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

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.

© 2015 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

"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).

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.