroids' [van den Oord et al., 2016], i. e. a means of leveraging machine learning to create a neo-parametric model that is *both* lifelike and locally flexible in terms of what it can say. However, it and similar neural TTS solutions are arguably more difficult to implement and represent a potentially high overhead. Moreover, since systems like WaveNet learn to synthesize speech in a holistic manner — with the linguistic model becoming an implicit and entangled component of the learning process — 'prosody realization is randomly chosen and cannot be easily altered' [Shechtman, Sorin, 2019: 275]. In other words, the system becomes more lifelike in *sound*, but loses the ability to alter intonation to the context of the spoken word[10]. Thus, 'congratulations, you won' and 'my apologies, you failed' would sound the same without additional intervention, and would consequently be contextually incongruent. This shortcoming can be overcome [Shechtman, Sorin, 2019] but requires additional resources compared to simply recording two messages with appropriate prosodic contours.

In sum, as is the case with the aspect of avatar presentation (presence and embodiment), good ECA design becomes a matter of *leveraging existing capabilities to ensure*

[10] For an example of how a complex artificial intelligence system's use of prosody prolongs the closing of a service encounter see Egorova, Klowait, this issue.

DOI: 10.14515/monitoring.2021.1.1793

congruence. That is, if the conversational system is designed with a focus on pre-determined sentences, human voice recordings would be the obvious direction to head towards. Conversely, if conversational flexibility is desired — this is especially true for ECAs that try to maximize realism and generalize its range of applications — the more flexible synthetic voice is preferable, especially since conversational interactivity can moderate the de-anthropomorphizing effect of robotic voice. A negative example of the latter case is described in Klowait [2017], where an automated computer-telephone interviewing system was created with realistic pre-recorded human voices, yet integrated with a contextually inflexible conversational system. The resulting negative user experiences — a considerable number of which never reached understanding that they were talking to a limited AI rather than a rude/incapable human — could have been avoided by opting for an obviously synthetic voice.

Turn-taking behavior

Beyond the choice of voice, the things *done* with the voice are of key importance to realism. In that respect, the field has moved beyond the informational model of conversation, whereby speech is a kind of message exchange [Quarteroni, 2018; Schröder et al., 2012; Bernard, 2017]. Instead, the field has embraced the socio-linguistic, anthropological and sociological insight that *talk* is *action* [Goodwin, 2000] whereby humans typically accomplish disparate goals, moment-by-moment. A key element of this action is *turn-taking* behavior, i. e. the organization of participants' talk-action in a recognizable turn-by-turn basis, and the demarcation of interactional phenomena such as backchanneling ('listener' activity during another speaker's turn at speech), interruptions, anticipations of transitions between speakers and overlap.

Some of those behaviors are currently beyond the limits of AI, such as the human ability to exploit the implicit rules of conversational organization in order to change a topic, take over as a speaker or gracefully avoid answering a direct question. However, especially 'attentive listening behavior', where a 'combination of head nods, vocalizations and facial feedback [shows] agreement and acknowledgment' [Yalçın, 2020: 124] is possible to implement, and is argued to increase realism and rapport.

For example, Yalcin's [Yalçın, 2020; Yalçın, DiPaola, 2019] work on empathic ECAs includes an interruptible state-based model of affective listening [Yalçın, 2020: 125—126]:

> *According to the state of the dialogue, the behavior of the agent can change and adapt to the user. While the interaction partner is speaking, the agent enters the listening state. Listening mode will be activated via the speech and video input from the agent […]. In this state, the agent is expected to provide proper backchanneling behavior as well as the emotional feedback. After the speech of the interaction partner is completed, the agent will enter the thinking state. In this state, the agent will be finished gathering information from the perceptual model and start processing the speech input for generating a response. This response generation process will make use of the context of the dialogue as well as the emotional content of the message. Lastly, the agent will enter the speaking state, where it executes the prepared response via its output channels including voice, facial expression and body gestures.*

DOI: 10.14515/monitoring.2021.1.1793

In other words, such a simplified system can interface with pre-existing machine states, such as information retrieval, idling, processing and behavior production. Moreover, as this framework is multimodal, i. e. incorporates synchronous and overlapping modalities (e. g. gaze, head and body movement, prosody, speech), it represents another way of transforming unrelated idling mechanisms (such as randomized avatar body shifts) into context-sensitive tools to improve presence. Lee, Badler and Badler's [2002] work on turn-taking-sensitive saccades represents evidence that this is a general trend: gaze that is interactionally purposeful enhances user experience. This observation is notably true for both random eye movement and constant direct eye contact, which has a tendency to become unnerving to the human participants.

Lastly, a system that is mutually integrated — i. e. where e. g. gaze is coordinated with body posture and speech — makes it possible to flexibly integrate higher-order multimodal behavior such as affective matching and emotional mimicry [Yalçın, DiPaola, 2019], both of which further aid the humanlike-ness of interaction.

Specialized Subsystems

Beyond the more modest strategies of implementing *at least* some listening behavior, research has increasingly shifted focus to implementations of higher-order interactional features, such as social status-specific multimodal conduct [Nixon, DiPaola, Bernardet, 2018] and emotion monitoring, classifying and mirroring [Yalçın, 2020]. The general trend can be represented by a hierarchy (see fig. 5):

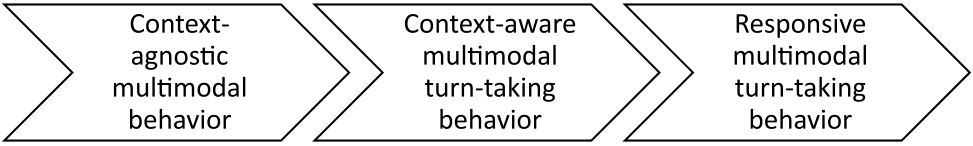

Fig. 5. Transition towards high-order interactional specificity

In other words, incongruent and random behavior (or the absence thereof) is rated worst, followed by behavior that displays a distinct (albeit teleologically fuzzy) orientation to ongoing interaction, followed by systems that attempt to respond to this ongoing interaction in a meaningful way (e. g. through empathy displays at the appropriate moments).

Towards a Multimodal Framework in HCI Research

As can be seen, current avatar design best practices are moving towards realism maximization. One consequence of this movement is the shifting focus towards the multimodality of communication when developing avatars. Despite this trend, ECA evaluations tend to focus on non-situational, non-ethnographic research methods [Kramer, ter Stal, Mulder, de Vet, van Velsen, 2020]. This methodological preference does not make it possible to investigate the actual use of multimodal resources for the design of ECAs, since experimental research does not rely on the study of what functions different modalities (gesture, gaze, etc.) perform in human communication.

For instance, a number of studies in the field of multimodal human-computer interaction show that people not only perceive ECA behavior in a certain way, but also adapt

DOI: 10.14515/monitoring.2021.1.1793

to it in real-time [Norouzi et al., 2018; Zhang at al., 2010; Bohus, Horvitz, 2010]. If the agent is set up to track an object at the center of the user's attention 90 % of the time, the user will pay less attention to the virtual agent. If this condition is changed to 50 % or 10 %, the user will begin to focus on the ECA [Zhang et al., 2010]. In other words, the level of involvement inscribed in the system affects the level of user involvement in said interaction: if an ECA has a single point of gaze-focus, users will tend to expend little resources for tracking the ECA's gaze; if, conversely, the ECA visibly distributes their attention to other objects in the world, users attend to that fact by tracking the ECA's gaze more attentively. However, this result in itself does not say anything about the need for this level of user involvement. In order to evaluate it, we need to understand what function attention management serves in interaction.

When it comes to complex multimodal interaction scenarios for state-of-the-art ECAs, it appears especially important to incorporate interaction-focused research methodologies at all stages of the design process. More specifically, there is a lack of a true multimodal framework for the development and final evaluation of the interactional outcomes of ECAs.

Given the multimodal complexity of the multiparty coordination of speech, gaze and other interactive resources, it would seem reasonable to assume that the research and evaluation included insights from state-of-the-art multimodality research. Yet, only classical gaze studies [Kendon, 1967] seem to find widespread purchase in the field.

As was elaborated in a previous section of this review, CA's current response to Button's critique (useful simulacra) goes hand in hand with technical developments in HCI on the whole. As such, the field is now at a stage where CA may make a third serious attempt at a methodological intervention[11]. This time, however, with the aid of multimodal conversation analysis, specifically.

Multimodal conversion analysis makes it possible to analyze the practices and resources used by people to interact with one another. Such resources are not limited to the domain of language but also include their embodied actions — such as gestures, gaze directions, and body movement [Goodwin, 2000; Mondada, 2019]. In contrast to the multimodal approach in psychology, communicative resources are thought to be applied situationally, i. e. within a specific context and for a specific task.

Research of human interaction with ECAs, carried out in line with multimodal conversation analysis, follows the logic proposed by Suchman [1987]: on the one hand, there is the object's design into which a particular mode of interaction is inscribed; on the other, there are the ways this object is interacted with *insitu*. In the process of analyzing video recordings of real-life interaction, discrepancies are found between what the system 'sees' (its interactive abilities) and what the user expects [Ewa, Abigail, 2016]. In particular, studies of interaction with conversational agents reveal problems with the turn-taking system (transfer of speakership from an ECA to a person and vice versa): people either start talking too early or 'skip' the allotted space for an answer, as a result of which their utterance is ignored, forcing participants repeat the sequence of interaction over and over [Arend, Sunnen, 2017; Pelikan, Broth, 2016; Pitsch, Gehle, Dankert, Wrede, 2017]. Furthermore, conversational agents may overlook the constant

[11] See Saul Albert's recent keynote: Albert S. (2020, June 29) Three Meeting Points between CA and AI. URL: https://saulalbert.net/blog/three-meeting-points-between-ca-and-ai/ (accessed: 03.02.2021).

DOI: 10.14515/monitoring.2021.1.1793

repair sequences performed by human speakers, with similar consequences [Klowait, 2017; Trott, Rossano, 2017].

The main practical conclusion of such studies is the idea that — for a smooth interaction with an ECA — the user needs to somehow be 'explained' how the virtual agent works, and the communicative abilities the system possesses. That said, the aforementioned learning processes occur within the interaction itself. From a theoretical point of view, this means that the social expectations of users dynamically change based on their reaction to the sequentially unfolding actions of the virtual agent.

Multimodality, understood in the conversation-analytic tradition, challenges both the media equation paradigm and the idea of realism maximization. Unlike the media equation, people interact with an ECA on the basis of a situational adaptation to the interface, and not on the basis of the unconscious illusion that we are interacting with a person rather than a computer (these theoretical assumptions were conceptualized by Klowait as a distinction between pragmatic and ontological anthropomorphism [see Klowait, 2018a]). As for the realism maximization theory, it merely manages to touch the tip of the iceberg of real engagement with an ECA.

Conclusion

This review traced a convergence of developments. On the one hand, we demonstrated the path of CA-HCI from an initial low-stakes intervention [Suchman, 1987], to a high-stake overenthusiasm [Frohlich, Luff, 1990] and, finally, to a marked shift towards pragmatic in-field concern [Moore, Arar, 2019]. On the other, we demonstrated how contemporary studies of ECAs have converged on the need to understand the multimodality of human interaction to facilitate the development of novel forms of human-computer encounters. We showed how there is currently a growing need for a congruence between an ECAs representational fidelity and interactional competence. We believe that multimodal conversation analysis can provide much needed answers, both in terms of insights and in terms of the appropriate ways to generate them.

References (Список литературы)

Aeschlimann S., Bleiker M., Wechner M., Gampe A. (2020) Communicative and Social Consequences of Interactions with Voice Assistants. *Computers in Human Behavior*. Vol. 112. 106466. https://doi.org/10.1016/j.chb.2020.106466.

Arend B., Sunnen P. (2017) Coping with Turn-Taking: Investigating Breakdowns in Human-Robot Interaction from a Conversation Analysis (CA) Perspective. In: N. Callaos, B. Sanchez, M. Savoie, F. Welsch, J. V. Carrasquero (eds.) *Proceedings: The 8th International Conference on Society and Information Technologies (ICSIT 2017)*. Orlando (FL): International Institute of Informatics and Systemics (IIIS). P. 149—154. URL: http://hdl.handle.net/10993/30952 (accessed: 05.02.2021).

Bernard D. (2017) Cognitive Interaction: Towards "Cognitivity" Requirements for the Design of Virtual Assistants (Working paper). In: *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Banff (CA): IEEE. P. 210—215. https://doi.org/10.1109/SMC.2017.8122604.

DOI: 10.14515/monitoring.2021.1.1793

Boehm B. W. (1988) A Spiral Model of Software Development and Enhancement. *Computer*. Vol. 21. No. 5. P. 61—72. https://doi.org/10.1109/2.59.

Bohus D., Horvitz E. (2010) Facilitating Multiparty Dialog with Gaze, Gesture, and Speech. In: *Proceedings of the 12th International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI—MLMI'2010)*. New York (NY): Association for Computing Machinery. Article No. 5. P. 1—8. https://doi.org/10.1145/1891903.1891910.

Bos N., Olson J., Gergle D., Olson G., Wright Z. (2002) Effects of Four Computer-Mediated Communications Channels on Trust Development. In D. Wixon (ed.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'02)*. New York, NY: ACM Press. P. 135—140. https://doi.org/10.1145/503376.503401.

Button G. (1990) Going Up a Blind Alley. Conflating Conversation Analysis and Computational Modelling. In: P. Luff, G. N. Gilbert, D. Frohlich (eds.) *Computers and conversation*. London: Academic Press. P. 67—90.

Button G., Lee J. R., Coulter J., Sharrock W. (1995) Computers, Minds and Conduct. Cambridge: Polity Press.

Cameron D., de Saille S., Collins E., Aitken J., Cheung H., Chua A., Loh E. J., Law J. (2021) The Effect of Social-Cognitive Recovery Strategies on Likability, Capability and Trust in Social Robots. *Computers in Human Behavior*. Vol. 114. P. 106561. https://doi.org/10.1016/j.chb.2020.106561.

Card S. K., Moran T. P., Newell A. (1983) The Psychology of Human-Computer Interaction. Hillsdale, NJ: Erlbaum.

Carroll J. M. (2010) Conceptualizing a Possible Discipline of Human — Computer Interaction. *Interacting with Computers*. Vol. 22. No. 1. P. 3—12. https://doi.org/10.1016/j.intcom.2009.11.008.

Chérif E., Lemoine J.-F. (2019) Anthropomorphic Virtual Assistants and the Reactions of Internet Users: An Experiment on the Assistant's Voice. *Recherche et Applications en Marketing (English Edition)*. Vol. 34. No. 1. P. 28—47. https://doi.org/10.1177/2051570719829432.

Couper-Kuhlen E. (2019) American English OKAY over Time: Challenge and Chance for Interactional Linguistics. Lecture given at the *Ninth Meeting of the Language and Social Interaction Working Group (LANSI)*. New York, NY: Teachers College, Columbia University.

Erofeeva M. (2019) On Multiple Agencies: When do Things Matter? *Information, Communication & Society*. Vo.l 22. No. 5. P. 590—604. https://doi.org/10.1080/1369118X.2019.1566486.

Ewa L., Abigail S. (2016) "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)*. San Jose, CA: Association for Computing Machinery. P. 5286—5297. https://doi.org/10.1145/2858036.2858288.

DOI: 10.14515/monitoring.2021.1.1793

Fordham A., Gilbert N. (1995) On the Nature of Rules and Conversation. *AI & SOCIETY.* No. 9. P. 356—372. https://doi.org/10.1007/BF01210587.

Frohlich D., Luff P. (1990) Applying the Technology of Conversation to the Technology for Conversation. In: P. Luff, G. N. Gilbert, D. Frohlich (eds.) *Computers and conversation*. London: Academic Press. P. 187—220.

Garfinkel H. (1967) Studies in Ethnomethodology. Englewood Cliffs, NJ: Prentice Hall.

Goffman E. (1981) Forms of Talk. Philadelphia, PA: University of Pennsylvania Press.

Goodwin C. (2000) Action and Embodiment within Situated Human Interaction. *Journal of Pragmatics*. Vol. 32. No. 10. P. 1489—1522. https://doi.org/10.1016/S0378-2166(99)00096-X.

Greenspan S., Goldberg D., Weimer D., Basso A. (2000) Interpersonal Trust and Common Ground in Electronically Mediated Communication. In: W. Kellogg, S. Whittaker (eds.) *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW'00).* New York (NY): ACM Press. P. 251—260. https://doi.org/10.1145/358916.358996.

Heritage J. (1984) A Change-of-State Token and Aspects of its Sequential Placement. In: J. M. Atkinson, J. Heritage (eds.) *Structures of social action: Studies in conversation analysis*. Cambridge: Cambridge University Press. P. 299—345.

Heritage J. (2016) On the Diversity of 'Changes of State' and their Indices. *Journal of Pragmatics*. Vol. 104. P. 207—210. https://doi.org/10.1016/j.pragma.2016.09.007.

Jones P. M., Mitchell C. M. (1994) Model-Based Communicative Acts: Human-Computer Collaboration in Supervisory Control. *International Journal of Human-Computer Studies*. Vol. 41. No. 4. P. 527—551. https://doi.org/10.1006/ijhc.1994.1072.

Kang S.-H., Watt J. H. (2013) The Impact of Avatar Realism and Anonymity on Effective Communication via Mobile Devices. *Computers in Human Behavior*. Vol. 29. No. 3. P. 1169—1181. https://doi.org/10.1016/j.chb.2012.10.010.

Pitsch K., Gehle R., Dankert T., Wrede S. (2017) Interactional Dynamics in User Groups: Answering a Robot's Question in Adult-Child Constellations. In *Proceedings of the 5th International Conference on Human Agent Interaction (HAI'17)*. Bielefeld (Germany): Association for Computing Machinery. P. 393—397. https://doi.org/10.1145/3125739.3132604.

Kendon A. (1967) Some Functions of Gaze-Direction in Social Interaction. *Acta Psychologica*. Vol. 26. P. 22—63. https://doi.org/10.1016/0001-6918(67)90005-4.

Klowait N. (2017) A Conceptual Framework for Researching Emergent Social Orderings in Encounters with Automated Computer-Telephone Interviewing Agents. *International Journal of Communication and Linguistic Studies*. Vol. 15. No. 1. P. 19—37.

Klowait N. (2018a) Reflexive Anthropomorphism: Ontological Ignorance, or Ignorant Ontology? *Sotsiologicheskii Zhurnal*. Vol. 24. No. 1. P. 8—33. https://doi.org/10.19181/socjour.2018.24.1.5711. (In Russ.)

DOI: 10.14515/monitoring.2021.1.1793

Кловайт Н. Рефлексивный антропоморфизм: неведение онтологии или невежественная онтология? *Социологический журнал*. 2018. Т. 24. № 1. С. 8—33. https://doi.org/10.19181/socjour.2018.24.1.5711.

Klowait N. (2018b) The Quest for Appropriate Models of Human-Likeness: Anthropomorphism in Media Equation Research. *AI & SOCIETY*. Vol. 33. No. 4. P. 527—536. https://doi.org/10.1007/s00146-017-0746-z.

Klowait N. (2019). Interactionism in the Age of Ubiquitous Telecommunication. *Information, Communication & Society*. Vol. 22. No. 5. P. 605—621. https://doi.org/10.1080/1369118X.2019.1566487.

Kramer L., ter Stal S., Mulder B. C., de Vet E., van Velsen L. (2020) Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review. *Journal of Medical Internet Research*. Vol. 22. No. 2. e14058. https://doi.org/10.2196/14058.

Krämer N. C. (2008) Social Effects of Virtual Assistants. A Review of Empirical Results with Regard to Communication. In: H. Prendinger, M. Ishizuka, J. Lester (eds.) *Intelligent Virtual Agents. IVA 2008. Lecture Notes in Computer Science*. Vol. 5208. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. P. 507—508. https://doi.org/10.1007/978-3-540-85483-8_63.

Lawson A. P., Mayer R. E., Adamo-Villani N., Benes B., Lei X., Cheng J. (2021) Recognizing the Emotional State of Human and Virtual Instructors. *Computers in Human Behavior*. Vol. 114. 106554. https://doi.org/10.1016/j.chb.2020.106554.

Lee S. P., Badler J. B., Badler N. I. (2002) Eyes Alive. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Technics (SIGGRAPH'02)*. New York (NY): Association for Computer Machinery. P. 637—644. https://doi.org/10.1145/566570.566629.

Luff P., Gilbert G. N., Frohlich D. (eds.) (1990) Computers and Conversation. London: Academic Press.

Mondada L. (2019) Contemporary Issues in Conversation Analysis: Embodiment and Materiality, Multimodality and Multisensoriality in Social Interaction. *Journal of Pragmatics*. Vol. 145. P. 47—62. https://doi.org/10.1016/j.pragma.2019.01.016.

Moore R. J., Arar R. (2019) Conversational UX design: A Practitioner's Guide to the Natural Conversation Framework (ACM Books). New York, NY: ACM Books.

Nass C., Moon Y. (2000) Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*. Vol. 56. No. 1. P. 81—103. https://doi.org/10.1111/0022-4537.00153.

Nass C., Moon Y., Green N. (1997) Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers with Voices. *Journal of Applied Social Psychology*. Vol. 27. No. 10. P. 864—876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x.

DOI: 10.14515/monitoring.2021.1.1793

Nass C., Steuer J., Tauber E. R. (1994) Computers Are Social Actors. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'94)*. New York, NY: ACM. P. 72—78. https://doi.org/10.1145/191666.191703.

Nixon M., DiPaola S., Bernardet U. (2018) An Eye Gaze Model for Controlling the Display of Social Status in Believable Virtual Humans. In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. Maastricht: IEEE. P. 1—8. https://doi.org/10.1109/CIG.2018.8490373.

Norouzi N., Kim K., Hochreiter J., Lee M., Daher S., Bruder G., Welch G. (2018) A Systematic Survey of 15 Years of User Studies Published in the Intelligent Virtual Agents Conference. In: *Proceedings of the 18th International Conference on Intelligent Virtual Agents*. Sydney (AUS): Association for Computing Machinery. P. 17—22. https://doi.org/10.1145/3267851.3267901.

Pelikan H. R. M., Broth M. (2016) Why That Nao? How Humans Adapt to a Conventional Humanoid Robot in Taking Turns-at-Talk. In: *Proceedings of the 34th Annual CHI Conference on Human Factors in Computing Systems (CHI'16)*. San Jose (CA): Association for Computing Machinery. P. 4921—4932.

Quarteroni S. (2018) Natural Language Processing for Industry. *Informatik-Spektrum*. Vol. 41. No. 2. P. 105—112. https://doi.org/10.1007/s00287-018-1094-1.

Reeves B., Nass C. I. (1996) The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. CSLI Publications. Stanford (CA): Cambridge University Press.

Sacks H. (1984) On Doing "Being Ordinary". In J. M. Atkinson, J. Heritage (eds.) *Structures of Social Action: Studies in Conversation Analysis*. Cambridge (UK): Cambridge University Press. P. 413—429. https://doi.org/10.1017/CBO9780511665868.024.

Sacks H., Jefferson G. (1992) Lectures on Conversation. Oxford (UK): Blackwell.

Schröder M., Bevacqua E., Cowie R., Eyben F., Gunes H., Heylen D., ter Maat M., McKeown G., Pammi S., Pantic M., Pelachaud C., Schuller B., de Sevin E., Valstar M., Wöllmer M. (2012) Building Autonomous Sensitive Artificial Listeners. *IEEE Transactions on Affective Computing*. Vol. 3. No. 2. P. 165—183. https://doi.org/10.1109/T-AFFC.2011.34.

Shechtman S., Sorin A. (2019) Sequence to Sequence Neural Speech Synthesis with Prosody Modification Capabilities. In: *Proceedings of the 10th ISCA Speech Synthesis Workshop*. ISCA. P. 275—280. https://doi.org/10.21437/SSW.2019-49.

Stokoe E. (2018) Talk: The Science of Conversation. London: Robinson.

Suchman L. A. (1987) Plans and Situated Actions: The Problem of Human-Machine Communication. Cambridge: Cambridge University Press.

Tan S.-M., Liew T. W. (2020) Designing Embodied Virtual Agents as Product Specialists in a Multi-Product Category E-Commerce: The Roles of Source Credibility and Social Presence. *International Journal of Human–Computer Interaction*. Vol. 36. No. 12. P. 1136—1149. https://doi.org/10.1080/10447318.2020.1722399.

DOI: 10.14515/monitoring.2021.1.1793

Trott S., Rossano F. (2017) Theoretical Concerns for the Integration of Repair. In: *Artificial Intelligence for Human Robot Interaction*. Arlington, VA: AAAI. P. 118—122.

van den Oord A., Dieleman S., Zen H., Simonyan K., Vinyals O., Graves A., Kalchbrenner N., Senior A., Kavukcuoglu K. (2016) WaveNet: A Generative Model for Raw Audio. URL: https://arxiv.org/abs/1609.03499 (accessed: 06.02.2021).

Walker J. H., Sproull L., Subramani R. (1994) Using a Human Face in an Interface. In: B. Adelson, S. Dumais, J. Olson (eds.) *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems Celebrating Interdependence (CHI'94)* New York, NY: ACM Press. P. 85—91. https://doi.org/10.1145/191666.191708.

Wooffitt R., Fraser N. M., Gilbert N., McGlashan S. (1997) Humans, Computers and Wizards: Analysing Human (Simulated) Computer Interactions. London: Routledge.

Yalçın Ö. N. (2020) Empathy Framework for Embodied Conversational Agents. *Cognitive Systems Research*. Vol. 59. P. 123—132. https://doi.org/10.1016/j.cogsys.2019.09.016.

Yalçın Ö. N., DiPaola S. (2019) Evaluating Levels of Emotional Contagion with an Embodied Conversational Agent. In: A. K. Goel, C. M. Seifert, C. Freska (eds.) *Proceedings of the 41st Annual Meeting of the Cognitive Science Society (CogSci'19)*. P. 3143—3149. URL: https://ivizlab.org/wp-content/uploads/sites/2/2019/09/0528.pdf (accessed: 06.02.2021).

Zhang H., Fricker D., Smith T. G., Yu C. (2010) Real-Time Adaptive Behaviors in Multimodal Human-Avatar Interactions. In: *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (IVA'10)*. Beijing: Association for Computing Machinery. Article 4. URL: https://dll.sitehost.iu.edu/papers/iva10_zhang.pdf (accessed: 06.02.2021).