Natural Behavior of a Listening Agent

Martijn Maatman, University of Twente
Jonathan Gratch, USC
Stacy Marsella, USC

Building Rapport with Virtual Humans

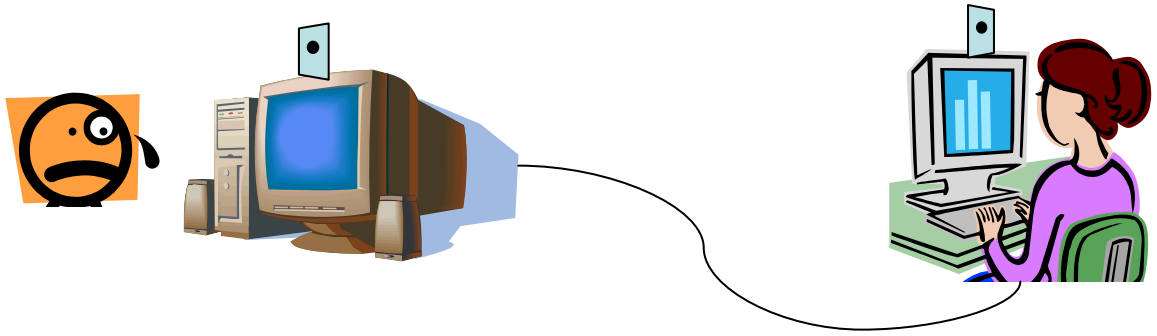
- Successful face-to-face encounters depend on establishing rapport
- Virtual Humans (ECAs) fail miserably
- Review psycholinguistic findings on how to get it
- Present system that attempts to achieve rapport with users via “active listening”

Building Rapport with Virtual Humans

- Successful face-to-face encounters are richly interactive
  - Involve tight reciprocity, coordination, feedback [Bavelas]
  - Interactions rapid and outside conscious awareness [Bargh]

Building Rapport with Virtual Humans

- Successful face-to-face encounters are richly interactive
  - Involve tight reciprocity, coordination, feedback [Bavelas]
  - Interactions rapid and outside conscious awareness [Bargh]



Trevarthen demonstrated a baby will interact happily with a live TV image of its mother who can also see the baby's image in real time but if a videotape of the mother showing exactly the same information is shown, the baby gets upset, presumably because its synchrony with the mother is lost.

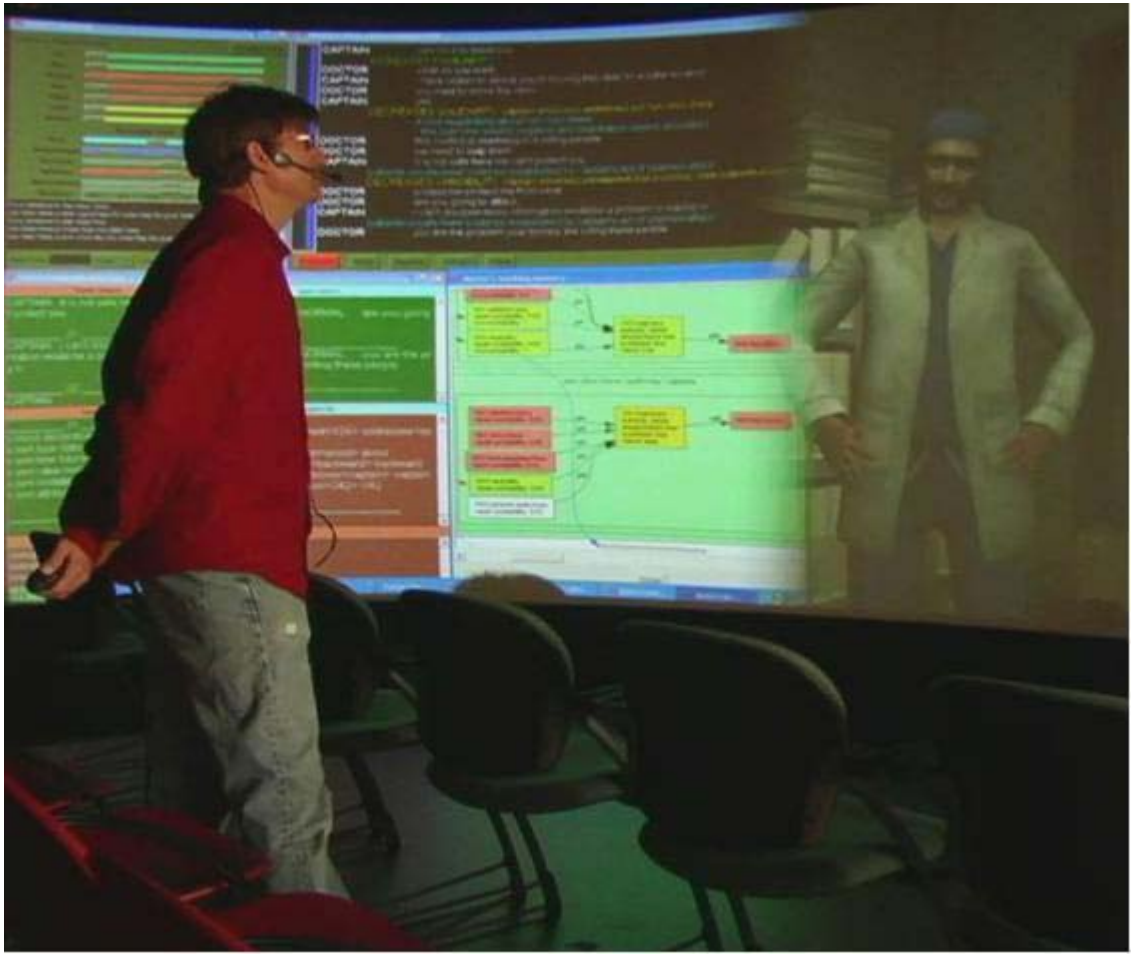
Building Rapport with Virtual Humans

- Successful face-to-face encounters are richly interactive
  - Involve tight reciprocity, coordination, feedback [Bavelas]
  - Interactions rapid and outside conscious awareness [Bargh]
- Presence of interactive behaviors predict “rapport”
  - Encounters with these behaviors seem qualitatively different
  - Participants strong agree if rapport established
  - Observers can detect it from “thin slices” of nonverbal behavior

Building Rapport with Virtual Humans

- Successful face-to-face encounters are richly interactive
  - Involve tight reciprocity, coordination, feedback [Bavelas]
  - Interactions rapid and outside conscious awareness [Bargh]
- Presence of interactive behaviors predict "rapport"
  - Encounters with these behaviors seem qualitatively different
  - Participants strong agree if rapport established
  - Observers can detect it from "thin slices" of nonverbal behavior
- Rapport correlates with socially desirably outcomes
  - Liking, affiliation [Chartrand, Lakin]
  - Communicative efficiency and comprehension
  - Greater social influence (success of therapeutic outcomes)

Rapport with Virtual Humans



Vhumans “interactionally challenged”

- Systems are blind and tone deaf

  Typical user input is a string of words
  - Tend to ignore prosody, intensity, affect
  - Tend to ignore posture, gesture and facial expression

- NLP techniques non-incremental
  - Semantic information only available after utterance processed

- Exceptions are few, piecemeal, and “loose”
  - Many systems detect “start-of-speech” & “end-of-speech”
  - STEVE (Rickel), MACK (Cassell) responded to user gaze
  - Kismit (Breazeal) and Neurbaby (Tosa) analyze speech intonation
  - Bickmore and Cassell’s REA detects pauses/disfluency
  - Kopp – Incremental speech recognition

- Consequence: low ratings of rapport (no one bothers to ask)

*Apologies to Sue Duncan

Challenge: Passing the "Duncan Test"*

Watch a Cartoon

Describe it to a human listener

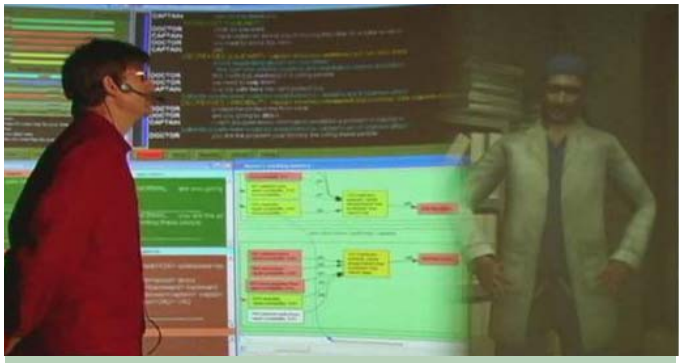
Demonstrate positive correlations between

- listening behaviors
- self reports of rapport
- observer judgments of rapport
- social outcomes: liking, affiliation….

*Apologies to Sue Duncan

Challenge: Passing the "Duncan Test"*

Watch a Cartoon



Describe it to **an agent listener**

Demonstrate positive correlations between

- listening behaviors
- self reports of rapport
- observer judgments of rapport
- social outcomes: liking, affiliation….

How to pass the test

- Could attempt rich model of information structure, collaborative monitoring and coordination: Cassell, Heylen05

- Could explore creating the “illusion of rapport”

How to cheat the test

- Review of psycholinguists of listening
  - Listening behavior associated with rapport (B)
  - Elicitation conditions (E)
  - The posited interactional function (F)
    - Where F consistent with “Keep talking”
      (in contrast to “take the turn”, “end the conversation”, etc.)
- Propose shallow behavioral mapping rules E→B
  - Abstract (and grossly simplify) psycholinguistic findings
  - Emphasize non-semantic features available in real time
  - “Compile in” the interactional function

Backchannel Continuers

- Nods and paraverbals (uh-huh) that occur during speech
- Several models of elicitation
  - Lowering of pitch (Ward)
  - Raised Loudness (Cassell)
  - Pause duration & part-of-speech frequency (Cathcart et al.)
- Signals that communication is working and that the speaker should continue (Yngve)

Rule: Lowering of pitch in speech signal → head nod

Rule: Raised loudness in speech signal → head nod

Disfluency Responses

- Behaviors that occur during speaker disfluency
  - Posture shifts
  - Gaze shifts
  - Frowns
- Disfluency includes
  - repetition, spurious words, pauses and filled pauses
- Signals that speaker should “take their time”
  - Disfluencies indicate speaker is experiencing processing difficulty (Clark&Wasow; Ward and Tsukahara)

Rule: Disfluency in speech signal → posture/gaze shift, frown

Mimicry (Behavioral Synchrony)

- Listeners often mimic speaker behavior
  - postures, gestures, mannerisms, rhythm, breathing
- Elicited when speaker agrees with message / messenger
  - Speaker is a friend [Duncan]
  - Speaker is trusted, respected
- Can increase liking, affiliation and social bonds
  - May synchronize conversational flow
  - May signal acceptance, trust

Rule: Speaker shifts posture → Mimic Posture

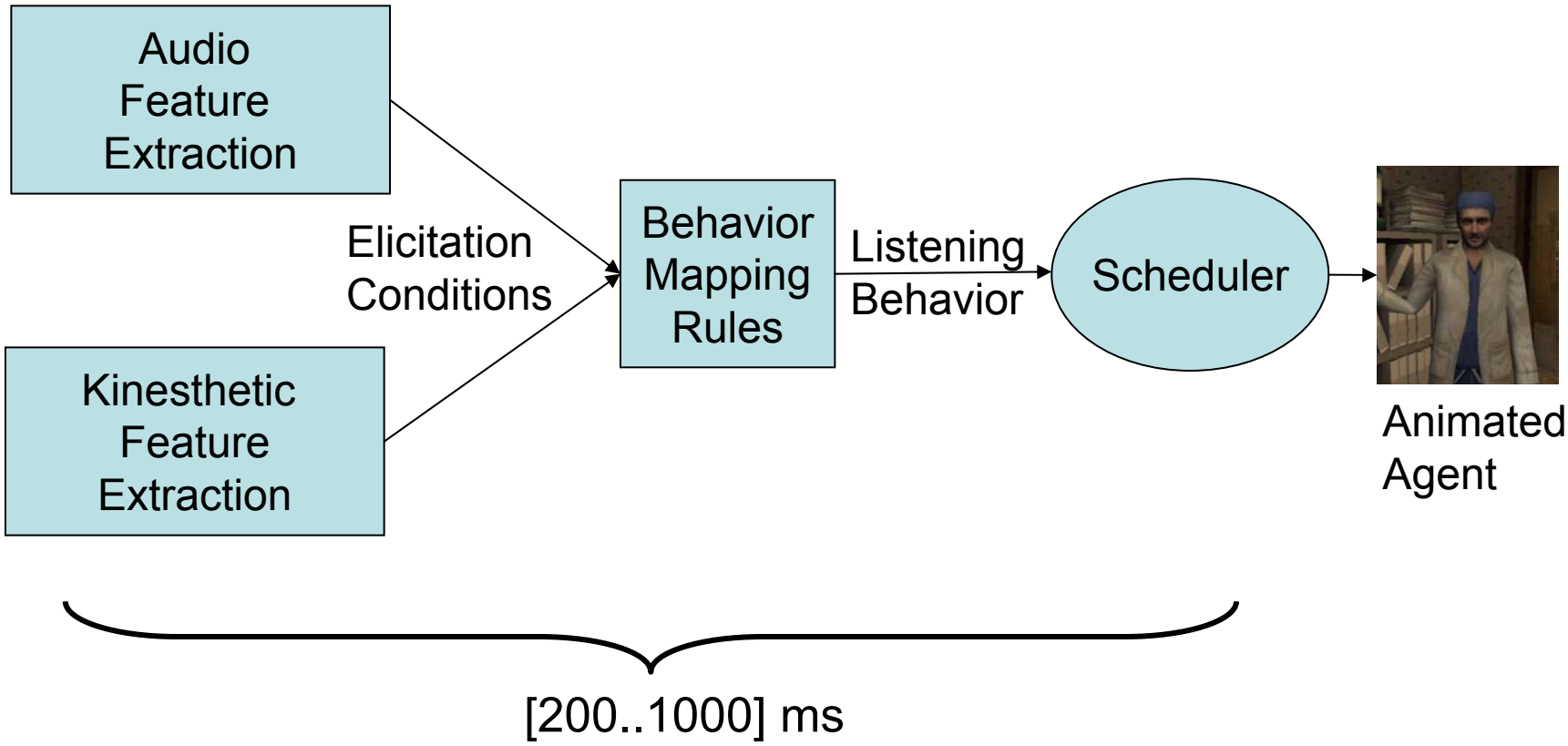
Rule: Speaker performs head gesture → Mimic Head Gesture

Other Factors

- Ignoring some understood behaviors
  - Facial expression
- Ignoring important contextual factors
  - Power
  - Gender
  - Race
  - Culture
  - Affiliation

Behavior mapping rules are “social context”-free

Towards a Listening Agent

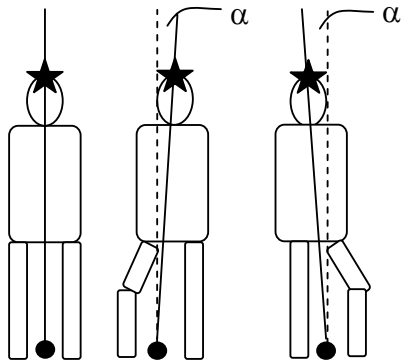


Acoustic Feature Detection

- Intensity Detection
  - Report intensities in upper-tail of baseline speech
- Pitch Detection
  - Adopt approach of Ward and Tsukahara 2000
  - Detect significant drop over last 120 milliseconds
- Disfluency Detection
  - Adopt method of by Shriberg (1999)
  - Detects filled or unfilled pauses
  - Standard deviation of frequencies over last 200ms < 1 Hz

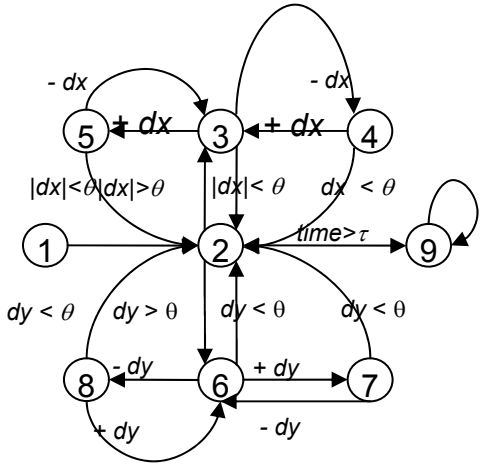
Kinesthetic Feature Detection

- Utilize 6 degree-of-freedom head tracker



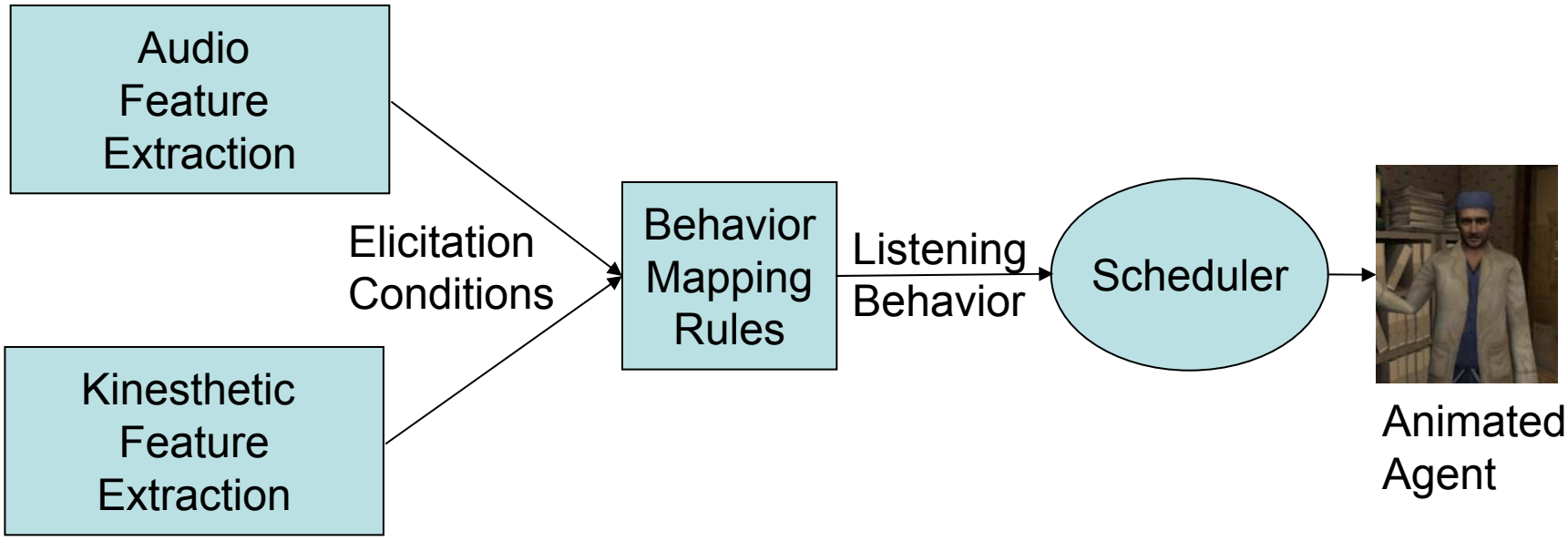
- Body posture
  - Fix location of feet
  - Just detects side-to-side posture shifts

- Head gestures
  - finite state transition graph
  - Transition on rotational angle



Scheduler

- Trivially integrates asynchronous behavior requests
  - 2 channels of behavior – head and posture
  - Discards behavior if channel currently performing a behavior



Limations

- Impact on Rapport untested
  - Formal evaluation pending
  - Some positive anecdotal feedback
  - Does great job of nodding in time to rock music
- Literally mindless feedback
  - Could be judged insincere
  - Considering how to integrate with semantic feedback
- Social context-free behavior rules
  - Ignores social, cultural, context

Status

- Just resumed this month
  - 3 interns starting September 15
- Replacing kinesthetic detection approach
  - Exploring camera based approaches (MIT, Twente)
  - Evaluating nod detection code of Moreny&Sidner,2005
- Evaluating alternative audio feature detection
  - Including affect detection by Narayanan
- Exploring cultural differences
- Formal evaluations in January

Conclusion

- Created a simple framework for face-to-face interaction
- Could help improve rapport, and arguably improve effectiveness of training, psychotherapy, etc.
- Could be a tool for testing psycholinguistic theory
  - Explore and test different mappings
  - More rigorously control social contextual factors (Race, etc.)

Acknowledgements

- We gratefully acknowledge the help and support of Sue Duncan, Roz Picard, Shri Narayanan, David Traum, Raul Galt and Regina Cabrera

- Thanks to Anton Nijholt for sending us interns

This work was sponsored by the U. S. Army Research, Development, and Engineering Command (RDECOM), and the content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Methodology

- 2x2 design
  - Contingent vs. Non-contingent (“yoked”) feedback
  - Agent vs. Human framing

  - Ultimatum game
  - Questionnaire: Manstead

Motivate “strong” interactivity

Rapport matters

Need to be deeply interactive to get it

Interactivity helps

- Pickering & Garrod explain via priming: I prime off of what you just said and through the priming I"m 50% of the way to understand you. People very quickly become in sync. Argue that speech in dialogue is much simplier, more reptitious than monologue/written speech.
- Malinowski "phatic communion" - looked at greetings, the things people do to come in sync w/ each other. Laver looked an not just intro stuff but stuff that goes throughout the interaction
- Feedback:Duncan: speaker gestures little until listener smiles
- Bargh, chartrand, larkin: "social glue" and "sense of affiliation“. impact on human evolution - gave us ability to form social groups

- Successful Face-to-face encounters are deeply interactive
- Interactive matters
  - Influence subsequent behavior
  - Help achieve schroncy

  Very tight interactive loop between speaker and listener (100ms)
  - Colwyn Trevarthen's demonstration that a baby will interact happily with a live TV image of its mother who can also see the baby's image in real time, but if a videotape of the mother showing exactly the same information is shown, the baby gets upset, presumably because its synchrony with the mother is lost. Apparently the baby directly perceives its synchrony with the mother and knows immediately when the mother's behavior is out of synch.

- Jon E. Grahe and Frank J. Bernieri: This study examined the relative impact different channels of communication had on social perception based on exposure to thin slices of the behavioral stream. Specifically, we tested the hypothesis that dyadic rapport can be perceived quickly through visual channels. Perceivers judged the rapport in 50 target interactions in one of five stimulus display conditions: transcript, audio, video, video+transcript or video + audio. The data demostrated that perceivers with access to nonverbal, visual information were the most accurate perceivers of dyadic rapport. Their judgements were found to covary with the visually encoded features that past research has linked with rapport expression. This suggests a presence of a nonverbally based implicit theory of rapport that more or less matches the natural ecology, at least as it occurs within brief samples of the behavioral stream.
- Such features as interpersonal proximity, synchrony and forward lean are linked to perceived rapport (Tickle-Dengen & Rosenthal, 1990). Nonverbal behavior may be relatively more important than verbal behavior in certain areas: in the expression and communication of spontaneous affect (Argyle, Salter); in the assessment of self-presentation and communication motives (DePaulo 1992); in the expression and communication of rapport and the related trait of extraversion (Funder and Colvin, 1998); and when perceptions are based on thin slices of behavior (Ambady and Rosenthal, 1992)

- The “chameleon effect” refers to the tendency to adopt the postures, gestures, and mannerisms of interaction partners (Chartrand & Bargh, 1999). This type of mimicry occurs outside of conscious awareness, and without any intent to mimic or imitate. Empirical evidence suggests a bi-directional relationship between nonconscious mimicry on the one hand, and liking, rapport, and affiliation on the other. That is, nonconscious mimicry creates affiliation, and affiliation can be expressed through nonconscious mimicry. We argue that mimicry played an important role in human evolution. Initially, mimicry may have had survival value by helping humans communicate. We propose that the purpose of mimicry has now evolved to serve a social function. Nonconscious behavioral mimicry increases affiliation, which serves to foster relationships with others. We review current research in light of this proposed framework and suggest future areas of research.