Communicating through Acting: The Role of Contextual Affordance in Intuitive Pantomimetic Gestural Communication

Siyi Gong[1] siyi.gong@ucla.edu
Kaiwen Jiang[2] jiangk11@msu.edu
Jessica G Li[3] jessli29@g.ucla.edu
Mireille Karadanaian[3] mireillekara@g.ucla.edu
Ziyi Meng[3,4] ziyimeng@g.ucla.edu
Tao Gao[1] taogao@ucla.edu

[1] Department of Communication, UCLA [2] Department of Statistics & Data Science, UCLA
[3] Department of Psychology, UCLA [4] Department of Political Science, UCLA

Abstract

When observing an action, how do people intuitively determine if it has communicative intent, such as in the case of pantomimes? We focus on two alternative theories: one suggests that instrumental intention competes with communicative intention, where the weaker the former is, the stronger the latter is; the other proposes that instrumental intention is nested within communicative intention, where the presence of the former facilitates the latter. To test these theories, we introduced the concept of contextual affordance, which manipulates the degree to which the instrumental components of an action are revealed. Through behavioral experimentation, we found a non-linear relationship between contextual affordance and communicativeness rating: partial affordance—providing implicit cues to an action's instrumental component without being fully rational—elicited the strongest perception of communicative intention, whereas full affordance or no affordance resulted in a weaker perception of communicative intention. Our study provides a novel definition of communicative intention and reveals that recognizing the instrumental goal and perceiving the suboptimality in achieving it work together to create a strong communicative signal. This work represents a step toward developing an integrated theory of pantomimes, specifying how the rationality principle can be applied to serve multiple purposes simultaneously.

Keywords: Communicative intention; Instrumental intention; Pantomimes; Rationality; Contextual affordance

Introduction

Among the various important purposes of human actions, two stand out as particularly crucial. Instrumental actions, or world-directed actions, aim to interact with and produce changes in the physical world, often immediately and efficiently (Weber, 1925; Habermass, 1984), such as kicking a ball. In contrast, communicative actions aim to convey information to others and influence their mental states (Searle, 1969; Habermass, 1984; Grice, 1957), like waving at a friend.

Pantomimes, a type of communicative action, involve intentionally creating bodily forms that imitatively represent something other than themselves (Żywiczyński, Wacewicz, & Sibierska, 2018), often depicting perceptually absent objects or actions by highlighting their distinctive features, functioning without spoken language (McNeill, 2012; Calbris, 2011; Tomasello, 2010). Pantomimes are notably iconic and transparent, allowing their meanings to be directly inferred from their forms (Jakobson, 1965), making them graspable even by novices. Pantomimes effectively communicate a range of complexities, from individual concepts (Gärdenfors, 2017) to events (Zlatev, Wacewicz, Zywiczynski, & van de Weijer, 2017) and narratives (Ferretti, 2021; Ferretti et al., 2017; Sibierska, 2017), making them a strong vehicle for information.

The capacity of pantomimes does not rely on pre-established meaning conventions (Żywiczyński et al., 2018) and is argued to be created and interpreted on the spot (Arbib, 2012, 2013; Poggi, 2007). Imagine you are in the car with your friend's car, in which the music is playing. Your friend has to pick up a phone call, and while she does that, she gestures to you the motion of adjusting a knob. No explicit communication was present in the scenario, but you understand perfectly her intention to ask you to turn down the volume. As demonstrated by this example, the motion trajectories of pantomimes may not occur frequently enough in daily life to become conventionalized. However, they are still intuitively recognized and universally communicable across different populations (McNeill, 1992; Goldin-Meadow, 2003; Tomasello, 2010; Żywiczyński et al., 2021).

At the same time, due to their "analogical link" to physical-cultural experiences (Calbris, 2011), pantomimes often share similar muscle movements with instrumental actions. Without understanding the underlying communicative intent, viewers may struggle to "quarantine" gestures from actions directing to change the world (Leslie, 1987; Tomasello, 2010). This raises a core question: How do observers intuitively discern whether an action is intended as a pantomime for communication?

We identify two approaches to addressing this question. One approach, which we termed the *parallel theory*, focuses on instances that violate the fundamental principles underlying instrumental actions, with one key principle being rationality. For humans, rational actions typically involve achieving external goals efficiently within the constraints of available resources (Simon, 1955; Bratman, 1987; Gopnik & Wellman, 1992; Dennett, 1989). Conversely, inefficiency in actions, as a violation of this principle, may signal that the actor has objectives beyond merely achieving their goal physically (Csibra & Gergely, 2007), such as intending to communicate. Exaggeration in actions is well documented across various communicative and educational contexts (Becchio, Sartori, & Castiello, 2010; Ho, Littman, MacGlashan, Cushman, & Austerweil, 2016; Trujillo, Simanova, Bekkering, & Özyürek, 2018; McEllin, Knoblich, & Sebanz, 2018). A recent study (Royka, Chen, Aboody, Huanca, & Jara-Ettinger, 2022) found that when repetitiveness in actions cannot be

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

easily explained by world-directed goals, observers are more likely to believe these actions serve communicative purposes. Building on similar research, this approach posits that communicative intent often runs parallel to instrumental intent: when an action clearly affects the physical world, it is typically interpreted instrumentally; however, deviations from this pattern prompt observers to attribute a communicative intent.

An alternative approach, termed *envelope theory* by us, emphasizes the importance of understanding the instrumental component of an action to accurately identify its communicative intention. In the volume-turning example, a communicative action operates on two levels: Firstly, it conveys an instrumental layer of world-directed information—the physical action of turning a knob. Secondly, it also communicates a layer of communicative intention, where the signaler desires the listener to recognize their intent to communicate—— the pragmatic meaning, "I want you to know that I need you to turn the knob" (Tomasello, 2010). To clarify, we assert that the joint inference of both levels—instrumental and communicative intentions—is critical for robust communication. This clarification does not specify the sequence in which these intentions occur. One implication of this hypothesis is that accurately recognizing the instrumental nature of an action supports its perceived communicativeness. For instance, identifying the action as adjusting the knob, given the context, helps interpret its communicative intent. Without this recognition, the joint inference weakens, making the gesture seem less like intentional communication directed at me. From this perspective, instrumental action is not an alternative to communicative actions, but a crucial foundation upon which communicative actions can be built.

Inspired by the envelope theory, we develop the concept of *contextual affordance*. Instead of focusing on the body movement of the actions themselves, contextual affordance specifies how much the physical environment provides cues about the instrumental purpose of an action performed, which can be operationally defined by the arrangement of objects in the scene where the action is observed. For example, the accessibility of a cup can provide different levels of affordance for the same action of drinking water: when the cup is in the actor's hand, it offers full affordance; when the cup is far from the actor's hand, it offers partial affordance; and when no cup is available, it provides zero affordance. For the same action, a stronger physical affordance indicates a stronger cue to its instrumental content. If communicative and instrumental intentions are simply two competing interpretations, then a weaker contextual affordance, where the action cannot be explained by a world-directed goal, should offer a stronger communicative power. Alternatively, according to the envelope theory, there can be an interesting non-linear relation between the context and communication strength: the communicative intent should reach its apex when the contextual affordance is barely sufficient to support some understanding of the instrumental intent of an action, since a full affordance maximally explains the action from a purely instrumental perspective, while no affordance weakens the understanding for instrumental content as the environment offers no clue.

This study aims to test the envelope and parallel theories by exploring how the perceived communicativeness of a mimetic action varies as a function of contextual affordance. Specifically, we manipulate different visual contexts of the same gestural action, defined by the accessibility of objects in a scene that is essential to the gestural action. A dataset of 30 action-object pairs was established, where each object can be operated by the action, such as drinking water (action) with a mug (object). For each action, we designed three contexts (Do, Far, Absent) to manipulate a scene's contextual affordance. The experiment aims to examine the patterns of perceived communicativeness and investigate the nuanced subjective experience for different contextual affordances.

Methods

The study was pre-registered on the Open Science Framework (OSF)[1] and received approval from the UCLA Office of the Human Research Protection Program. All methods were performed in accordance with the relevant guidelines and regulations.

Participants

Thirty-eight adults participated in this study. The data from thirty participants were used for analysis, while the data from the remaining eight participants were included in the training dataset (see Data pre-processing section). The sample size adopted in this study is based on a similar study (Royka et al., 2022). All participants were undergraduates at the University of California, Los Angeles, recruited from the Communication Department Subject Pool in exchange for course credit. This in-person experiment lasted approximately 20 minutes.

Materials

Thirty sets of stimulus videos were created, each set consisting of three videos of the same action, and each was filmed in one of the three context conditions, resulting in a total of ninety videos (see appendix for all actions in the Far context). Each video has a duration of one second. All videos were filmed in a lab room, featuring a blue background, an actress's upper body below her eyes, and a table surface in front of her. The filming angle was fixed for all videos, creating the impression that the viewer was sitting face-to-face with the actress, with the table in between.

Each set of videos consists of an object and an action corresponding to the object (see appendix). In the same set of videos, the action trajectory was kept identical, with the only difference between videos being the object's distance relative to the actress (Fig. 1). In the Do context, the object is placed in the actress's hand at the beginning of the video and moves

[1]Data, code, stimuli, registration, and appendix are available at the Files section of the project's OSF repository https://osf.io/5tby9/?view_only=a11ca168e1674e37a447856bb61a3037

with her hand. In the Far context, the object is within the actress's reach but remains stationary on the table without any contact with the actress. In the Absent context, the object is absent from the scene, leaving an empty table. Most actions involve the movement of the actress's hands and arms, with a few involving mouth movements (e.g., "blowCandle").

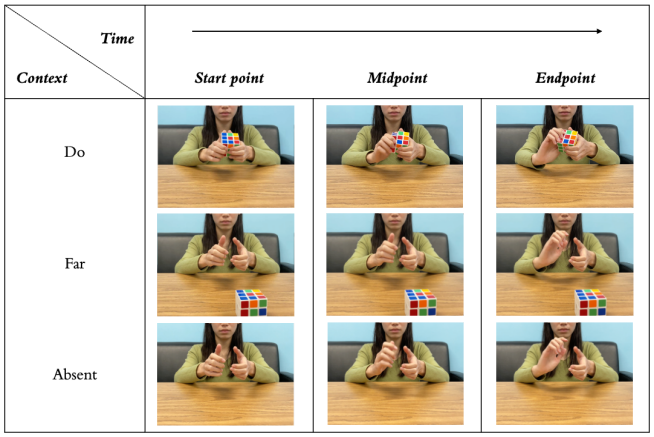

Figure 1: Illustration of the video screenshots for timeline start point (0 second), midpoint (0.5 second), and endpoint (1 second) of action "twistCube" in the Do (top), Far (middle), and absent (bottom) contexts. The action trajectory of the actress is strictly controlled for the same action across contexts.

The experiment was coded in jsPsych (de Leeuw, Gilbert, & Luchterhandt, 2023). Participants interacted with the experiment using a mouse, keyboard, and monitor (24-inch, 60 Hz, 1920 x 1080 screen resolution) from a distance of approximately 68 cm. The display entered full-screen mode for the entire study. Videos were displayed within an 18° x 14° area centered on the screen.

Design

Each participant viewed 30 videos, covering all 30 actions, with 10 videos presented in each context condition (Do, Far, and Absent). In other words, participants saw each action only once but experienced each context 10 times. For instance, if a participant watched the "twistCube" action in the Do context, they would not see it in the Far or Absent contexts. The selection of videos for each context was counterbalanced across every three participants using a Latin Square design. Once the videos were selected, their display order was randomly shuffled for each participant to minimize any potential confounding effects. This means that while one in every three participants saw the same set of videos, the display order differed for each of them.

Procedure

Participants were invited to the lab and seated in front of a monitor equipped with a mouse and keyboard. After providing informed consent, they were instructed to watch videos of a person performing certain actions and imagine themselves sitting in front of the person. They were also instructed to answer questions about this person and her actions. A comprehension check was administered before proceeding to the experiment phase.

The experiment phase consisted of 30 trials. Each trial began with a one-second fixation screen, followed by a one-second stimulus video displayed at the center of the screen, which played automatically without a playback option. Below the video, a slider bar allowed participants to use the mouse and rate "how communicative this video clip is" on a Likert scale (1 - definitely not communicating, 7 - definitely communicating). Note that it was up to the participants to decide the definition of communication. On the subsequent page, participants were presented with a text box where they were asked to type in a language description of "what was the person doing in the video?" They were instructed that there were no right or wrong answers and encouraged to provide their honest, subjective impressions. After typing their interpretation of the video, the participants were allowed to take as much time as they needed before pressing the "continue" button, which led to the next trial.

After the experiment phase, participants completed a questionnaire that inquired about their perceived difficulty and understanding of the experiment's purpose.

Data pre-processing

We conducted pre-processing on participants' language descriptions of their interpretations of the videos. We annotated the 900 free responses using a Large Language Model (LLM). The training dataset (see Participants section) was randomized and manually annotated by an experimenter who was blinded to the context and participant details. We fine-tuned a GPT-4o model on this annotation data. The responses were first evaluated to determine if they suggested an instrumental intention—actions that could potentially change the world. For example, "moving a paper" gets a "yes," while "no idea" or "shaping hand into a cone" gets a "no." This initial evaluation is termed the "any" response. If a description initially received a "yes" for identifying an instrumental intention, it was further assessed to determine if it correctly identified the intended action. This involved a binary decision on whether the described action resembled the intended one in terms of outcomes, allowing for differences so long as the semantics match. For instance, "tear crisps" would receive a "yes" for the intended action "openSnack," while "flip page" would get a "no." See appendix for more details.

Results

Communicativeness rating

The effect of context on the average subjective rating of communicativeness was examined using a mixed-effect linear model (R. H. Baayen, 2008; R. Baayen, Davidson, & Bates, 2008) performed in R (R Core Team, 2023) using the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015). The context conditions were entered as fixed effects to predict com-

municativeness rating, while both the action and participants were entered as random effects. The model was fitted using Restricted Maximum Likelihood (REML). As shown in Fig.2, the model reveals that the Far context ($M = 5.29/7$) has significantly higher ratings than Do context ($M = 4.23/7$, $b = -1.06$, $CI_{95\%} : [-1.33, -0.79]$, $t(839.00) = -7.63$, $p < 0.001$), which has a significant higher rating than the Absent context ($M = 3.87/7$, $b = -0.36$, $CI_{95\%} : [-0.63, 0.09]$, $t(839.00) = -2.59$, $p = 0.010$). These findings demonstrate that the Far context was rated as significantly more communicative compared to the other two contexts, with the Absent context being perceived as the least communicative.

These findings contradict the predictions of the parallel theory but aligns with those of the envelope theory: zero contextual affordance does not lead to stronger communicative intention. Meanwhile, the highest ratings for the Far context particularly support the envelope theory that partial affordance supports the perceived communicativeness of actions. These findings support that instrumental intention is nested within communicative intention, suggesting a more nuanced relation between the two.

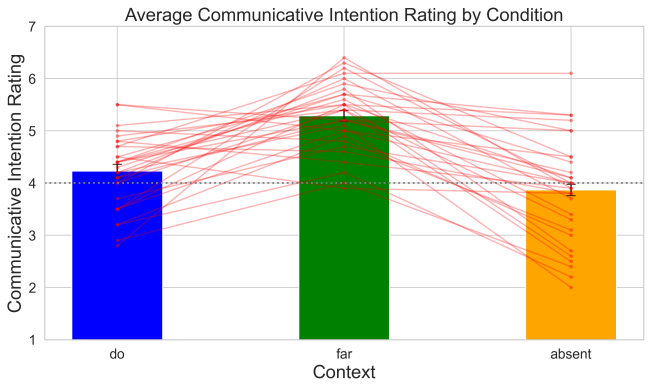

Figure 2: Each bar represents the average communicative intention ratings in a context condition. The vertical black lines represent standard error. Each red dot represents the average rating of an individual action for each context. The red lines connect the ratings of the same actions in different contexts.

Action-level analysis

Did these effects truly arise from context manipulation alone? We zoom into action-level analysis for communicativeness rating to find out whether certain actions were inherently more communicative due to their body motions. If so, we would expect some actions to consistently receive higher communicativeness ratings across all context conditions. Conversely, certain actions might only appear highly communicative in one context but not others, in which case we would not anticipate any correlation between the communicativeness ratings of actions and their context conditions.

We correlated the communicativeness rating of the same action of any two pairs of context conditions, making a total of three correlations. The analysis reveals no statistically significant correlations between the Do-Far context pair or Do-Absent context pair, but a weak positive correlation for the Far-Absent context pair (see Fig.3 for full statistical results).

In general, these results support the interpretation that the variance of communicativeness cannot be attributed to the body motion per se but largely relies on the interplay between the action and the context in which it is performed. In addition, certain actions, such as "closeJar,""flipPage,""openSnack," were rated as more communicative when the contextual affordance was not full, indicating potential avenues for further exploration. However, it's important to treat such a correlation with caution as it is marginal. The explicit source of the variance in the Do context will be further explored and discussed in later sections.

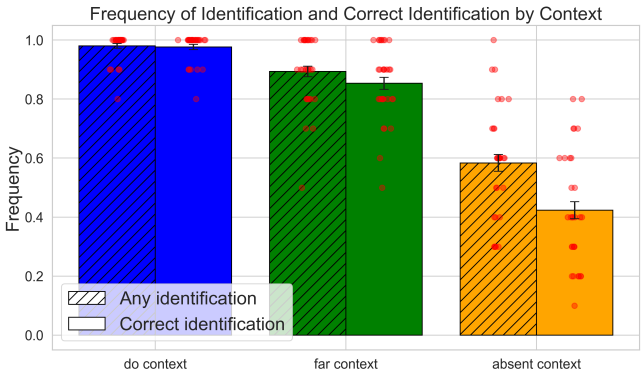

Figure 4: Hatched bars represent the frequency of any identification in free responses across contexts, while clear bars represent the frequency of correct identification in free responses across contexts. Each red dot represents an individual action, and vertical lines indicate standard errors.

Action interpretation

The envelope theory offers clear insight that understanding the instrumental components of the pantomime action in an ambiguous context supports the communicativeness of an action. We explore whether participants' interpretations of the instrumental intentions behind actions varied across different levels of contextual affordance as hypothesized. Free responses were pre-processed to include the "any" and "correct" identifications (see Data pre-processing sections for details). The two measurements reflect different mechanisms by which understanding the instrumental component enhances recognition of communicative intention: the correct-identification rate is more fine-grained, capturing an accurate recognition of the intended instrumental component, while the any-identification rate explores whether simply identifying any instrumental intention, even if not the correct one, is sufficient to enhance communicative intent recognition.

As shown in Fig.4, the frequencies of identifying any instrumental intention (hatched bars) or the correct instrumental intention (clear bars) were highest in the Do context ($M_{any} = 0.98$, $M_{correct} = 0.98$), followed by the Far condition ($M_{any} = 0.89$, $M_{correct} = 0.85$), and lowest in the Absent

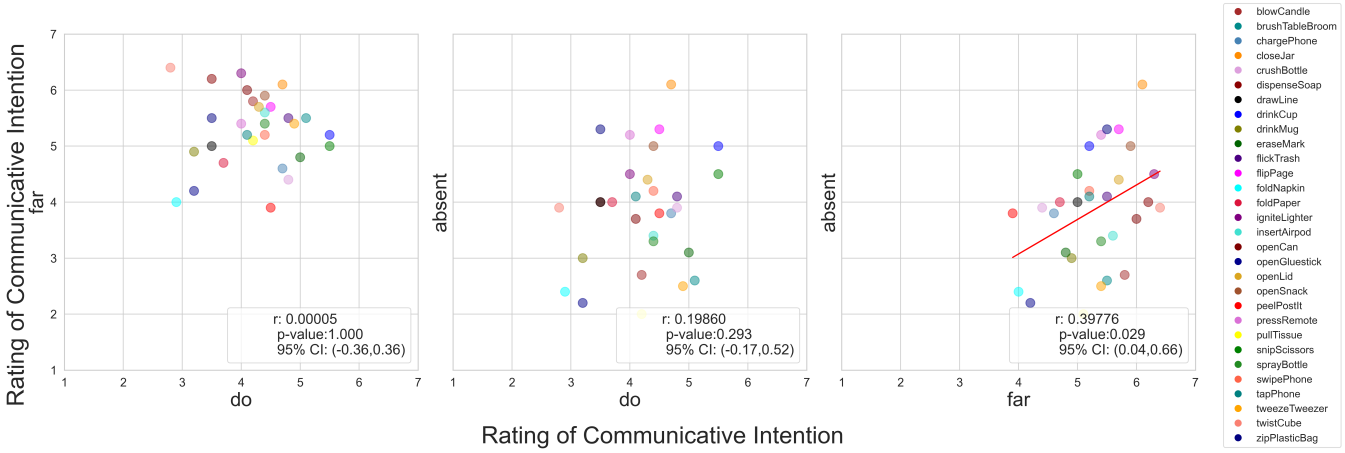

Figure 3: Correlation of average subjective communicativeness ratings for the same actions between Do and Far contexts (left), Do and Absent contexts (middle), and Far and Absent contexts (right). Each dot represents an action, with color codes displayed on the right. A linear regression line in red is shown for the significant correlation.

condition ($M_{any} = 0.58$, $M_{correct} = 0.42$). Additionally, a two-way ANOVA (3 contexts × 2 identification measurements) revealed a significant main effect of context ($F(2,174) = 98.33$, $p < 0.001$, $\eta_p^2 = 0.53$), a significant main effect of identification measurement ($F(1,174) = 5.44$, $p = 0.021$, $\eta_p^2 = 0.03$), but a non-significant interaction between context and measurement ($F(2,174) = 2.65$, $p = 0.073$, $\eta_p^2 = 0.03$).

These results indicate that when the context fully supported the action, participants had little difficulty identifying an instrumental intention, and their identifications largely aligned with the filming intention. Participants identified slightly less instrumental intention when the context affordance was partial, but their identification was mostly correct. However, when the object was missing and thereby the contextual affordance was absent, these identification abilities were significantly impaired, and participants' interpretations of the actions were often inconsistent with the intended filming purpose. These findings highlight the strong contextual influence on understanding instrumental intention from actions.

Communicativeness and instrumental intention

Building on previous analyses that show both communicativeness ratings and instrumental understanding of actions vary by context, we aim to further investigate whether the degree of perceived instrumental intention can explicitly explain the variance in communicativeness ratings, thereby providing a more direct test of our hypothesis.

First, we fitted two mixed-effect linear models predicting communicativeness ratings from both types of identification, with action and participants as random effects for both models. The models reveal a significant main effect for both "any identification" ($b = 1.51$, $CI_{95\%} = [1.20, 1.83]$, $t(874.83) = 9.63$, $p < 0.001$) and "correct identification" ($b = 1.37$, $CI_{95\%} = [1.10, 1.65]$, $t(863.69) = 9.83$, $p < 0.001$). The results indicate that successfully identifying the correct instrumental component of the action, or even recognizing any instrumental component, can predict higher communicative intention ratings.

Further, we focused on each context individually and analyzed how the identification rates of all actions (any or correct) correlate with their average communicativeness ratings within that context. The parallel theory states that failing to identify the world-directedness makes an action more communicative, therefore predicting no or a negative correlation. In contrast, the envelope theory argues that an action's instrumental component is largely intertwined with and provides support for its communicative components, therefore predicting a positive correlation.

For the any-identification rate, significant positive correlations were found in the Far and Absent contexts, while none were found in the Do context. The same pattern was observed for the correct-identification rate (see Fig.5 for full statistics). These results suggest that the more frequently viewers have at least some, or an accurate, understanding of the instrumental content of the actions, the more likely they are to rate them as highly communicative. The finding strongly supports that recognizing the instrumental goal of actions can facilitate the perception of communicative intention.

Discussion

This study introduces a novel perspective on defining communicative action and its connection to instrumental goals by focusing on pantomimes. Exploring how people discern communicativeness from pantomiming actions, we draw from two branches of literature, the parallel and the envelope theories, that offer different views on the relationship between communicative and instrumental aspects of actions. This study is particularly inspired by the envelope theory, which posits that communication is a joint inference of both instrumental intention and communicative intention, while the presence of one supports the other. Our research hypothesizes

that contextual affordances influence the perceived communicative intention of an action by affecting how its instrumental components are recognized, thereby impacting the support they offer for communicative interpretations. We investigate this by assessing the impact of contextual affordances on subjective ratings of communicative intention and analyzing open-ended explanations of the actions observed.

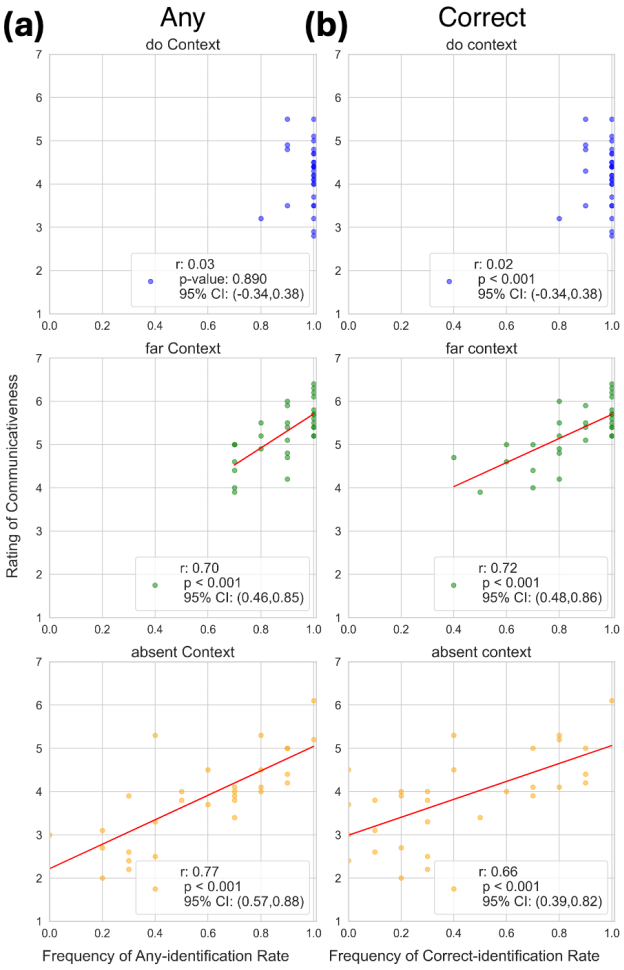

Figure 5: Subjective communicativeness ratings as a function of the frequency of (a) any identification and (b) correct identification responses for Do (top), Far (middle), and Absent (bottom) contexts. Each dot represents an action, with linear regression lines shown for significant correlations in red.

Rather than defining "communicative" for participants, we asked them to rely on their intuition. This approach ensures that the results reflect their naive perceptions of communication, allowing us to derive an understanding of what communicativeness means from their responses. The hypothesis was supported by results across context conditions: the communicativeness ratings were lower when the object was absent or directly manipulated, but highest when the object was present just enough to offer partial affordance for the action. This result illustrates a non-linear relationship between the amount of contextual affordance in a scene and how likely an action will be perceived as communicative. One explanation is that participants in the Far condition had a clearer understanding of the actions' instrumental intentions than in the Absent condition, but these intentions were not as overwhelmingly evident as in the Do condition. This nuanced balance, in turn, leads to the most robust interpretation of communication.

To further explore this idea, we closely examined individual actions within the context conditions by analyzing the correlation between recognizing the instrumental intention of an action and its communicativeness rating. As expected, we found no significant correlations in the Do context. This is largely due to the small variance in both types of intention identification rates, leading to a ceiling effect for all actions. However, we did observe positive correlations in the Far and Absent contexts. This result suggests actions that more overtly and accurately reveal their world-directedness are considered more communicative. Taken together, these results shed light on the idea that simultaneously recognizing the instrumental goals and realizing the suboptimality in achieving the goal work together to give rise to strong communicative action.

Additionally, we want to emphasize that these effects did not primarily result from differences in the specific body movements of actions across contexts. The rating of an action in the Do context did not predict its rating in the Far or Absent context. However, we did observe similar ratings for certain actions in the Far and Absent contexts, suggesting the degree these actions appeal to communication may depend on whether the objects are directly manipulated. Future research could analyze sentiments in the free responses to further explore this phenomenon. Overall, this result underscores the context-sensitivity of these cognitive processes and highlights that the interaction between action and context, rather than the action alone, is key to effective, intuitive gestural communication.

How can humans naturally recognize pantomiming actions as communication? We believe both the challenge for and solution to this question lie in the fact that they often resemble other goal-directed actions. On one hand, recognizing gestures intuitively relies on reasoning and planning abilities that are not based on conventions, a key component being the naïve utility calculus (Jara-Ettinger, Gweon, Schulz, & Tenenbaum, 2016), which allows us to observe a new environment and form intuitive expectations about how bodies should move to directly change the world efficiently. On the other hand, for these gestural actions to be communicative, they must deviate from the rationality principle (Royka et al., 2022), diverging from the expected actions that people would intuitively predict as the most rational path. Given these two key points, there is a need for an integrated theory of non-conventionalized gestural communication that specifies how the rationality principle can be carefully applied to serve multiple purposes simultaneously. This study represents a step toward achieving that goal, and there is much more to explore.

References

Arbib, M. A. (2012). *How the brain got language: The mirror system hypothesis* (Vol. 16). Oxford University Press.

Arbib, M. A. (2013). Complex imitation and the language-ready brain. *Language and Cognition*, *5*(2-3), 273–312.

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390-412. (Special Issue: Emerging Data Analysis) doi: 10.1016/j.jml.2007.12.005

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using r*. Cambridge university press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Becchio, C., Sartori, L., & Castiello, U. (2010). Toward you: The social side of actions. *Current Directions in Psychological Science*, *19*(3), 183–188.

Bratman, M. (1987). *Intention, plans, and practical reason*. Cambridge: Harvard University Press.

Calbris, G. (2011). *Elements of meaning in gesture*. John Benjamins Publishing Company.

Csibra, G., & Gergely, G. (2007). 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*(1), 60–78. doi: 10.1016/j.actpsy.2006.09.007

de Leeuw, J. R., Gilbert, R. A., & Luchterhandt, B. (2023). jspsych: Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, *8*(85), 5351. doi: 10.21105/joss.05351

Dennett, D. C. (1989). *The intentional stance*. MIT press.

Ferretti, F. (2021). The narrative origins of language. In *Oxford handbook of human symbolic evolution.* Oxford University Press. doi: 10.1093/oxfordhb/9780198813781.013.33

Ferretti, F., Adornetti, I., Chiera, A., Nicchiarelli, S., Magni, R., Valeri, G., & Marini, A. (2017). Mental time travel and language evolution: a narrative account of the origins of human communication. *Language Sciences*, *63*, 105–118.

Gärdenfors, P. (2017). Demonstration and pantomime in the evolution of teaching. *Frontiers in psychology*, *8*, 415.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. Harvard University Press.

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language*, *7*(1-2), 145-171. doi: 10.1111/j.1468-0017.1992.tb00202.x

Grice, H. P. (1957). Meaning. *The Philosophical Review*, *66*(3), 377–388. doi: 10.2307/2182440

Habermass, J. (1984). *The theory of communicative action-reason and the rationalization of society (Theorie des kommunikativen Handelns-Handlungsrationalität und gesellschaftliche Rationalisierung)*. Cambridge: Polity Press.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 29). Curran Associates, Inc.

Jakobson, R. (1965). Quest for the essence of language. *Diogenes*, *13*(51), 21–37.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*(8), 589–604. doi: 10.1016/j.tics.2016.05.011

Leslie, A. M. (1987). Pretense and representation: The origins of "theory of mind.". *Psychological Review*, *94*(4), 412-426.

McEllin, L., Knoblich, G., & Sebanz, N. (2018). Distinct kinematic markers of demonstration and joint action coordination? evidence from virtual xylophone playing. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 885.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (2012). *How language began: Gesture and speech in human evolution*. Cambridge University Press.

Poggi, I. (2007). Mind, hands, face and body. *A goal and belief view of multimodal communication. Weidler, Berlin*.

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Royka, A., Chen, A., Aboody, R., Huanca, T., & Jara-Ettinger, J. (2022). People infer communicative action through an expectation for efficient communication. *Nature Communications*, *13*(1), 4160.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language*. Syndics of the Cambridge University Press.

Sibierska, M. (2017). Storytelling without telling: The non-linguistic nature of narratives from evolutionary and narratological perspectives. *Language & Communication*, *54*, 47–55.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118. doi: 10.2307/1884852

Tomasello, M. (2010). *Origins of human communication*. MIT press.

Trujillo, J. P., Simanova, I., Bekkering, H., & Özyürek, A. (2018). Communicative intent modulates production and comprehension of actions and gestures: A kinect study. *Cognition*, *180*, 38–51.

Weber, M. (1925). *Wirtschaft und gesellschaft*. JCB Mohr (P. Siebeck).

Zlatev, J., Wacewicz, S., Zywiczynski, P., & van de Weijer, J. (2017). Multimodal-first or pantomime-first? communicating events through pantomime with and without vocalization. *Interaction Studies*, *18*(3), 465–488.

Żywiczyński, P., Sibierska, M., Wacewicz, S., van de Weijer, J., Ferretti, F., Adornetti, I., ... Deriu, V. (2021). Evolution of conventional communication. a cross-cultural study of pantomimic re-enactments of transitive events. *Language & Communication*, *80*, 191–203.

Żywiczyński, P., Wacewicz, S., & Sibierska, M. (2018). Defining pantomime for language evolution research. *Topoi*, *37*, 307–318.