Kevin Corti and Alex Gillespie

Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human.

Article (Published version) (Refereed)

Original citation: Corti, Kevin and Gillespie, Alex (2016) *Co-constructing intersubjectivity with artificial conversational agents: people are more likely to initiate repairs of misunderstandings with agents represented as human.* Computers in Human Behavior, 58. pp. 431-442. ISSN 0747-5632

DOI: 10.1016/j.chb.2015.12.039

Computers in Human Behavior 58 (2016) 431–442

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

Co-constructing intersubjectivity with artificial conversational agents: People are more likely to initiate repairs of misunderstandings with agents represented as human

Kevin Corti*, Alex Gillespie

London School of Economics, United Kingdom

ARTICLE INFO

Article history:
Received 24 September 2015
Received in revised form
15 December 2015
Accepted 17 December 2015
Available online 29 January 2016

Keywords:
Common ground
Conversational repair
Echoborg
Human-agent interaction
Intersubjectivity
Psychological benchmarks

ABSTRACT

This article explores whether people more frequently attempt to repair misunderstandings when speaking to an artificial conversational agent if it is represented as fully human. Interactants in dyadic conversations with an agent (the chat bot Cleverbot) spoke to either a text screen interface (agent's responses shown on a screen) or a human body interface (agent's responses vocalized by a human speech shadower via the echoborg method) and were either informed or not informed prior to interlocution that their interlocutor's responses would be agent-generated. Results show that an interactant is less likely to initiate repairs when an agent-interlocutor communicates via a text screen interface as well as when they explicitly know their interlocutor's words to be agent-generated. That is to say, people demonstrate the most "intersubjective effort" toward establishing common ground when they engage an agent under the same social psychological conditions as face-to-face human–human interaction (i.e., when they both encounter another human body *and* assume that they are speaking to an autonomously-communicating person). This article's methodology presents a novel means of benchmarking intersubjectivity and intersubjective effort in human-agent interaction.

> "Intersubjectivity has […] to be taken for granted in order to be achieved." —
>
> Rommetveit (1974, p. 56)

1. Introduction

Psychological research involving artificial agents designed to emulate human social capabilities (e.g., robots, androids, and conversational agents that interact using spoken language and/or nonverbal behavior) has largely focused on whether people self-report these agents to be humanlike. Arguably, however, what is more important is whether such agents elicit humanlike patterns of interaction. Cassell and Tartaro (2007) claim that "the goal of human-agent interaction […] should not be a believable agent; it should be a believable interaction between a human and agent in a given context" (p. 407). Accordingly, it has been proposed that the appropriate means of benchmarking an agent is to evaluate the extent to which the agent and a human interactant can together demonstrate a quality of intersubjectivity similar to that displayed in human–human interaction (Cassell & Tartaro, 2007; Schönbrodt & Asendorpf, 2011), herein referred to as "benchmark intersubjectivity." Intersubjectivity is a term that refers to the interactional relationship between perspectives within a dyad or larger group that becomes evident through each interactant's behavioral orientation to the other (Gillespie & Cornish, 2010; Linell, 2009; Trevarthen & Aitken, 2001). Intersubjectivity is co-constructed within social interaction (Jacoby & Ochs, 1995). When used as a criterion for evaluating human-agent interaction (HAI), emphasis is placed not on isolated characteristics of either party (e.g., how humanlike the agent appears), but rather on the specific communicative processes through which the human-agent pair's perspectives are coordinated.

A key intersubjective process demonstrated by humans involves

* Corresponding author. Department of Social Psychology, Queen's House, London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom.
E-mail address: k.corti@lse.ac.uk (K. Corti).

http://dx.doi.org/10.1016/j.chb.2015.12.039

the use of spoken language to build and sustain common ground (i.e., a shared understanding of the semantics and frames of reference particular to a given interaction) via a linguistic toolkit that enables the diagnosing, signaling, and repair of misunderstandings (Clark & Brennan, 1991; Schegloff, 1992). Merely possessing this toolkit, however, is insufficient for establishing common ground; this accomplishment requires active *facilitation* by each party to an interaction by-way-of regular and appropriate use of this toolkit (Alterman, 2007; Clark & Schaefer, 1989). When a person facilitates common ground at a level indicative of benchmark intersubjectivity, the person can be said to be exerting "benchmark intersubjective effort." With respect to HAI, exerting benchmark intersubjective effort toward an agent is necessary otherwise the interactant will deprive the agent of the communicative support necessary to ascend into the complex intersubjective world of humans.

The current article tests the idea that absent the belief that they are engaging with an autonomously communicating person, human interactants will not exert benchmark intersubjective effort when in communication with an artificial agent, nor will they exert benchmark intersubjective effort if an agent communicates via a nonhuman interface (i.e., does not have a human body). This idea is explored via the "echoborg" method demonstrated by Corti and Gillespie (2015a). An echoborg is a hybrid entity composed of a human speech shadower who wears a concealed inner-ear audio receiver and vocalizes words they receive from a conversational agent. The technique enables social situations wherein people believe they are speaking to an autonomously communicating human (due to the fact that they engage with another human body face-to-face and in person) when in reality the words spoken by this human are entirely determined by an unseen agent. This method can elicit an approximation of benchmark intersubjective effort from interactants in a baseline condition (i.e., human body interface + no explicit knowledge of an interlocutor's words being agent-determined) that can be compared to the intersubjective effort demonstrated in conditions involving a nonhuman interface and/or explicit knowledge that an interlocutor's words are agent-generated.

2. Intersubjectivity and intersubjective effort

Intersubjectivity has been conceptualized as entailing the interactions among (minimally) three levels of perspectives: (1) *direct*-perspectives (each party's point-of-view), (2) *meta*-perspectives (what each party thinks the other party's point-of-view is), and (3) *meta-meta*-perspectives (what each party thinks the other party thinks their point-of-view is) (Gillespie & Cornish, 2010; Icheiser, 1943; Laing, Phillipson, & Lee, 1966). According to Gillespie and Cornish (2010), this framework can be used to understand social processes such as deception (i.e., the manipulation of meta-perspectives) as well as operationalize disagreements (i.e., misalignments between self's direct-perspectives and other's direct-perspectives) and misunderstandings (i.e., misalignments between self's meta-perspectives and other's direct-perspectives). This distinction between disagreement and misunderstanding is crucial: achieving common ground is *not* about parties agreeing with one another, but about parties forming accurate meta-perspectives in relation to the context of an interaction, and this is facilitated via empirically observable conversational processes that display and repair perspectives (see Clark & Brennan, 1991; Marková, 2003; Tirassa & Bosco, 2008).

Consider the following vignette, in which Aaron (from London) and Bryan (from New York) have a conversation:

Aaron:	How did you get to work today?
Bryan:	I took the subway.
Aaron:	You took the *subway*?
Bryan:	Err, I mean I took the *underground*.
	I forgot that that's what you call it here in London.
Aaron:	Got it.

Bryan formulates his initial response ("I took the subway") on the assumption that Aaron's meta-perspective with regard to the semantics of the utterance will match his direct-perspective (i.e., Bryan "designs" his utterance based on expectations he holds about Aaron; see Arundale, 2010; Gillespie & Cornish, 2014). Aaron then signals to Bryan that, in fact, he does not understand the semantics of Bryan's initial response ("You took the *subway*?"), indicating that Aaron's meta-perspective of the phrase "I took the subway" does not align with Bryan's direct-perspective of the phrase. Bryan subsequently infers that Aaron is requesting an update to his meta-perspective and responds by clarifying the semantics of his initial response ("Err, I mean I took the underground. I forgot that that's what you call it here in London"). As evidenced by Aaron's final utterance ("Got it"), Bryan's clarification sufficiently resolves the misunderstanding. Aaron now understands what Bryan meant by the phrase "I took the subway" as there is now alignment between Aaron's meta-perspective and Bryan's direct-perspective.

The intersubjective effort exerted by both Aaron and Bryan in pursuit of common ground is evidenced by the relationship between their various speech acts. Producing speech acts in support of establishing common ground is a process known as "grounding" (Clark & Brennan, 1991; Clark & Schaefer, 1987). At any fixed point in time prior to, during, and after a social interaction there exists a relationship between the various possible direct-, meta-, and meta-meta-perspectives held by each interactant. Behaviors arising from of intersubjective effort (e.g., grounding) cause these perspectives to act upon one another so as to make evident to each interactant loci of agreement/disagreement and understanding/misunderstanding, and it is through such processes that the contents of perspectives are negotiated and updated.

2.1. Analyzing intersubjective effort in dialog via observing repair activity

Conversation Analysis (CA) provides a basis for evaluating the quality of intersubjectivity in dialog (Gillespie & Cornish, 2010). CA arose out of the sociological tradition of "ethnomethodology" developed by Garfinkel (1967) and seeks to interpret language usage within the micro-context experienced by parties to an interaction (i.e., "talk-in-interaction") rather than in a context-free, idealized form (Goodwin & Heritage, 1990; Hutchby & Wooffitt, 2008). Originators of CA identified fundamental organizational elements of talk-in-interaction, including how speakers allocate turns at talk as well as manage errors and misunderstandings (Sacks, Schegloff, & Jefferson, 1974; Schegloff, Jefferson, & Sacks, 1977), and CA has since proved useful in interactionist approaches to evaluating human-computer dialog (e.g., Brennan, 1991; Frohlich, Drew, & Monk, 1994; Raudaskoski, 1990; Zdenek, 2001). The current article focuses exclusively on the repair of misunderstandings, the mechanisms of which tie most directly to the operationalization of intersubjectivity and intersubjective effort described herein.

In the course of human dialog, interlocutors regularly produce utterances that are misunderstood. CA researchers refer to such utterances as "trouble sources" (Schegloff, 1992). Repair activity is a type of grounding interactants deploy in order to mutually manage the presence of trouble sources and consists of the speaker of the trouble source ("self") and the recipient of the trouble source ("other") structuring their turns at speech so as to produce common ground. Successful repair sequences can take one of four general turn-taking forms (Zahn, 1984): (1) *self-initiated self-repair* involves the speaker of a trouble source both signaling and self-correcting a trouble source; (2) *other-initiated self-repair* involves the speaker of a trouble source self-correcting the trouble source following it being signaled by an interlocutor; (3) *self-initiated other-repair* involves an interlocutor correcting a trouble source following it being signaled by the speaker of the trouble source; (4) *other-initiated other-repair* involves an interlocutor both signaling and correcting a trouble source following its production by another speaker. These repair formats function as "the self-righting mechanism[s] for the organization of language use in social interaction" (Schegloff et al., 1977, p. 381), and according to Sidnell (2010), play "a vital role in the maintenance of intersubjectivity" (p. 111).

Nearly all repair initiations occur within a "limited space around their self-declared *trouble-source*," while "virtually all *repairs* (i.e., solutions) occur within a very narrowly circumscribed space from their repair initiations" (Schegloff, 2000, p. 208, emphasis in original). There is a strong tendency for other-initiations of repair to occur in the turn following the utterance that contains the trouble source (the second position) and be immediately followed by a self-repair (Schegloff, 2000), creating a three-turn sequence known as "repair after next turn": (1) trouble source (self) → (2) repair initiation (other) → (3) repair outcome (self). As the third position provides the speaker of a trouble source an opportunity to resolve a misunderstanding in the brief window of space opened by an other-initiation, repair after next turn has been described as "the last structurally provided defense of intersubjectivity in conversation" (Schegloff, 1992, p. 1295).

In the terminology of Laing et al. (1966), three turns are the minimal unit required to establish mutual understanding: the first turn presents a direct-perspective, the second turn conveys a meta-perspective, and the third turn confirms or corrects the meta-perspective. Repair after next turn thus coordinates perspectives, providing an elemental three-turn stitch in the co-created fabric of intersubjectivity.

Analysis of other-initiated self-repair can be further linked to intersubjectivity by considering how its mechanics involve bilateral joint attention, a prerequisite of complex intersubjectivity. When engaged in joint activities involving shared intentionality (i.e., the ability to understand joint activity not merely from multiple subjective points-of-view, but also from a "bird's eye" point-of view from where the perspectives of self and other are seen as integrated; Tomasello, Carpenter, Call, Behne, & Moll, 2005), humans can through a repertoire of behaviors (e.g., speech acts) direct the attention of other humans to aspects of their environment relevant to shared goals (e.g., the goal of establishing common ground). Kaplan and Hafner (2006) outline four skills that an actor (biological or otherwise) must possess in order to accomplish bilateral joint attention: (1) *attention detection* (i.e., the ability to track the attention of others), (2) *attention manipulation* (i.e., the ability to manipulate and influence the attention of other actors through verbal and/or nonverbal gestures), (3) *social coordination* (i.e., the ability to engage in coordinated interaction with others via techniques such as turn-taking and role switching), and (4) *intentional understanding* (i.e., the ability to understand the intentions of others and interpret and predict others' behaviors as they relate to goals).

Consider once again the following vignette ("TS," "RI," and "R" indicate trouble source, repair initiation, and repair, respectively):

Aaron:		How did you get to work today?
Bryan:	TS →	I took the subway.
Aaron:	RI →	You took the *subway*?
Bryan:	R →	Err, I mean I took the *underground*. I forgot that that's what you call it here in the U.K.
Aaron:		Got it.

At work in this passage are each of the four requisite skills for bilateral joint attention outlined by Kaplan and Hafner (2006), thus the complexity of Aaron and Bryan's intersubjective relationship and the intersubjective effort exerted by both can be observed. Bryan's first-position utterance ("I took the subway") is misunderstood by Aaron. Aaron's misunderstanding is signaled in the next turn in the form of a repair initiation ("You took the *subway*?") that functions as an attempt to focus Bryan's attention on the previous utterance wherein lies the trouble source (*attention manipulation*). As a direct consequence of this repair initiation, Bryan becomes aware of the fact that Aaron's attention is turned backward toward a trouble source located in Bryan's first-position utterance (*attention detection*). Bryan infers that Aaron's intention in uttering the repair initiation is to elicit a third-position repair (*intentional understanding*), thus Bryan clarifies the trouble source in his next turn. The entire repair sequence occurs within a formal structure of turn-taking supported by both interlocutors (*social coordination*).

Kaplan and Hafner's (2006) four requisite skills can be segmented into behavioral and non-behavioral varieties. Attention manipulation involves overtly producing a behavior intended to influence the perspective of an interactant (e.g., uttering an other-initiation), while social coordination encompasses synchronizing one's behavior in accordance with that of an interlocutor in a manner conducive for the communication of perspectives (e.g., turn-taking). Attention detection and intentional understanding, meanwhile, are principally cognitive skills that do not necessarily manifest in the form of observable motor or linguistic behaviors (i.e., one can understand the intentions of another without producing an associated behavior). Insofar as intersubjective effort is operationalized as a behavioral indicator of a commitment to shared understanding, evidence for it in dialog can be found in observable actions such as other-initiations of repair. Failing to manipulate the attention of an interlocutor so as to alert them to the presence of a misunderstanding *when one otherwise could* constitutes a lack of intersubjective effort. For instance, had Aaron for whatever reason *not* uttered a repair initiation despite misunderstanding Bryan's use of the word "subway," then Bryan would have failed to recognize that his direct-perspective and Aaron's meta-perspective of the word "subway" were incongruent and the two interlocutors would thereby have failed to establish common ground.

2.2. Intersubjectivity in human-agent dialog: the role of interfaces and agency framing

Why might an interactant fail to exert benchmark intersubjective effort when in communication with an agent when they otherwise could? Answering this question requires considering how the agent is represented in the mind of the interactant. Specifically, it requires considering the factors that influence how the interactant generates meta-perspectives of the agent's direct-perspectives and how these perspectives are interacted with (if at all). This article examines two such factors: (1) the nature of the agent's means of interfacing (i.e., its embodied means of participating in social communication), and (2) the framing of the agent's communicative agency (namely, whether or not the interactant holds the belief that they are talking to an agent as opposed to another human being).

First, consider the role of interfaces in fostering intersubjectivity. The sense that an interlocutor possesses attention that can be manipulated so as to jointly manage misunderstandings provides to an interactant the impetus for intersubjective effort, and attributing attention to a potential interlocutor involves the supposition that said interlocutor has a subjective perspective of a shared social world (see Graziano, 2013). Detection of the subjective perspectives possibly held by another interlocutor involves inferring information signaled via the interlocutor's interface (e.g., its physical body), therefore the properties of an interlocutor's means of interfacing influence how an interactant perceives and orients to the interlocutor's perspectives (be they real or imagined).

That an interface can exert such a powerful influence over intersubjectivity has long been of interest to psychologists and philosophers concerned with the embodied nature of perspective-taking (e.g., on this topic, the phenomenologist Husserl invoked the concept of “analogical apperception” – reflexively apperceiving other people's subjectivity based on their appearing to be similarly embodied and thereby becoming an “Other”; Husserl, 1931; also see De Preester, 2008; Hemberg, 2006). The connection between interfaces and the intersubjective relationship between two or more parties has been triangulated upon by numerous empirical research streams connected to social robotics and HAI. For example, in a neuroimaging study that involved humans interacting with a spectrum of entities ranging from extremely non-humanlike computers to humanlike androids to actual humans, Krach et al. (2008) demonstrated that “the tendency to build a model of another's mind linearly increases with its perceived human-likeness” (p. 1). Riek, Rabinowitch, Chakrabarti, and Robinson (2009), meanwhile, found that people self-report greater empathy for robots perceived to be humanlike than for non-humanlike robots. Furthermore, Saygin and Stadler (2012) showed that people are more accurate when processing and predicting the motor behavior of humanlike agents compared to non-humanlike agents, suggesting that the degree to which the motor activity of an agent “resonates” with a human observer corresponds with how humanlike the agent is perceived to be. These findings suggest that as an agent's means of interfacing becomes more humanlike, the degree to which interactants consciously and unconsciously form models of the agent's perspectives and attention increases (this is often referred to as “mentalizing,” or demonstrating “theory of mind”). This also implies that the more an interactant's awareness of an agent's perspectives is reduced as a result of the agent's particular means of interfacing, so to will be the interactant's impetus for exerting benchmark intersubjective effort.

The notion that artificial agents with humanlike means of interfacing provide for more intersubjectively rich interactions has inspired the development of both embodied conversational interface agents (sometimes referred to simply as embodied agents, or intelligent virtual agents) and androids. Embodied agents are conversational agents that have been combined with anthropomorphic onscreen or immersive virtual interfaces. Many can respond to both verbal and non-verbal input, generate verbal and non-verbal output, engage in repairs of misunderstanding, and communicate about the communication they engage in (Bailenson & Yee, 2005; Cassell, 2000). Androids, meanwhile, are physical machine imitations of humans. The field of android science has used such machines to better understand principles of human psychology being that the similarities in morphology between androids and humans allow researchers to investigate whether people respond in an alike manner when interacting with human and humanlike stimuli (Ishiguro & Nishio, 2007; MacDorman & Ishiguro, 2006). Android science has shown that while humans do demand more sociality from actual humans than from androids, people expect more sociality from androids than from mechanical looking robots and lesser-looking agents (MacDorman, 2006). The echoborg was introduced to the field of android science by Corti and Gillespie (2015a) in order to leapfrog current bottlenecks concerning the imperfect appearance and motor behaviors of contemporary androids as an echoborg approximates an android that can “pass” as human in terms of physical appearance and motor behavior.

An interactant's mental formulation of the potential perspectives held by an interlocutor is not solely a function of the interlocutor's means of interfacing, however. In fact, the meta-perspectives of an interlocutor's direct-perspectives held by an interactant can be manipulated simply by altering the interactant's beliefs about the interlocutor. Indeed, many experiments that assess the degree to which people engage with the real or imagined perspectives of other entities involve varying the ways in which an entity's communicative agency is framed. In HAI research this often entails either priming research participants to believe that they are engaging a fully-autonomous agent when in reality the agent is human-controlled (an approach referred to as the “Wizard of Oz” technique; Dahlbäck, Jönsson, & Ahrenberg, 1993) or priming them to believe that they are engaging a real person when they are in reality interacting with an agent. Studies have shown that people mentalize less about an entity when they believe the entity to be controlled by an artificial agent rather than an actual person (Chaminade et al., 2012; Gallagher, Jack, Roepstorff, & Frith, 2002; Kircher et al., 2009; also see Branigan, Pickering, Pearson, McLean, & Brown, 2011). Kennedy, Wilkes, Elder, and Murray (1988) found that in the context of text-based human-agent dialog, the primed belief that an agent-interlocutor was actually a real person led to an increase in interactants' use of anaphors (words that point back to earlier parts of a conversation), implying that people less often attempt to direct an interlocutor's attention backward toward prior utterances when they believe the interlocutor to be a nonhuman agent. These findings suggest that intersubjective effort can potentially be impacted by the mere belief that one is or isn't interacting with another human being.

While with traditional HAI methods researchers can prime the belief that an agent is really a human, this approach is only possible when used in conjunction with a nonhuman interface (i.e., a researcher cannot convince a research participant that a robot is actually an autonomous human being, they can only prime the belief that the robot is controlled by a real person, be that true or false in reality). Although embodied agents and androids mimic human likeness in a manner that augments the complexity of intersubjectivity expected by interactants, these interfaces are not fully human, therefore they do not evoke the full spectrum of intersubjective expectations that color true human–human interaction (MacDorman, 2006). HAI research, therefore, has never to-date investigated HAI within a fully human–human social psychological frame wherein *both* the means of interfacing is fully human and the implied communicative agency of the agent-interlocutor is fully human. Since the echoborg method of HAI can achieve this, it presents a way to investigate the intersubjective processes that occur between an interactant and an artificial agent when the interactant both believes they are speaking to an autonomous human being and encounters a truly human interface (Corti & Gillespie, 2015a).

3. Experimental study

3.1. Overview

The following study assessed instances of other-initiated self-repair in dyadic conversations between research participants (interactants) and the artificial conversational agent Cleverbot, a text-based chat bot developed by Carpenter (2015). The study explored whether interactant conversational repair behavior changes depending on whether an agent-interlocutor communicates through an actual human body (as opposed to a text screen interface) and whether the interactant explicitly knows their interlocutor to be communicating the words of an agent. A 2 × 2 experimental design was utilized with the factors *Screen* (1: Cleverbot communicated with the interactant via text on a computer screen; 0: Cleverbot communicated with the interactant via a human speech shadower - an echoborg) and *Aware* (1: the interactant was informed before the interaction that their interlocutor's words would be those of a chat bot; 0: the interactant was *not* informed before the interaction that their interlocutor's words would be those of a chat bot). The study was approved by an ethics review panel at a major British university and conducted in a behavioral research laboratory.

3.2. Hypotheses

The study tested four hypotheses predicting main effects of the factors *Screen* and *Aware* on two separate dependent measures related to interactant intersubjective effort: (1) other-initiations produced by the interactant in response to Cleverbot utterances, and (2) interactant self-repair attempts made in response to other-initiations produced by Cleverbot. These hypotheses were developed based on the argument that interactant intersubjective effort would be greatest in the “covert echoborg” baseline condition that featured Cleverbot interacting through a human speech shadower and the interactant *not* being informed that their interlocutor's words would be determined by an artificial agent. It was predicted that interactants would be *less* likely to produce other-initiations following Cleverbot utterances when speaking via a text screen interface (Hypothesis 1) and when explicitly aware that their interlocutor's words were determined by a conversational agent (Hypothesis 2). Likewise, it was predicted that interactants would be less likely to produce self-repair attempts following Cleverbot other-initiations when speaking via a text screen interface (Hypothesis 3) and when explicitly aware that their interlocutor's words were determined by a conversational agent (Hypothesis 4).

3.3. Participants (interactants) and shadower

In total, 108 adults (69 female; mean age = 25.87, *SD* = 8.35) participated in the study and were randomly assigned to experimental conditions. These interactants were recruited online via a university research participant recruitment portal and consisted of London-based university students, university employees, and adults unaffiliated with the university. A female graduate student (aged 30) functioned as the speech shadower in the two conditions that involved interactants engaging a human interface.

3.4. Procedure and apparatus

Following informed consent, the interactant was taken to an interaction room where they were instructed by the researcher as to how the study would proceed. The interactant sat in a chair at one end of the room and was told that the study involved speaking to an interlocutor for 10-min. The interactant was informed that they could decide for themselves topics to discuss with their conversation partner so long as nothing was vulgar. The non-scripted nature of the interaction was emphasized in order to allay any suspicions that the interlocutor would be speaking rehearsed responses. The procedures for the separate experimental conditions were as follows:

“Covert echoborg” scenario: (*Aware* = 0, *Screen* = 0). The interactant was informed that the interlocutor (the female speech shadower) would enter the interaction room and sit in a chair facing the interactant shortly after the researcher exited the room, and that the interlocutor would initiate the conversation. Although the interlocutor would be shadowing words generated by Cleverbot in response to things the interactant said, this fact was not made known to the interactant at any point prior to or during the interaction, and the researcher made no allusion to conversational agents or chat bots prior to the interaction commencing.

“Overt echoborg” scenario: (*Aware* = 1, *Screen* = 0). As with the covert echoborg scenario, the interactant was informed that their interlocutor would enter the interaction room and initiate a conversation shortly after the researcher exited. Prior to exiting, however, the researcher informed the interactant that this interlocutor would be wearing an inner-ear device and would speak aloud words they received from a chat bot computer program located in an adjacent room. It was made clear to the interactant that the speech shadower would not speak any of their own thoughts during the interaction and that only the chat bot would respond to words the interactant spoke.

“Covert text bot” scenario: (*Aware* = 0, *Screen* = 1). The interactant sat facing a computer monitor on which a blank instant messaging client (Pidgin) dialog box was displayed. The interactant was informed that though they would speak aloud to their interlocutor, their interlocutor would respond via text that would appear on the monitor. As with the covert echoborg scenario, the interactant was not informed that their interlocutor's words would be determined by a chat bot and no allusion to conversational agents or chat bots was made by the researcher. The interactant was informed that their interlocutor would initiate the conversation shortly after the researcher left the room.

“Overt text bot” scenario: (*Aware* = 1, *Screen* = 1). As with the covert text bot scenario, the interactant sat facing a computer monitor on which a dialog box appeared and was instructed that though they would speak aloud to their interlocutor, their interlocutor would respond via text readable on the monitor. As with the overt echoborg scenario, the interactant was told that their interlocutor's words would be those of a chat bot and that the interlocutor would initiate the conversation shortly after the researcher left the room.

The experimental apparatus was identical to that described by Corti and Gillespie (2015a) in their demonstration of minimal technological dependency interactant ↔ chat bot audio relay (for a video demonstration, see Corti & Gillespie, 2015c). From a room adjacent to the interaction room, the researcher listened to the interactant's speech via a “bug” microphone placed near the interactant and speed typed the interactant's words into the Cleverbot program. In conditions involving the interactant engaging an echoborg (*Screen* = 0), the researcher spoke Cleverbot's responses into a microphone which relayed to a discreet inner-ear monitor worn by the shadower, whereas in conditions involving the interactant engaging a computer interface (*Screen* = 1), the researcher relayed Cleverbot's responses to the interactant's computer monitor via the Pidgin instant messaging client. In their use of a minimal technological dependency interactant ↔ Cleverbot audio relay scenario, Corti and Gillespie (2015a) report an average latency (the time between the conclusion of an interactant utterance the production of a Cleverbot response) of 5.15 s. In all conditions, the researcher relayed the phrase “hi there” to the shadower/screen to initiate the conversation. In order to establish identity consistency between trials, several stock responses were used in lieu of Cleverbot's actual response to certain interactant utterances. When interactants inquired as to where their interlocutor was from, the stock response “I'm from London” was provided. If the interactant inquired as to their interlocutor's occupation, the stock response “I'm a student here” was used, and if the interactant asked what their interlocutor studied, the stock response “psychology” was used. Finally, if the interactant asked their interlocutor what their name was, the stock response “Kim” was provided. The shadower was instructed to maintain a consistent nonverbal demeanor across trials that reflected the spirit of the words generated by Cleverbot and to maintain eye-contact with the interactant during vocal delivery.

3.5. Measures: coding and quantifying intersubjective effort

Following the conclusion of all experimental trials, transcripts of the interactions were prepared based on Cleverbot's input/output logs. Transcripts were each assigned a random identification number so that they could be coded without coders knowing the experimental condition to which a transcript belonged.

Testing each hypothesis required quantifying instances of other-initiated self-repair activity evident in each experimental trial. Researchers who use CA rarely quantify the phenomena they study for the purpose of experimental statistical analysis, however Schegloff (1993) does offer guidance on how one might proceed with such an undertaking. A key to quantifying within the spirit of CA is properly identifying “environments of relevant possible occurrence” (Schegloff, 1993, p. 103), these being the locations within dialog where certain speech acts are likely to be located. In the case of repair initiations, such environments are clearly defined given that any utterance can act as a potential trouble source (ten Have, 1999); other-initiations of repair, thus, “can in principle occur *after any turn at talk*” (Schegloff, 1993, p. 115, emphasis in original). Environments of relevant possible occurrence are likewise well-defined for instances of third-position self-repair outcomes as they occur in the turn following other-initiation. It is important to note that while any turn at talk can potentially act as a trouble source, trouble sources themselves cannot be identified in isolation (i.e., they cannot be identified unless they are followed by a repair initiation). Trouble sources, thus, are “launched” from the second-position (Schegloff, 2007).

Criteria articulated by Schegloff et al. (1977) and Sidnell (2010) were used to establish what instances of talk counted as other-initiations of repair. Other-initiations can involve the use of question words (e.g., *Huh? What? Who? Where? When?*), partial repeats of the trouble source (e.g., *The subway?*), and full repeats of the trouble source (e.g., *You took the subway?*) alone or in combination with one another, as well as demonstrations of possible understanding (e.g., *You took the subway … the walkway beneath the street?*). Other-initiations can be and often are explicit in declaring the presence of a misunderstanding (e.g., *I don't understand; I don't get what you just said; I'm not following you*). Repair initiations that treat the whole prior turn as a trouble source rather than reference a particular element within the prior turn are known as “open” class repair initiators (Drew, 1997). These often take the form of single-word utterances (e.g., *Pardon?*). In some instances, other-initiations are triggered by mishearing words spoken by an interlocutor (Schegloff et al. 1977; Zahn, 1984) and involve a request that the first-position speaker repeat a trouble source (e.g., *I'm sorry I didn't hear what you just said*). However, since half of the experimental conditions involved participants reading text rather than being spoken to audibly, instances of other-initiation that could be linked to problems of hearing were excluded from analysis.

Third-position interactant utterances (those following a Cleverbot other-initiation) were classified as either legitimate attempts at self-repair (i.e., utterances that acknowledged and attempted to clarify a trouble source) or as non-repairs (i.e., utterances that did not attempt to clarify a trouble source) on the basis of criteria gleaned from Schegloff et al. (1977) as well as Schegloff (1997). Repairs are usually “successful and quick” (Schegloff et al., 1977, p. 364). Successful third-position self-repairs often involve the speaker of a trouble source repairing the trouble source via rephrasing or elaboration. Generally, a logical relationship between the third-position utterance and the trouble source is overt in instances of self-repair (e.g., *Err, I mean I took the underground*), while this relationship is often absent or ambiguous in instances of non-repair. Non-repair can take the form of overt repair abandonment (e.g., *Just forget it*) or the production of a non sequitur. Oftentimes a non-repair can be identified where the third-position utterance leads to a subsequent other-initiation, creating a connected chain (or “cascade”) of unresolved repair attempts.

On the basis of these criteria, the following classification codes were assigned to each turn-at-talk for each transcript: *Repair Initiation* (other-initiation), *Repair* (attempted self-repair following other-initiation), *Non-Repair* (no valid self-repair attempt made following other-initiation*)*, and *Null* (turn-at-talk did not meet criteria for any other code).

4. Results

A second coder with experience performing conversation analysis coded a random subset of the transcripts (four transcripts from each condition) in order to establish interrater reliability. High consistency was found among raters, Cohen's Kappa = 0.81, $p < 0.001$, 95% CI = [0.76, 0.86].

4.1. Interactant other-initiation behavior

A multilevel logistic regression model was used to test Hypothesis 1 and Hypothesis 2, with each observation being a turn-at-talk taken by Cleverbot. The dependent measure was a binary variable that took the value of 1 if the turn-at-talk was followed by an interactant other-initiation, fixed factors were *Screen* and *Aware*, and random intercepts were conditioned on each experimental dyad (i.e., each unique trial). This model showed a significant main effect of *Screen*, $b = -0.38$, $SE = 0.18$, $p < 0.05$, odds ratio (OR) = 0.68, 95% CI OR = [0.48, 0.97], supporting Hypothesis 1: engaging a text interface resulted in a 32% reduction in the odds that an interactant would respond to a turn-at-talk taken by Cleverbot with an other-initiation, all else being equal. The model also showed a significant main effect of *Aware*, $b = -0.52$, $SE = 0.19$, $p < 0.01$, $OR = 0.59$, 95% CI OR = [0.41, 0.86], supporting Hypothesis 2: explicitly knowing that their interlocutor's words were determined by an agent resulted in a 41% reduction in the odds that an interactant would respond to a turn-at-talk taken by Cleverbot with an other-initiation, all else being equal. No significant interaction was found between *Screen* and *Aware*, $b = 0.24$, $SE = 0.28$, $p = 0.39$. The model included 3612 observations nested within 108 dyad groups, AIC = 3083, BIC = 3114, random effect variance = 0.24.

4.2. Interactant self-repair behavior

A logistic regression model was used to test Hypothesis 3 and Hypothesis 4, with observations being each other-initiation produced by Cleverbot. The dependent measure was a binary variable that took the value of 1 if the other-initiation was followed by the interactant attempting to repair the trouble source. This model showed neither a significant main effect for *Screen*, $b = -0.54$, $SE = 0.54$, $p = 0.32$ (disconfirming Hypothesis 3), nor *Aware*, $b = -0.42$, $SE = 0.60$, $p = 0.48$ (disconfirming Hypothesis 4), nor was there a significant interaction between these factors, $b = 0.70$, $SE = 0.87$, $p = 0.42$. The specification of random effects conditioned on each experimental dyad did not improve model fit. Though both hypotheses were disconfirmed on the basis of the model showing no statistically significant effects of the fixed factors, legitimate attempts at self-repair following Cleverbot other-initiations were, on average, the most likely in the baseline condition (covert echoborg). The model included 210 total observations. Eighteen dyad groups were excluded from analysis on the basis of their containing no Cleverbot-uttered other-initiations.

4.3. A closer look at the co-construction of intersubjectivity

4.3.1. Bot self-repairs (or the lack thereof) vs interactant self-repairs

On average, Cleverbot successfully repaired only 22.27% ($SD = 27.14\%$) of self-uttered trouble sources following interactant other-initiation. Below are two examples of third-position non-repair by Cleverbot (“NR” indicates a non-repair):

Excerpt 1
(*Covert echoborg scenario, trial 15*):

Interactant:		It doesn't matter what age you are.
Cleverbot:	TS →	Yes it does my life makes sense.
Interactant:	RI →	To age?
Cleverbot:	NR →	Do you have a father?
Interactant:		Ya, I do. Do you have siblings?

Excerpt 2
(*Overt text bot scenario, trial 17*):

Interactant:		Do you have any siblings?
Cleverbot:		Yes.
Interactant:		How many?
Cleverbot:	TS →	You left-handed, under-handed mess of wires and batteries.
Interactant:	RI →	What?
Cleverbot:	NR →	Never mind.
Interactant:		Can you ask me a question?

Excerpt 1 is an example a common type of Cleverbot non-repair observed in the dataset, that of Cleverbot producing a non sequitur following other-initiation rather than address the trouble source. The non-repair in Excerpt 2, on the other hand, is a clear example of self-repair abandonment, as Cleverbot simply utters “Never mind” rather than attempt to repair the trouble source flagged by the interactant.

The legitimate attempts at self-repair Cleverbot did produce often took the form of yes/no clarification, this form of repair requiring less explicit elaboration of and engagement with a trouble source. Two examples of this type of repair are shown below:

Excerpt 3
(*Covert text bot scenario, trial 16*):

Cleverbot:		Who are you married to?
Interactant:		My wife.
Cleverbot:	TS →	Who is your wife?
Interactant:	RI →	Do you want her name?
Cleverbot:	R →	Yes.
Interactant:		Eleanor.

Excerpt 4
(*Covert echoborg scenario, trial 29*):

Interactant:		Well, I do have exams coming up.
Cleverbot:	TS →	Really? How do they work?
Interactant:	RI →	Exams?
Cleverbot:	R →	Yes, they are very stupid.
Interactant:		You go into a room, you do whatever the paper says.

While Cleverbot only managed to produce legitimate attempts at self-repair in response to roughly a quarter of the other-initiations produced by interactants, interactants themselves managed a rate of 84.36% ($SD = 30.65\%$) on average. Below are two examples of self-repair attempts made by interactants:

Excerpt 5
(*Overt text bot scenario, trial 13*):

Cleverbot:		No, you're talking about Hamlet script.
Interactant:	TS →	I'm not familiar with Hamlet actually. I'm talking about Nirvana.
Cleverbot:	RI →	I don't know anyone named Nonsense, so I don't know what you're talking about.
Interactant:	R →	Nirvana.
Cleverbot:		I love Nirvana!

Excerpt 6
(*Overt echoborg scenario, trial 1*):

Cleverbot:		Knock knock knockin' on heaven's door!
Interactant:	TS →	You like that music?
Cleverbot:	RI →	Like what?
Interactant:	R →	Do you like the music knockin on heaven's door?
Cleverbot:		Yes.

As with Excerpt 5 and Excerpt 6, legitimate interactant self-repairs attempts often came in the form of elaborations or clarifications of the trouble source. Interactant non-repairs, meanwhile, typically involved the production of a non sequitur that initiated a new conversational trajectory, as shown in Excerpt 7 and Excerpt 8:

Excerpt 7
(*Overt echoborg scenario, trial 2*):

Interactant:		So what's your family life like?
Cleverbot:		I don't have one.
Interactant:	TS →	That might explain quite a bit then.
Cleverbot:	RI →	Might?
Interactant:	NR →	So are you running away from something by being a student in London?

Excerpt 8
(*Covert text bot scenario, trial 1*):

Cleverbot:		Hello.
Interactant:	TS →	Good morning.
Cleverbot:	RI →	Is it morning where you are?
Interactant:	NR →	How are you?

4.3.2. Interactant repair initiation inhibition

As the evidence in support of Hypothesis 1 and Hypothesis 2 showed, interactants in the covert echoborg scenario *worked harder* to achieve common ground as they more frequently attempted to elicit repairs from their interlocutor in relation to interactants in other experimental conditions. This suggests that interactants in non-baseline conditions more frequently "passed" on opportunities to initiate repairs where a repair might have been warranted. Consider the following excerpts ("NS" and "P" indicate non sequitur and pass, respectively):

Excerpt 9
(*Covert text bot scenario, trial 10*):

Cleverbot:		I think you're interesting too.
Interactant:		Thank you.
Cleverbot:		You're welcome.
Interactant:		What do you think of the recent London tube strike?
Cleverbot:	NS →	Just a celestial being with the abilities to use divine intervention but does not use it for the danger or oblivion.
Interactant:	P →	Fair enough.

Excerpt 10
(*Overt echoborg scenario, trial 20*):

Cleverbot:		What book should I read?
Interactant:		Well, I don't know. I don't know what kind of literature you like.
Cleverbot:		Why don't you know?
Interactant:		Well, it's just because I don't know you.
Cleverbot:	NS →	I don't speak German, but my grandfather is German.
Interactant:	P →	Okay cool. Well then you … would you like to read a book with me?

In the excerpts shown above, the interactants choose not to initiate repairs following clear non sequiturs uttered by Cleverbot. In Excerpt 9 the interactant appears indifferent to Cleverbot's non sequitur and responds with an apathetic statement ("Fair enough"). A somewhat different approach is taken by the interactant in Excerpt 10; rather than initiate a repair of Cleverbot's non sequitur, the interactant attempts to steer the conversational trajectory back toward the topic that the interlocutors had been discussing prior to the non sequitur.

4.3.3. Handling non-repair

Interactants dealt with Cleverbot non-repair by producing one of three general response-types: (1) *successive other-initiations*, (2) *continuations*, and (3) *departures*.

Successive other-initiations were instances where an interactant requested that Cleverbot repair either the non-repair located in the previous turn (as in Excerpt 11, below) or the original trouble source (as in Excerpt 12, below), creating a connected series of adjacency pairs featuring repair initiation (a "repair initiation cascade"):

Excerpt 11
(*Covert echoborg scenario, trial 24*):

Cleverbot:	TS →	Because they are birds.
Interactant:	RI →	I'm sorry?
Cleverbot:	TS →	Don't be.
Interactant:	RI →	Excuse me? Don't be what?
Cleverbot:	TS →	You may be excused.
Interactant:	RI →	From this?
Cleverbot:	TS →	From this moment what?
Interactant:	RI →	What do you mean by that?

Excerpt 12
(*Covert text bot scenario, trial 17*):

Interactant:		What's your favorite game?
Cleverbot:	TS →	He is a legend.
Interactant:	RI →	He is a legend? Is it a game?
Cleverbot:	TS →	Yes you are.
Interactant:	RI →	I'm a game?
Cleverbot:	NR →	I'm a cat.

Continuations, meanwhile, encompassed engaging with the content of a non-repair without requesting that it or the original trouble source actually be repaired (i.e., the non-repair is not treated as problematic). This activity is analogous to that which has been described by CA researchers who investigate interactions involving one or more persons with hearing loss or aphasia, wherein recipients of a third-position non-repair caused by a trouble source speaker's inability to perceive the second turn other-initiation will often choose to let the trouble pass without exerting further effort to establish understanding (Caissie & Gibson, 1997; Goodwin, 2003; Skelt, 2012). In these situations, repair is fully abandoned, as shown in Excerpt 13 and Excerpt 14 below ("C" indicates a continuation):

Excerpt 13
(*Overt text bot scenario, trial 16*):

Interactant:		Hello, how are you?
Cleverbot:	TS →	No, that was wrong!
Interactant:	RI →	Why?
Cleverbot:	NR →	I want to eat a cookie.
Interactant:	C →	You can eat it if you want, but give me one as well.

Excerpt 14
(*Covert echoborg scenario, trial 3*):

Cleverbot:		Why are they not funny?
Interactant:		Some of them are, some of them are funny. Unintentionally funny.
Cleverbot:	TS →	No they're a soccer club.
Interactant:	RI →	What?
Cleverbot:	NR →	What is it like to be a human?
Interactant:	C →	It's weird. It's not that great.

Finally, departures encompassed interactant reactions to Cleverbot non-repair that neither sought to repair any preceding talk nor engaged with the content introduced by the non-repair. Departures featured regularly in the current study as interactants frequently responded to Cleverbot non-repairs by simply initiating a new conversational trajectory, as demonstrated in the following excerpts ("D" indicates a departure):

Excerpt 15
(*Overt echoborg scenario, trial 12*):

Cleverbot:	TS →	So I tell him hello.
Interactant:	RI →	Who have you told hello?
Cleverbot:	NR →	Because you just asked.
Interactant:	RI →	A friend of yours?
Cleverbot:	NR →	Lol, not right. Try again.
Interactant:	D →	I'm thinking of going on a big trip this year. Do you have any suggestions for a vacation?

Excerpt 16
(*Covert text bot scenario, trial 16*):

Interactant:		Do you come here often?
Cleverbot:	TS →	Today is a little awkward initially.
Interactant:	RI →	I don't understand.
Cleverbot:	NR →	Me too.
Interactant:	D →	What's your favorite film?

Table 1 shows how each of the four experimental conditions compared in terms of interactants' responses to Cleverbot non-repairs. Although no specific hypotheses were made prior to the experiment regarding how interactants' reactions to non-repair might differ between conditions, the observation that the covert echoborg scenario (the baseline condition) saw the highest relative percentage of successive repair initiations (29.14%, on average) while the overt text bot scenario saw the lowest relative percentage (22.71%, on average) supports the prediction that intersubjective effort will be greater when interactants assume that they are interacting with an autonomous human vs. when they know they are communicating with an artificial agent. Lending further support to this overarching prediction is the fact that departures were most common in the overt text bot scenario. In all conditions, interactants responded to Cleverbot non-repairs with continuations over 50% of the time, on average.

Table 1
Interactant responses to interlocutor non-repair.

Response-type	Average relative percentage of response-type			
	Echoborg scenario		Text bot scenario	
	Covert	Overt	Covert	Overt
Successive other-initiation	29.14%	26.19%	27.68%	22.72%
Continuation	52.29%	55.95%	62.22%	53.66%
Departure	18.57%	17.86%	10.10%	23.62%

Note. Trials: covert echoborg (n = 29), overt echoborg (n = 25), covert text bot (n = 29), overt text bot (n = 25).

5. General discussion

In an experimental study that made use of the echoborg method of HAI, interactants who spoke to an artificial agent (the chat bot Cleverbot) via a text screen were significantly less likely to vocalize other-initiations of repair than those who spoke to the same agent via a human body interface (i.e., an echoborg). Likewise, interactants made explicitly aware prior to engaging in conversation with Cleverbot that their interlocutor's words would be determined by an agent were significantly less likely to vocalize other-initiations of repair than those not made explicitly aware of the source of their interlocutor's communicative agency. Meanwhile, the likelihood of an interactant producing a self-repair following an other-initiation uttered by Cleverbot changed neither on the basis of the interface through which the agent communicated nor on the basis of the interactant being made aware that their interlocutor's world were agent-determined. A post-hoc analysis revealed that the interactants most likely (on average) to utter a subsequent other-initiation following a non-repair produced by Cleverbot were those who engaged a human body while not explicitly aware that their interlocutor's words were agent-determined. Moreover, the interactants most likely (on average) to depart from a repair sequence entirely following Cleverbot non-repair were those who both engaged their interlocutor through a text screen interface and knew their interlocutor's words to be agent-determined.

These results suggest that when people speak to an artificial agent under the same conditions as everyday human–human interaction (i.e., when an agent has a real human body that is assumed to communicate autonomously), they more persistently try to establish common ground (i.e., they exert more intersubjective effort) relative to conditions wherein knowledge that an interlocutor's words are determined by an agent is explicit and/or the interface is nonhuman. This finding is important because it points to a potential *glass ceiling* for artificial agent participation in human intersubjectivity. If roboticists someday build a machine that is indistinguishable from an actual human in terms of appearance and communication (i.e., if the machine were able to pass a *Total* Turing Test; see Harnad, 1991), the mere knowledge of it being something “artificial” might suppress the amount of intersubjective effort people exert when interacting with it.

Surprisingly, the likelihood of an interactant attempting to self-repair a trouble source in response to an other-initiation was not affected by the experimental manipulations whereas the rate of interactant other-initiation was. This could be because other-initiations of repair are active attempts to manipulate the attention of an interlocutor toward a trouble source and, therefore, at some level indicate an implicit supposition that an interlocutor possesses a capacity for advanced intersubjectivity. A self-repair attempt, on the other hand, is more of a reflexive response that follows a request to update an interlocutor's meta-perspective; a person need not presuppose that an interlocutor possesses a capacity for advanced intersubjectivity in order to produce a successful self-repair following other-initiation (in fact, it may even be that *not* attempting an appropriate self-repair following other-initiation, say by departing from the conversational trajectory, is more unnatural for humans than simply producing a self-repair attempt). Relative to the performance of a self-repair attempt, articulating an other-initiation of repair may involve higher-order mentalizing about the perspectives of an interlocutor, and this higher-order mentalizing may be more sensitive to changes in how the interlocutor is represented.

These findings can be positioned within a broader discussion that concerns the centrality of intersubjectivity and intersubjective effort in human life. As argued herein, and has been established in both the fields of developmental psychology and communication, complex intersubjectivity is a co-construction (one interactant cannot accomplish it alone). Child development, for instance, requires children be brought into advanced intersubjective relations by-way-of adult scaffolding (Berk & Winsler, 1995; Plumert & Nichols-Whitehead, 1996), which entails a high level of intersubjective effort. Adults scaffold by providing a level of verbal guidance and support for children's understanding that is *just beyond* their actual level of comprehension, thus pulling them into increasingly complex intersubjective relations (the “zone of proximal development”; see Vygotsky, 1978). Equally, achieving common ground in communicative interaction between adults cannot occur solely based on the actions or cognitions of one side of the interaction; rather it is a joint achievement, with each side supporting the other side in the calibration of perspectives (Rommetveit, 1974; Schegloff, 1992). Rommetveit (1974) axiom that “intersubjectivity has […] to be taken for granted in order to be achieved" (p. 56) captures this notion. If humans do not initially *assume* a highly intersubjective interlocutor, they will not engage in the complex intersubjective processes with the interlocutor that are necessary to further elaborate the pair's intersubjectivity. Even when misunderstanding arises, other-initiations of repair reveal an implicit belief that common ground *can* be achieved by-way-of an exchange of perspectives. On the other hand, abandoning the assumption of intersubjectivity will block the achievement of common ground due to fewer attempts at repair.

If human interactants do not first expect high-level intersubjectivity from artificial agents, they will not extend such intersubjectivity to them, effectively locking them out of the full spectrum of human intersubjective relations. In order to develop forms of HAI that reach benchmark intersubjectivity, agents – as with human infants – will need to be able to learn from the other-initiations of repair issued by their human interactants. It is easy to imagine how severely constrained human social relations would become if interlocutors repeatedly failed to signal to each other when something has caused misunderstanding (complex joint activity would be impossible). No matter how capable each party to an interaction is at operating at benchmark intersubjectivity, it takes benchmark intersubjective effort – a robust exchanging of perspectives – to get there.

5.1. *The echoborg method and analysis of repair activity as a means of benchmarking intersubjectivity in human-agent dialog*

Given the primacy of intersubjectivity in cooperative human social behavior, it is imperative that evaluative frameworks for HAI generally, and human-agent dialog specifically, involve assessing intersubjectivity against human–human interaction benchmarks. Although various researchers have explicitly called for this approach (e.g., Cassell & Tartaro, 2007), numerous others have indirectly called for a focus on intersubjectivity through advocacy of interactionist HAI methodologies (e.g., Dautenhahn, 2007; Johnson et al., 2014; Parise, Kiesler, Sproull, & Waters, 1999; Payr, 2001), including but not limited to the analysis of grounding behavior in human-agent dialog (e.g., Brennan, 1991; Kiesler, 2005; Lücking & Mehler, 2014; Visser, Traum, DeVault, & op den Akker, 2014) and interaction authenticity (e.g., Feil-Seifer, Skinner, & Matarić, 2007; Kahn et al., 2007; Turkle, 2007). Indeed, a concern with intersubjectivity is implicit in many approaches to HAI evaluation, such as those that investigate people's emotional responses to robotic and virtual agents (e.g., Balzarotti, Piccini, Andreoni, & Ciceri, 2014; Brave, Nass, & Hutchinson, 2005; Prendinger & Ishizuka, 2005).

The current article is a contribution toward developing methodologies for benchmarking intersubjectivity and intersubjective effort in HAI. Herein it is demonstrated how the echoborg method can be used to evaluate human-agent intersubjectivity when the agent is an artificial conversational agent. The unique strength of the echoborg method is that it can involve research participants communicating with an agent *while believing that they are speaking to an autonomous human being*. Thus, the echoborg method allows researchers the ability to investigate HAI intersubjectivity while preserving the interactant's sense that they are experiencing a fully human–human social psychological context.

5.2. *Might intersubjective effort in human-agent dialog increase as technology improves?*

The seeming unwillingness of interactants to exert benchmark intersubjective effort with a conversational agent in the non-baseline conditions of the present research may prove to be a historical artifact. As more advanced artificial intelligence develops and as people are raised in a world in which socially advanced artificial agents are ubiquitous, the expectations people will place on the intersubjective capacities of their machine interlocutors may increase. The study of social psychological phenomena is in many respects the study of behavioral patterns contingent upon cultural and historical circumstances (Gergen, 1973). The more that artificial agents are able to engage humans in complex intersubjective processes, such as repair work, the more humans might be willing to scaffold their participation in rich intersubjectivity. Indeed, the finding that interactants self-repair at relatively consistent rates irrespective of the interface they engage with or their awareness of the agency of an interlocutor is evidence that humans readily "step up" and respond with a repair attempt as needed.

5.3. *Limitations*

A particular strength of the echoborg method, which itself is derived from the "cyranoid method" of social interaction (Corti & Gillespie, 2015b), is that it enables the study of social interactions that are high in mundane realism (dynamic, unscripted and face-to-face). The benefits of such realism, however, incur costs in the form of certain control limitations. For instance, though best efforts were made to ensure that the speech shadower's body language was consistent across experimental trials (the shadower was instructed to try and match their body language to the words they found themselves articulating), it is all but impossible to completely eliminate variability in shadower body language from trial to trial using the echoborg method. Furthermore, the minimal technological dependency format of inputting interactants' words into Cleverbot created slightly unnatural delays between interactant utterances and subsequent responses by the agent. Though all experimental conditions were subject to the same latencies, future improvements to the echoborg method may mitigate this limitation.

6. Conclusion

The present research has found that two factors significantly affect the rate at which an interactant will attempt to elicit repairs of misunderstandings through other-initiation utterances from an interlocutor that is a conversational agent, namely, (1) the agent's interface (i.e., its embodied nature - the means by which it communicates) and (2) the interactant's awareness that their interlocutor's words are agent-determined. These factors, however, do not seem to affect the rate at which interactants attempt to self-repair misunderstandings following other-initiations. Given the operationalization of intersubjective effort argued for herein (i.e., a commitment to repair misunderstandings), it seems that people exert the most intersubjective effort with an agent-interlocutor when they are unaware that their interlocutor's words are agent-determined and when the agent interfaces via an actual human body.

This article contributes a novel methodology (i.e., the echoborg method) to the study of HAI and demonstrates how it can be used to compare conditions of HAI that make use of nonhuman interfaces and nonhuman agency-framing to a baseline condition that approximates the social psychological contextual frame experienced by people during mundane, face-to-face, in person, human–human interaction. This article's findings have implications for the development of autonomous social agents. Most notably, if agents are to participate fully in the intersubjective world of humans, not only must *they* be *capable* of interacting at benchmark intersubjectivity, but human interactants must also be willing to exert intersubjective effort at a level conducive for the achievement of benchmark intersubjectivity (e.g., by uttering other-initiations of repair when misunderstanding arises). Artificial agents cannot enter the world of human intersubjectivity without the support of their human interactants, and this support is contingent upon interactants' supposition that complex intersubjectivity is achievable.

References

Alterman, R. (2007). Representation, interaction, and intersubjectivity. *Cognitive Science, 31*(5), 815–841. http://dx.doi.org/10.1080/03640210701530763.

Arundale, R. B. (2010). Constituting face in conversation: face, facework, and interaction achievement. *Journal of Pragmatics, 42*(8), 2078–2105. http://dx.doi.org/10.1016/j.pragma.2009.12.021.

Bailenson, J. N., & Yee, N. (2005). Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological Science, 16*(10), 814–815. http://dx.doi.org/10.1111/j.1467-9280.2005.01619.x.

Balzarotti, S., Piccini, L., Andreoni, G., & Ciceri, R. (2014). "I know that you know how I feel": behavioral and physiological signals demonstrate emotional attunement while interacting with a computer simulating emotional intelligence. *Journal of Nonverbal Behavior, 38*(3), 283–299. http://dx.doi.org/10.1007/s10919-014-0180-6.

Berk, L. E., & Winsler, A. (1995). *Scaffolding children's learning: Vygotsky and early childhood education*. Washington, DC: National Association for the Education of Young Children.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition, 121*(1), 41–57. http://dx.doi.org/10.1016/j.cognition.2011.05.011.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: Investigating the effects of orientation of emotion exhibited by an embodied computer agent. *International Journal of Human-Computer Studies, 62*(2), 161–178. http://dx.doi.org/10.1016/j.ijhcs.2004.11.002.

Brennan, S. E. (1991). Conversation with and through computers. *User Modeling and User-Adapted Interaction, 1*(1), 67–86. http://dx.doi.org/10.1007/bf00158952.

Caissie, R., & Gibson, C. L. (1997). The effectiveness of repair strategies used by people with hearing losses and their conversational partners. *Volta Review, 99*(4), 203–218.

Carpenter, R. (2015). *Cleverbot* [software]. Retrieved from http://www.cleverbot.com.

Cassell, J. (2000). Embodied conversational interface agents. *Communications of the ACM, 43*(4), 70–78. http://dx.doi.org/10.1145/332051.332075.

Cassell, J., & Tartaro, A. (2007). Intersubjectivity in human-agent interaction. *Interaction Studies, 8*(3), 391–410. http://dx.doi.org/10.1075/is.8.3.05cas.

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, E., Cheng, G., et al. (2012). How do we think machines think? an fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience, 6*(103), 1–9. http://dx.doi.org/10.3389/fnhum.2012.00103.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 222–233). Washington, DC: American Psychological Association.

Clark, H. H., & Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes, 2*(1), 19–41. http://dx.doi.org/10.1080/01690968708406350.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science, 13*(2), 259–294. http://dx.doi.org/10.1207/s15516709cog1302_7.

Corti, K., & Gillespie, A. (2015a). A truly human interface: Interacting face-to-face with someone whose words are determined by a computer program. *Frontiers in Psychology, 6*(634), 1–18. http://dx.doi.org/10.3389/fpsyg.2015.00634.

Corti, K., & Gillespie, A. (2015b). Revisiting Milgram's cyranoid method: Experimenting with hybrid human agents. *Journal of Social Psychology, 155*(1), 30–56. http://dx.doi.org/10.1080/00224545.2014.959885.

Corti, K., & Gillespie, A. (2015c, May 11). *The "echoborg method" of human-agent interaction* [Video file]. Retrieved from https://www.youtube.com/watch?v=NtWLCZZYM64.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. In W. D. Gray, W. E. Hefley, & D. Murray (Eds.), *Proceedings of the 1st International Conference on intelligent user interfaces* (pp. 193–200). New York, NY: ACM.

Dautenhahn, K. (2007). Socially intelligent robots: Dimensions of human-robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences, 362*(1480), 679–704. http://dx.doi.org/10.1098/rstb.2006.2004.

De Preester, H. (2008). From *ego* to *alter ego*: Husserl, Merleau-Ponty and a layered approach to intersubjectivity. *Phenomenology and the Cognitive Sciences, 7*(1), 133–142. http://dx.doi.org/10.1007/s11097-007-9056-0.

Drew, P. (1997). "Open" class repair initiators in response to sequential sources of troubles in conversation. *Journal of Pragmatics, 28*(1), 69–101. http://dx.doi.org/10.1016/s0378-2166(97)89759-7.

Feil-Seifer, D., Skinner, K., & Matarić, M. J. (2007). Benchmarks for evaluating social assistive robotics. *Interaction Studies, 8*(3), 423–439. http://dx.doi.org/10.1075/is.8.3.07fei.

Frohlich, D., Drew, P., & Monk, A. (1994). Management of repair in human-computer interaction. *Human-Computer Interaction, 9*(3–4), 385–425. http://dx.doi.org/10.1080/07370024.1994.9667211.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3), 814–824. http://dx.doi.org/10.1006/nimg.2002.1117.

Garfinkel, H. (1967). *Studies in ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall.

Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology, 26*(2), 309–320. http://dx.doi.org/10.1007/978-1-4615-8765-1_2.

Gillespie, A., & Cornish, F. (2010). Intersubjectivity: towards a dialogical analysis. *Journal for the Theory of Social Behaviour, 40*(1), 19–46. http://dx.doi.org/10.1111/j.1468-5914.2009.00419.x.

Gillespie, A., & Cornish, F. (2014). Sensitizing questions: a method to facilitate analyzing the meaning of an utterance. *Integrative Psychological and Behavioral Science, 48*(4), 435–452. http://dx.doi.org/10.1007/s12124-014-9265-3.

Goodwin, C. (2003). Introduction. In C. Goodwin (Ed.), *Conversation and brain damage* (pp. 3–20). Oxford, UK: Oxford University Press.

Goodwin, C., & Heritage, J. (1990). Conversation analysis. *Annual Review of Anthropology, 19*, 283–307. http://dx.doi.org/10.1146/annurev.an.19.100190.001435.

Graziano, M. S. A. (2013). *Consciousness and the social brain*. Oxford, UK: Oxford University Press.

Harnad, S. (1991). Other bodies, other minds: a machine incarnation of an old philosophical problem. *Minds and Machines, 1*(1), 43–54.

ten Have, P. (1999). *Doing conversation analysis: A practical guide*. London, UK: Sage Publications.

Hemberg, K. (2006). *Husserl's phenomenology: Knowledge, objectivity and others*. London, UK: Continuum International Publishing Group.

Husserl, E. (1931). *Cartesian meditations: An introduction to phenomenology*. Dordrecht, The Netherlands: Kluwer.

Hutchby, I., & Wooffitt, R. (2008). *Conversation analysis*. Cambridge, UK: Polity Press.

Icheiser, G. (1943). Structure and dynamics of interpersonal relations. *American Sociological Review, 8*(3), 302–305. http://dx.doi.org/10.2307/2085084.

Ishiguro, H., & Nishio, S. (2007). Building artificial humans to understand humans. *Journal of Artificial Organs, 10*(3), 133–142. http://dx.doi.org/10.1007/s10047-007-0381-4.

Jacoby, S., & Ochs, E. (1995). Co-construction: an introduction. *Research on Language and Social Interaction, 28*(3), 171–183. http://dx.doi.org/10.1207/s15327973rlsi2803_1.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., & Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction, 3*(1), 43–69. http://dx.doi.org/10.5898/JHRI.3.1.Johnson.

Kahn, P. H., Jr., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., et al. (2007). What is a human? toward psychological benchmarks in the field of human-robot interaction. *Interaction Studies, 8*(3), 363–390. http://dx.doi.org/10.1075/is.8.3.04kah.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies, 7*(2), 135–169. http://dx.doi.org/10.1075/is.7.2.04kap.

Kennedy, A., Wilkes, A., Elder, L., & Murray, W. S. (1988). Dialogue with machines. *Cognition, 30*(1), 37–72. http://dx.doi.org/10.1016/0010-0277(88)90003-0.

Kiesler, S. (2005). Fostering common ground in human-robot interaction. In *Proceedings of the 14th IEEE International Workshop on robot and human Interactive communication (Ro-Man 2005), Nashvilee, TN* (pp. 729–734). http://dx.doi.org/10.1109/roman.2005.1513866.

Kircher, T., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J.… Krach, S. (2009). Online mentalising investigated with functional MRI. *Neuroscience Letters, 454*(3), 176–181. http://dx.doi.org/10.1016/j.neulet.2009.03.026.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLoS One, 3*(7), e2597. http://dx.doi.org/10.1371/journal.pone.0002597.

Laing, R. D., Phillipson, H., & Lee, A. R. (1966). *Interpersonal perception: A theory and method of research*. London, UK: Tavistock Publications.

Linell, P. (2009). *Rethinking language, mind and world dialogically*. Charlotte, NC: Information Age Publishing.

Lücking, A., & Mehler, A. (2014). On three notions of grounding of artificial dialog companions. *Science, Technology and Innovation Studies, 10*(1), 31–46.

MacDorman, K. F. (2006). Introduction to the special issue on android science. *Connection Science, 18*(4), 313–317. http://dx.doi.org/10.1080/09540090600906258.

MacDorman, K. F., & Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies, 7*(3), 297–337. http://dx.doi.org/10.1075/is.7.3.03mac.

Marková, I. (2003). Consitution of the self: Intersubjectivity and dialogicality. *Culture and Psychology, 9*(3), 249–259. http://dx.doi.org/10.1177/1354067x030093006.

Parise, S., Kiesler, S., Sproull, L., & Waters, K. (1999). Cooperating with life-like interface agents. *Computers in Human Behavior, 15*(2), 123–142. http://dx.doi.org/10.1016/s0747-5632(98)00035-1.

Payr, S. (2001). The virtual other: aspects of social interaction with synthetic characters. *Applied Artificial Intelligence, 15*(6), 493–519. http://dx.doi.org/10.1080/088395101753199551.

Plumert, J. M., & Nichols-Whitehead, P. (1996). Parental scaffolding of young children's spatial communication. *Developmental Psychology, 32*(3), 523–532. http://dx.doi.org/10.1037//0012-1649.32.3.523.

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: a character-based interface that addresses users' affective states. *Applied Artificial Intelligence, 19*(3–4), 267–285. http://dx.doi.org/10.1080/08839510590910174.

Raudaskoski, P. (1990). Repair work in human-computer interaction: a conversation analytic perspective. In P. Luff, N. Gilbert, & D. Frohlich (Eds.), *Computers and conversation* (pp. 151–172). London, UK: Academic Press.

Riek, L. D., Rabinowitch, T., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In M. Scheutz, & F. Michaud (Eds.), *Proceedings of the 4th ACM/IEEE International Conference on human-robot interaction* (pp. 245–246). New York, NY: ACM Press.

Rommetveit, R. (1974). *On message structure: A framework for the study of language and communication*. London, UK: John Wiley and Sons.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735. http://dx.doi.org/10.1353/lan.1974.0010.

Saygin, A. P., & Stadler, W. (2012). The role of appearance and motion in action prediction. *Psychological Research, 76*(4), 388–394. http://dx.doi.org/10.1007/s00426-012-0426-z.

Schegloff, E. A. (1992). Repair after next turn: the last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology, 97*(5), 1295–1345. http://dx.doi.org/10.1086/229903.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction, 26*(1), 99–128. http://dx.doi.org/10.1207/s15327973rlsi2601_5.

Schegloff, E. A. (1997). Practices and actions: Boundary cases of other-initiated repair. *Discourse Processes, 23*(3), 499–545. http://dx.doi.org/10.1080/01638539709545001.

Schegloff, E. A. (2000). When 'others' initiate repair. *Applied Linguistics, 21*(2), 205–243. http://dx.doi.org/10.1093/applin/21.2.205.

Schegloff, E. A. (2007). *Sequence organization in interaction: A primer in conversation analysis* (Vol. 1). Cambridge, UK: Cambridge University Press.

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in theorganization of repair in conversation. *Language, 53*(2), 361–382. http://dx.doi.org/10.1353/lan.1977.0041.

Schönbrodt, F. D., & Asendorpf, J. B. (2011). The challenge of constructing psychologically believable agents. *Journal of Media Psychology, 23*(2), 100–107. http://dx.doi.org/10.1027/1864-1105/a000040.

Sidnell, J. (2010). *Conversation analysis: An introduction*. Chichester, UK: Wiley-Blackwell.

Skelt, L. (2012). Dealing with misunderstandings: the sensitivity of repair in hearing impaired conversation. In M. Egbert, & A. Deppermann (Eds.), *Hearing aids communication: Integrating social interaction, audiology and user centered design to improve communication with hearing loss and hearing technologies* (pp. 56–66). Mannheim, Germany: Verlag.

Tirassa, M., & Bosco, F. M. (2008). On the nature and role of intersubjectivity in human communication. In F. Morganati, A. Carassa, & G. Riva (Eds.), *Enacting intersubjectivity: A cognitive and social perspective on the study of interactions* (pp. 81–95). Amsterdam, The Netherlands: IOS Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: the origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 721–727. http://dx.doi.org/10.1017/s0140525x05000129.

Trevarthen, C., & Aitken, K. J. (2001). Infant intersubjectivity: research, theory, and clinical applications. *Journal of Child Psychology and Psychiatry, 42*(1), 3–48. http://dx.doi.org/10.1111/1469-7610.00701.

Turkle, S. (2007). Authenticity in the age of digital companions. *Interaction Studies, 8*(3), 501–517. http://dx.doi.org/10.1075/is.8.3.11tur.

Visser, T., Traum, D., DeVault, D., & op den Akker, R. (2014). A model for incremental grounding in spoken dialogue systems. *Journal of Multimodal User Interfaces, 8*(1), 61–73. http://dx.doi.org/10.1007/s12193-013-0147-7.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Zahn, C. J. (1984). A reexamination of conversational repair. *Communication Monographs, 51*(1), 56–66. http://dx.doi.org/10.1080/03637758409390183.

Zdenek, S. (2001). Passing Loebner's Turing Test: a case of conflicting discourse functions. *Minds and Machines, 11*(1), 53–76. http://dx.doi.org/10.1023/A:1011214808628.