Toward Scalable Measures of Quality of Interaction: Motor Interference

FRANK FÖRSTER, KERSTIN DAUTENHAHN, and CHRYSTOPHER L. NEHANIV,
University of Hertfordshire

Motor resonance, the activation of an observer's motor control system by another actor's movements, has been claimed to be an indicator for quality of interaction. Motor interference as one of the consequences of the presence of resonance can be detected by analyzing an actor's spatial movements. It has therefore been used as an indicator for the presence of motor resonance. Unfortunately, the experimental paradigm in which motor interference has been shown to be detectable is ecologically implausible both in terms of the types of movements employed and the number of repetitions required. In the presented experiment, we tested whether some of these experimental constraints can be relaxed or modified toward a more naturalistic behavior without losing the ability to detect the interference effect. In the literature, spatial variance has been analytically quantified in many different ways. This study found these analytical variations to be nonequivalent by implementing them. Back-and-forth transitive movements were tested for motor interference; the effect was found to be more robust than with left-right movements, although the direction of interference was opposite to that reported in the literature. We conclude that motor interference, when measured by spatial variation, lacks promise for embedding in naturalistic interaction scenarios because the effect sizes were small.

CCS Concepts: • **Human-centered computing → User studies**; **Collaborative and social computing design and evaluation methods**; *Interactive systems and tools*; *Empirical studies in interaction design*; *Systems and tools for interaction design*;

Additional Key Words and Phrases: Human-robot interaction, social robotics, interaction measures

ACM Reference format:
Frank Förster, Kerstin Dautenhahn, and Chrystopher L. Nehaniv. 2019. Toward Scalable Measures of Quality of Interaction: Motor Interference. *Trans. Hum.-Robot Interact.* 9, 2, Article 8 (December 2019), 25 pages.
https://doi.org/10.1145/3344277

1 INTRODUCTION

Detecting in an automatic manner whether a particular interaction between human and machine "works" is an unsolved problem in human-machine interaction. This is unsurprising given that human-machine interactions may be classified along many "degrees of freedom," ranging from the functional to the social, the embodied to the disembodied, and the collaborative to the

This material was based on work supported by the Air Force Office of Scientific Research, Air Force Material Command, USAF under award no. FA9550-15-1-0063.
Authors' address: F. Förster, K. Dautenhahn, and C. L. Nehaniv, University of Hertfordshire, College Lane, Hatfield, AL10 9AB, UK; emails: {f.foerster, k.dautenhahn, C.L.Nehaniv}@herts.ac.uk.
Authors current address: K. Dautenhahn and C. L. Nehaniv, University of Waterloo, 200 University Avenue W., Waterloo, N2L 3G1, Canada; emails: {chrystopher.nehaniv, kerstin.dautenhahn}@uwaterloo.ca.

2573-9522/2019/12-ART8
https://doi.org/10.1145/3344277

competitive, to name a few. Yet even when narrowing down this range to, say, the subset of embodied, collaborative interactions, the same statement holds true: no computational technique exists by which the artificial agent could perceive whether the interaction works from the viewpoint of the human or whether interactional breakdown is likely to occur, meaning that a frustrated human "user" gives up on the machine on that particular occasion.

1.1 Motor Resonance, Motor Interference, and Quality of Interaction

In human-robot interaction (HRI), motor resonance has been proposed as a potential candidate for computationally assessing what might be termed 'quality of interaction' [12]. Chaminade and Cheng [11] assert that "the measure of resonance indicates the extent to which an artificial agent is considered as a social inter-actor," which appears to be a somewhat more static qualification than the on-the-fly assessment of an interaction in progress that we are alluding to here. More importantly the word 'consider' seems to imply a conscious, rather than automatic, evaluation of the interaction partner. The motor resonance phenomenon has been more generally been referred to as "a plausible foundation for higher-order social cognition" [10], and motor interference may be conceptualized as its metric [29]. The preceding statements attribute to motor resonance its suitability as a general measure for the potential of an artificial agent to be conceived of as a social entity. And although motor interference has been successfully used in the past for investigating the factors involved in triggering motor resonance in principle, the question remains whether it can be used as a (soft) real-time measure for the quality of an ongoing interaction.

In the standard paradigm for assessing the presence of motor interference introduced by Kilner et al. [22], the effect is measured in a constrained interaction scenario. Participants execute intransitive, or target-less, vertical or horizontal waving motions while observing similar congruent or incongruent motions of a model. The models employed in previous studies range from moving dots on a screen [44] to full-fledged humanoid robots [12, 43]. In many, although not all, studies performed within this paradigm, participants are trained before the trials proper to ensure their movements adhere to certain constraints. Exemplary contraints are a set movement length, or a set frequency of movement apart from the defined movement direction [21, 44]. Figure 1 visualizes the effect of motor interference.

Most motor interference studies were designed with the purpose of identifying the factors involved in triggering the interference effect. These can be divided into top-down and bottom-up factors. Typical bottom-up factors pertain to biological properties of the model, such as joint configurations [24], movement dynamics [12, 21], or the overall visual appearance [32]. Top-down factors are psychological, with examples being beliefs about the models' agency or intentionality [26, 43, 44], or the mental states resulting from calling certain properties of the model to attention [28].

1.2 Unresolved Issues Surrounding Motor Interference Measures

Qualifications of motor resonance such as the ones cited in the previous section are too vague if our ultimate aim is to find a quantiative measurement tool that, at least partially, measures the quality of interaction with a sufficiently small temporal granularity. The cited qualifications do not spell out the precise role that motor interference could have in assessing an artificial agent. The fact that motor interference in particular has only been measured in an ecologically rather implausible scenario raises a number of questions with regard to its potential utility as a practical "benchmark measure."

On a more fundamental level, it is unclear how precisely motor interference measurements relate to alternative indices of motor resonance, such as measures of motor synchronization [36, 46]. By extension, it is unclear how motor interference measures relate to outcome measures that are more closely linked to motor synchrony, such as rapport [2] or the fluidity of interaction [8]. In

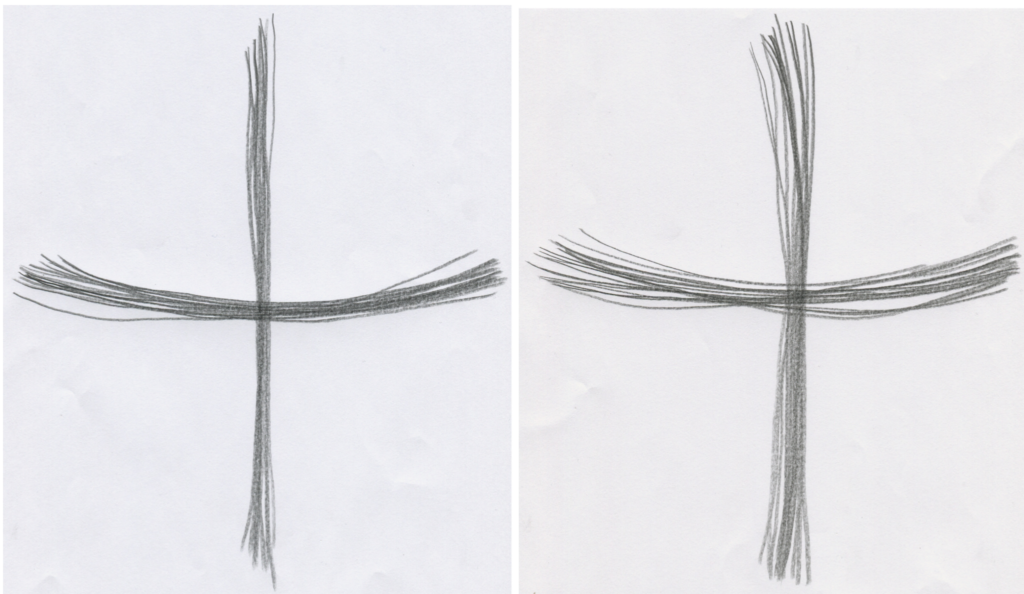

Fig. 1. Reproductions of exemplary plots shown in Kilner et al. [22] depicting the impact of motor interference with a human model. Left: A participant's movements in congruent trials. Right: Movements in incongruent trials—that is, under the influence of motor interference.

contrast to the interactionally simultaneous setup in which motor interference is typically measured, interactional fluidity and motor synchronization can be measured and discussed based on turn-based interaction formats that resemble more closely the way humans naturally interact with one another [13, 45]. The following thought experiment may help to illustrate the issue.

Let us assume we were to assess artificial agents in terms of their capacity to provoke motor resonance in human observers in the established motor interference paradigm. In other words, assume we used intransitive movements and a large number of repetitions to benchmark artificial agents in terms of their capacity to be considered social interactors. Let us refer to this measure as the "social capacity benchmark." Now imagine we had two robots, R1 and R2, with different benchmark scores and R1 had a higher score than R2. Now let us further assume we were to equip both robots with turn-taking controllers such as the one developed by Chao and Thomaz [13], and we deployed them in some open-ended interaction involving turn taking.

Given that the two robots have two different benchmarks based on motor interference but use the same controller governing their turn-based behavior, what outcome with regard to the ensuing interaction would we expect? Would we expect the human interactor to interact longer with R1 than with R2? Would we expect the human to be more tolerant with respect to interactional missteps on part of R1? Would we expect the human interactor to feel more engaged when interacting with R1 as compared to R2?

If the answer to these questions is negative, then it is questionable whether motor interference has much utility as a benchmark in terms of informing concrete robot designs or HRI designs. If we answer these questions positively, however, then we need to clarify how this benchmark relates to measures such as motor synchronization. In other words, we need to clarify the precise relationship between motor interference, of which we only obtain a snapshot measure in the established paradigm, and measures of motor synchrony, which are usually lifted from turn-based interaction formats. Our working hypothesis is that motor interference effects are, in principle, dynamic in the sense that their effect size fluctuates within the time window of an interaction akin to what has been observed with measures of interpersonal coordination. We also assume that the "granularity" of fluctuation is comparable to that of motor synchronization measures. In other words, we assume that the issue is one of measurement rather than a difference in the way in which these different indices—motor interference and motor sychronization—are modulated by motor resonance.

1.3 Motivation and Research Questions

The published research on motor resonance in human-machine dyads focused by and large on the identification and disentanglement of factors attributed to the artificial agent that are either necessary for or contributing to the evocation of motor resonance in the human interactor. Although the current work still attempts to contribute to this line of research, a second and equally important research target is to assess whether motor interference measures can be used in less artificial setups.

There are at least three axes along which motor interference would need to be "scaled up" to render it a relevant candidate measure for assessing the quality of an unfolding interaction.

First, it would need to scale in terms of complexity: the measurement techniques need to be applicable to motor actions more complex than linear intransitive movements. We address this issue by having participants use interactionally more plausible transitive—that is, goal-directed *grasp and translate*—movements: grabbing an object and placing it somewhere else.

Second, it would need to scale in terms of the speed with which it can be detected. Currently, motor interference, when measured via behavioral artifacts such as variation of a movement trajectory, is only detectable after the interaction has already ended. Although not exactly in real time, motor resonance would need to be detectable in relatively quick succession to the relevant interaction segment if it ought to inform the behavior of the robot. This issue is only tangentially addressed in the present experiment and remains an open problem.

Third, motor interference is currently only measured in simultaneous interaction formats, where agents are acting at the same time. We know from the conversation analytical literature that those human interactions that involve conversations typically unfold in turns: the actors take turns in (inter-)acting rather than acting simultaneously [41]. We therefore attempt to assess whether motor interference scales from sequential to turn-based interaction formats.

Based on these considerations, the research questions underlying our experimental design are the following:

- (Q1) Given the proposed *grasp and translate* paradigm,[1] do the established techniques for detecting motor interference using spatial variation scale to our more complex motor actions?
- (Q2) If these techniques do scale, which are the relevant factors that impact upon potentially occurring motor interference?
- (Q3) Are those same techniques applicable to turn-based interaction formats? In other words, does the interference effect last long enough to be detectable in the next turn?
- (Q4) Does the presence of motor interference, widely considered a sub-conscious phenomenon, impact on humans' conscious, subjective evaluation of the interaction?

[1]The paradigm is new with respect to the measurement of motor interference. Bisio et al. [4] have used it before in the context of detecting motor contagion.

The experimental factors used in the present study are *congruency*, *interaction mode*, *movement direction*, and *prime*.

Congruency is the common and defining factor in all motor interference studies where the effect is measured in terms of spatial variation of the human arm movement. Motor interference is said to occur if a certain spatial variation of human arm movement is significantly larger during the execution of *incongruent* movements between model and human as compared to *congruent* movements. The spatial variation in question is the one orthogonal to the main axis of movement.

The use of *interaction mode* as a factor formalizes our question as to whether motor interference "survives" the transition to consecutively executed movements rather than simultaneous ones.

Priming is a top-down factor that has been investigated in past studies. However, the established literature is still divided as to whether it does modulate the motor interference effect. We therefore include it as an explicit factor in our study. On a practical level, *priming* is interesting because it is comparatively "cheap" in the sense that, rather than requiring a potentially costly re-design of the (interaction) system, it consists of making a robot-related statement concerning its mental state or cognitively relevant behavior toward the human user or interactor. If efficacious as modulator of motor interference, it would be an easy factor to apply and test before re-designing a given robot behavior.

Movement direction is a somewhat incidental factor in that it only arises through the presence of a horizontal surface in our setup. Although theoretically it is the least interesting factor, it has great practical relevance. Horizontal surfaces such as tables, shelves, or book cases are ubiquitous in the human environment, and transitive *forth-back* movements are no less likely and plausible than *left-right* movements. As we are investigating the scalability of motor interference measures toward more realistic interactions, a relevant question is whether it can be detected in *forth-back* movements as well.

2 MATERIALS AND METHODS

2.1 Study Design

The presented study was designed as a factorial experiment with the four factors *priming*, *interaction mode*, *congruency*, and *movement direction*. *Priming* was the only between-group factor with two levels: *primed* and *not-primed*. *Primed* participants were told that 'Deechee', the employed humanoid robot 'iCub' [30], would be watching them during the experiment. *Not-primed* participants were not told anything in this respect. Each of the other three within-group factors consisted of two levels such that each participant completed eight experimental runs.[2] This means that each factor combination was realized precisely once for each participant.

In terms of the *interaction mode* (*iMode*), a run could be either *simultaneous* or *consecutive*. In *simultaneous* runs, the participant and robot would perform their respective movements simultaneously. In *consecutive* runs, the participant would first watch the robot perform its movements and start his or her own performance immediately afterward.

In each run, the movements of the respective participant were either *congruent* or *incongruent* to the robot's movements. If the robot moves its toy forth and back, and the participant moves it from left to right (and back), they are moving in an incongruent manner. If they both move their respective toy left and right, they are moving congruently. Because each of the two actors can move along two main axes (forth-back and left-right), there are four different combinations of movement for the dyad, two of which are congruent and two of which are incongruent. We might

[2] In the following, we will refer to the experimental unit corresponding to the actions of a single participant in one particular condition as an (experimental) *run*.

refer to these as the two *instantiations* of congruency or incongruency. Some authors, such as Shen et al. [43], only use one instantiation of incongruency.

Movement direction (*mDir*). In previous research on motor interference, the participants and/or model performed intransitive horizontal (*left-right-left*) or vertical (*up-down-up*) repetitive movements. Our setup involving transitive actions on a table surface affords us naturally two horizontal but no vertical movements. Whereas the *left-right-left* grasp-and-translate actions on the table surface share a high degree of similarity to the established horizontal waving motions, there is no equivalent in the literature to the *forth-back* movements. We decided to tentatively include *forth-back* movements nevertheless, as this movement direction is plausible and natural in the given setup. Moreover, we were curious as to whether motor interference can also be observed in this movement direction. Analytically, however, we will separate the data in part of our analysis along this factor, as we do not consider *movement direction* to be a genuine experimental factor but rather an implementation detail originating in the specifics of the physical layout.

In addition to participants' movement data, we were also interested in their subjective assessment of the interaction. This assessment was realized via the design of a new questionnaire (cf. Section 2.5) that was completed by participants four times, once for each instantiation of the combination of *congruency* x *iMode*. The order of conditions was counterbalanced and randomly assigned to participants with one constraint: each of the four possible instantiations of *congruency* x *iMode* were blocked—that is, the two instantiations of congruency were executed sequentially. After each block, participants were asked to complete a questionnaire before they could continue with the next run.

Participants' arm movements were recorded with the help of the motion tracker Polhemus Liberty, where a tracker is fixed to a participant's wrist and sends signals to a receiver located on the table between the participant and robot (cf. Figure 4). Details to the dependent variables and the data analysis of the motion data are provided in Section 2.6.

2.2 Recruiting and Distribution of Participants

We recruited 24 participants, 22 of whom were right-handed and 2 of whom were ambidextrous. Twelve of the 24 participants were female, the other 12 were male, and the average age was 31.2 years (SD = 12.4). Participants were recruited by the use of flyers and posters on campus, in the area surrounding the university and from attendees of a series of scientific talks called 'Cafe Scientifique'.

2.3 Instructions to Participants

Prior to the experiment, participants were handed an information sheet and consent form. In the information sheet, we explained that they would interact with the humanoid robot 'Deechee'. We explained the overall purpose and general structure of the experiment, namely that there would be eight runs and a questionnaire after every second run. Within these explanations we did not make any mention of motor resonance, as we did not want to heighten participants' attention to the specifics of their own movements. The overall aim was to have participants move as naturally as possible. Apart from the priming of those participants in the *primed* group and some explanation about the humanoids' grasping capabilities, we did not make any further statements with respect to *Deechee*.

Participants were told that their role in the experiment was to move the toy located in front of them either in a forth-back or left-right-left direction. Before the start of each run, participants were told the expected *movement direction* of their movements, as well as whether the run was a *simultaneous* or a *consecutive* one. For *simultaneous* runs, they were instructed to start their movement as soon as they saw Deechee starting to move, and that they could stop once Deechee stopped. For *consecutive* runs, we told participants that they should watch Deechee during its turn, and that they should start their execution as soon as Deechee finished its turn. If participants did not start performing their turn immediately after Deechee finished its turn, the experimenter told participants that they should start.

2.4 Robot Behavior and Spatial Setup

When participants entered the room, the robot was in its start position, meaning its right arm was located over the toy and its gaze fixating the toy. No eye contact with participants was established (Figure 2, upper left). Once participants were given the explanations outlined in the previous section and signed the consent form, they had the motion sensor attached to their wrists. Then they were seated opposite the iCub (cf. Figures 4 and 5), and the experiment would start. The robot's behavior was semi-autonomous. That means the start of the robot's behavior was triggered manually by the experimenter for each run. Thereafter, the robot acted autonomously until it received a stop signal, or until the required number of repetitions had been executed.

Upon receiving the initial start command, 'Deechee' would grasp the toy and start moving it *forth-back-forth* or *left-right-left* repeatedly depending on the predetermined *movement direction* of that particular run. An example of the spatial trajectory and velocity profile of the robot's hand movement are shown in Figure 3. Just before each respective movement segment, the robot moved its head and eyes such that it looked at the next target position. No explicit time delay was introduced into the high-level control code, yet the mere sequencing leads to a relatively naturalistic behavior in that the gaze is shifted toward the target a few hundred milliseconds before the onset of the arm movement.

In *consecutive* runs, Deechee would execute 15 movement cycles and subsequently stop at the start position with its gaze fixated on the toy. If the run was a *simultaneous* one, Deechee would execute the respective movements until it received a stop signal from the experimenter. It would then end its movement with the toy in the start position and its gaze fixated on it. If Deechee accidentally dropped the toy, the experimenter sent it a signal to that effect and Deechee would go into its start position. In this case, its hand was slightly higher, and its fingers opened such that a *grasp* action could be executed in the next move. The experimenter would then place the toy at the start position on the table and send Deechee another signal triggering it to grasp the toy and restart the movement execution for this particular run. Additional technical details are provided in the supporting materials (Section A.1.2).

2.5 Questionnaire Design

For the purpose of assessing participants' conscious evaluation of the interaction, we created a new questionnaire that is an amalgamation of two existing questionnaires: the *Temple Presence Inventory* (*TPI*) [27] and the *Networked Minds Social Presence Inventory* (*NMI*) [3]. To our knowledge, there is no existing questionnaire that would assess the perceived quality of an experienced interaction. Parts of the aforementioned two questionnaires appeared to be the closest match in this respect. From both questionnaires, we included all factors and items that applied to our experimental setup, such as factors and items that were not obviously non-sensical. Items that assumed the robot and human were not physically co-located in the same space, for example, were not applicable.

The *TPI* was designed to assess different dimensions of presence in tele-presence scenarios. This means that some factors and items were not applicable to our scenario due to the robot and human being physically co-located in the same room. The *TPI* assesses eight different dimensions of presence with 48 items in total. We incorporated 36 items from four of these eight dimensions into our questionnaire.

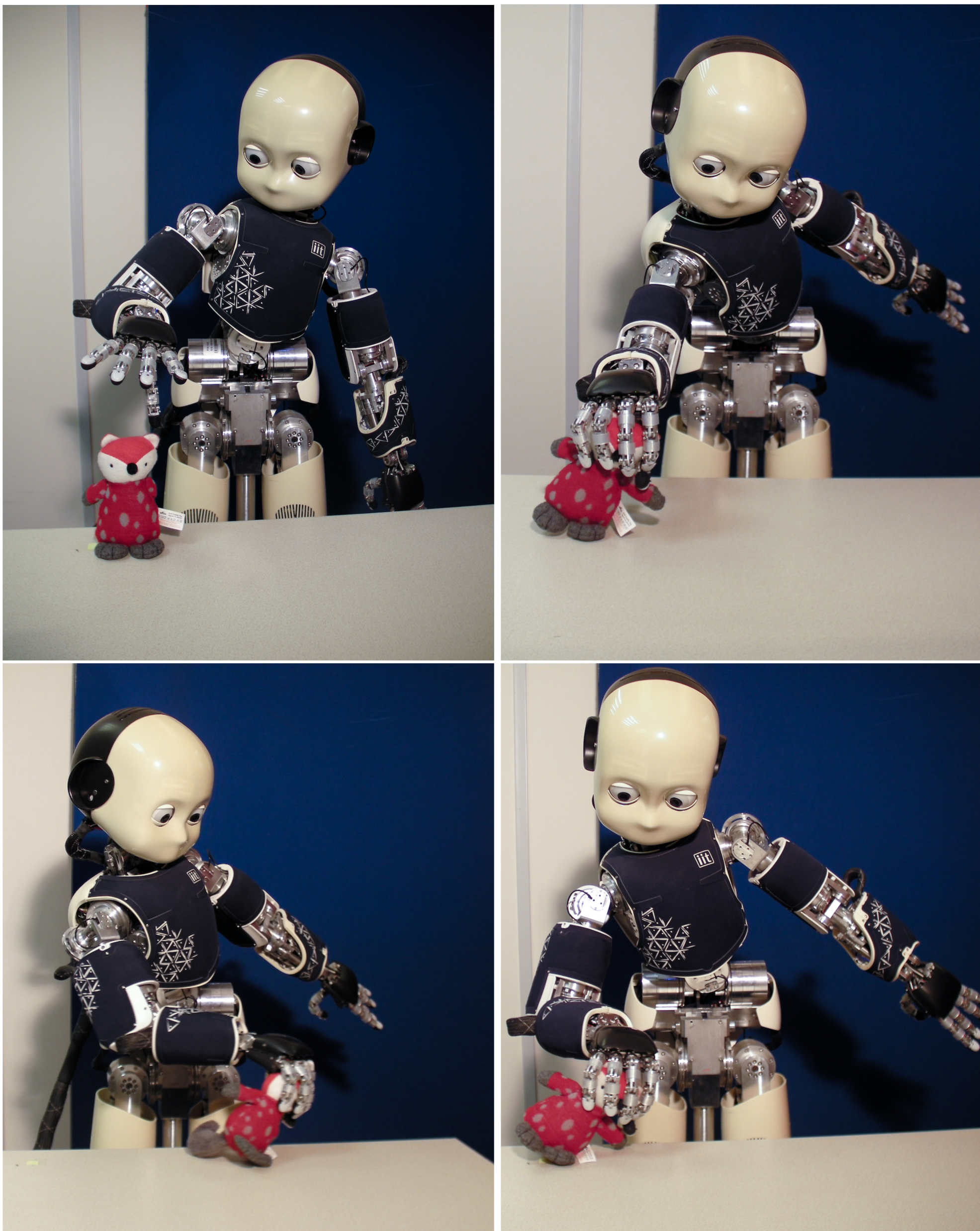

Fig. 2. iCub humanoid robot in various phases of movement. Top left: Start position prior to the start of the first run. Top right: After a forward move (half of the *forward-backward* movement cycle). Bottom left: After a left move (half of the *left-right* movement cycle). Bottom right: After having completed one *left-right* cycle.

TPI-Social Richness (*TPI-R*). We used all items from this dimension that consists of opposed pairs such as *Remote-Immediate* or *Unsociable-Sociable*. *Social Richness* assesses "the extent to which a medium (or an artificial agent) is perceived as sociable, warm, sensitive, personal or intimate" ([27], clause in parentheses added by us).

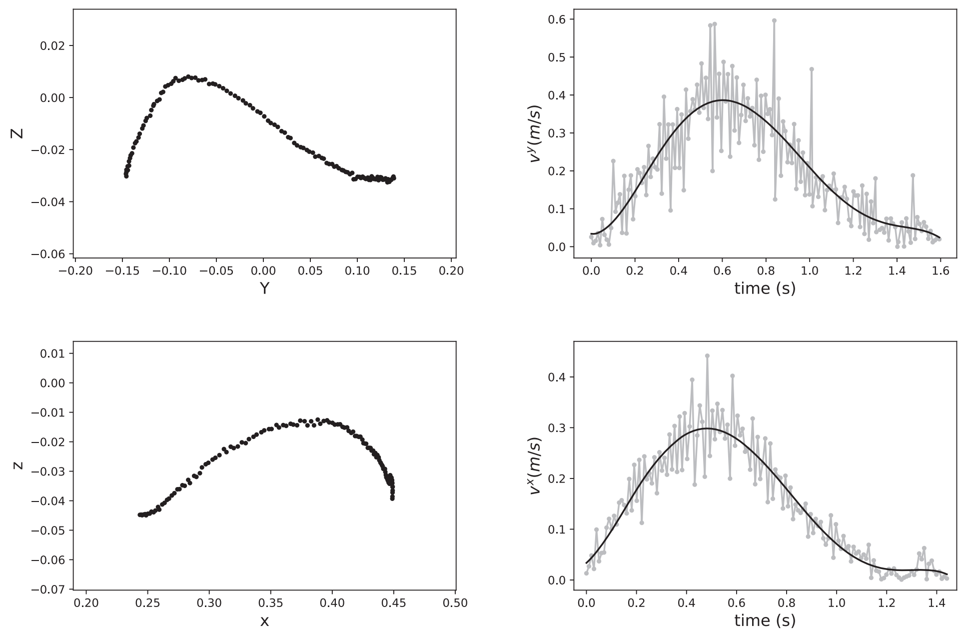

Fig. 3. Spatial and velocity profiles of single arm movements of the iCub robot (right arm). Top and bottom left: Robot's hand trajectory in the sagittal and frontal plane of a single forth-back and left-right movement, respectively. Points correspond to the center of the robot's palm as reported by the robot's Cartesian interface. Top and bottom right: Velocity profiles of the same left-right and forth-back movements, respectively. The velocity profile was calculated based on the position information and the associated timestamps (gray curve); the black line depicts a regression curve.

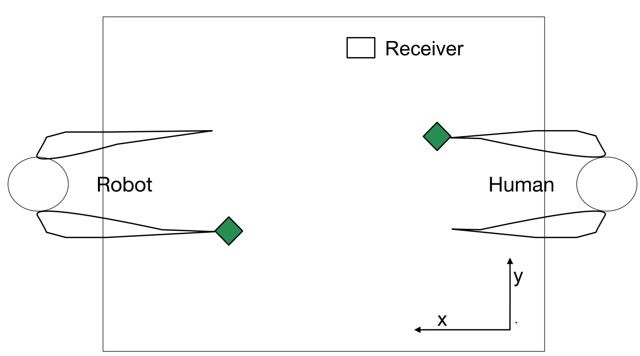

Fig. 4. Spatial setup of the experiment and coordinate system. The participant and robot are located on opposite ends of a small office table. They move the provided toy both with their right hand and arm either simultaneously or consecutively, either in a congruent or incongruent manner.

TPI-Social Presence–Active Interpersonal (*TPI-SPA*). All items from this construct were incorporated, including items such as "How often did you smile in response to Deechee?" This construct assesses whether a medium or a robot is conceived of as a social actor or the degree of perceived anthropomorphism.

TPI-Social Presence–Parasocial Interaction (*TPI-SPP*). This dimension of presence assesses to what degree users of a medium ignore the mediated nature of the interaction, or whether the media

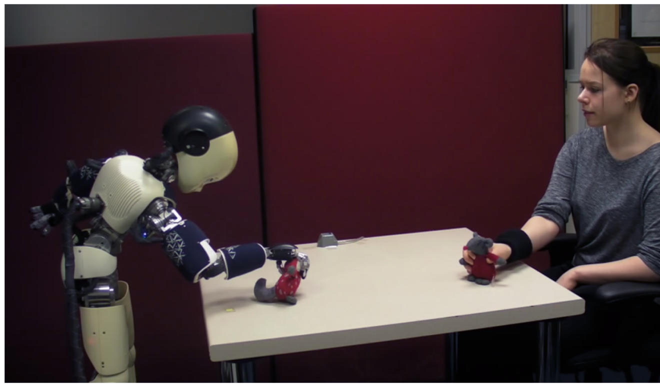

Fig. 5. Still image of one instantiation of the simultaneous, incongruent condition. The participant moves the toy from left-to-right and back, whereas the robot moves its toy forth-and-back.

personality is perceived as a genuine social actor. Despite the interaction not being mediated in our setup, three of seven items still appeared suitable, with one example being "How often did you want to or did you make eye contact with Deechee?"

TPI-Engagement (*TPI-Eng*). Three of six items from this construct were transferred into our questionnaire. *TPI-Eng* was originally designed to assess the degree of engagement with or mental immersion into a virtual reality. We chose those three items that applied to our non-virtual scenario—items such as "How relaxing or exciting was the experience?"

Similar to the *TPI*, the *NMI* was designed to assess a user's sense of social presence when the interaction is mediated. The *NMI* explicitly mentions its applicability in "quasi-social relationships with new forms of artificially intelligent beings" [3]. Again, only a subset of items applied to our experimental setup.

NMI-Co-Presence (*NMI-CoP*). We incorporated all eight items of this scale into our questionnaire. *Co-Presence* assesses the participant's perception of whether she felt that she was together in the same space with the robot. "Deechee hardly noticed me" is an exemplary item from this scale.

NMI-Perceived Attentional Engagement (*NMI-PAE*). We only used this sub-scale from the construct *perceived psychological engagement*. The sub-scale consists of six items, all of which were incorporated into our questionnaire. Example items would be "I paid close attention to Deechee" or "Deechee tended to ignore me."

NMI-Perceived Behavioral Interdependence (*NMI-PBI*). This scale consists of six items, and we incorporated all of them. *Behavioral interdependence* refers to the coupling of one's own actions to those of another. Such couplings happen in collaborative forms of interaction but also in competitive ones. Although participants' movements were objectively not coupled to the ones of the robot in our setup, this does not preclude the possibility that participants might perceive them as being coupled, especially in the presence of motor resonance. Examples of items in this scale are "My behavior was often in direct response to Deechee's behavior" or "What I did often affected what Deechee did."

2.6 Overview of Analysis of Movement Data

The motion data recorded from participants' arm movements needs to undergo several steps of pre-processing before any statistical analysis can be conducted. As we will see in Section 3, the particular choices made in any of these steps can have a considerable impact on the results. For this reason, we provide a more detailed account of the various pre-processing options than is usual in the established literature on motor interference. Instead of just presenting one particular combination of choices, we executed a multitude of analyses to determine how sensitive the detection of motor interference is to these data processing choices.

The data processing pipeline can be decomposed into the following operations and is independent of the particular type of motion tracker employed:

(1) Low-pass filtering of motion data
(2) Decomposition of single-run motion records into motion segments corresponding to single left-right or forth-back movements
(3) Pruning of boundary data at both ends of each motion segment
(4) Removal of entire motion segments at either end of the recording for each run (optional)
(5) Calculation of baseline proxy data for spatial variation—either variance/standard deviation in the orthogonal plane or curvature
(6) Outlier detection and removal from baseline statistics (optional)
(7) Calculation of baseline descriptive statistics based on the variation proxies
(8) Normalization of the descriptive statistics (optional)
(9) Inferential statistics operating on normalized descriptive statistics data.

The order of steps 3 and 4 may be inverted, and step 4 may not be necessary. Step 6 (outlier removal) is often not discussed in the literature, so it may or may not be optional. Step 8 is typically only performed if curvature is chosen as the proxy.

2.6.1 Pre-Processing of Motion Data. With 'pre-processing', we refer here to steps 1 through 6 of the pipeline sketched earlier. Given the motion recordings of single runs, these recordings are first low-pass filtered. We followed Oztop et al. [34] and filtered the data using a 25-Hz fifth-order Butterworth filter.[3] Once filtered, the run-level records were split into segments corresponding to single motion segments. In the present study, we employed a semi-automatic segmentation process.

In step 3, individual motion segments typically are pruned. This is necessary because participants appeared to momentarily rest their hand on the extreme end of movement segments exhibiting what Oztop et al. [34] referred to as "small drifts in the hand location." Again, a semiautomatic process was employed where an automatic boundary detection is followed by a phase of manual checks and corrections. As this type of pruning may remove important parts of the data, we made the decision to analyze both pruned and unpruned sets of motion segments. *Gaps* data in Section 3 refers to the set of pruned motion segments, whereas *No Gaps* refers to the set of unpruned segments.

In step 4, we removed the very first and the very last motion segments of a run such that only 20 motion segments remained, which corresponds to 10 complete *left-right-left* or *forth-back-forth* movements.

Hereafter, following Gowen et al. [19], we will refer to the plane that is orthogonal to the one that corresponds to participants' main direction of arm movements as the *error plane*. Two kinds of proxies quantifying the variation in the error plane are common in motor interference research (step 5):

- Variance or standard deviation of the movement with respect to the error plane (cf. [19, 21, 22, 43])
- A particular kind of curvature (cf. [12, 18, 34]).

[3]In our case, low-pass filtering did not make much of a difference in terms of the ensuing statistical results. We applied it nonetheless to stay as close as possible to the data processing pipelines documented in the literature.

Table 1. 2^4 Tests on the Full Dataset

Conf. #	*GAPS*	*Participants rm*	*Outliers rm*	*Norm. Type*	*Test Type*	*Significant Effects and Interactions* (**bold**) *& Test Statistics*		
						movement-direction.	**congruency × movement-dir.**	
1	no	yes	yes	N2	para	($F = 5.74$, $p = .026$, $\eta^2_G = .030$)	($F = 9.84$, $p = .005$, $\eta^2_G = .037$)	*
2	no	no	yes	N2	para	($F = 5.53$, $p = .028$, $\eta^2_G = .027$)	($F = 8.07$, $p = .010$, $\eta^2_G = .030$)	*
3	no	no	no	N2	para	($F = 4.56$, $p = .044$, $\eta^2_G = .023$)	($F = 7.75$, $p = .011$, $\eta^2_G = .030$)	*
4	yes	yes	yes	none	para	($F = 31.86$, $p < .001$, $\eta^2_G = .400$)		*
5	yes	no	yes	none	para	($F = 35.37$, $p < .001$, $\eta^2_G = .390$)		*
6	yes	no	no	none	para	($F = 38.93$, $p < .001$, $\eta^2_G = .390$)		*
7	no	yes	yes	none	para	($F = 57.15$, $p < .001$, $\eta^2_G = .521$)	($F = 10.65$, $p = .004$, $\eta^2_G = .014$)	
8	no	no	yes	none	para	($F = 62.94$, $p < .001$, $\eta^2_G = .510$)	($F = 8.92$, $p = .007$, $\eta^2_G = .011$)	
9	no	no	no	none	para	($F = 65.64$, $p < .001$, $\eta^2_G = .504$)	($F = 8.59$, $p = .008$, $\eta^2_G = .013$)	

Shapiro-Wilk: * $p < 0.001$.

Note: Listed are tests and data configurations yielding significant results for both curvature and variance proxies and the results of the so-characterized test. *Curvature proxy data* per motion segment is aggregated using the mean. Only significant ($p < .05$) results are listed. *Conf. #*: Number of test configuration. *GAPS* (yes/no): Basis for analysis is the *Gaps* dataset and *No Gaps* otherwise. *Participants Removed* (yes/no): Outliers on the level of entire participants were removed if "Yes." *Outliers Removed* (yes/no): Outliers on the level of single runs were removed if "yes." *Norm. Type*: *N2* or *none*. See the main text for an explanation of normalization *N2*. *Test Type* specifies whether the stated results are based on a parametric ANOVA or a non-parametric test. Asterisks (*) at the very end of a row indicate that the distribution of residuals was found to be non-normal. For these cases, the p-values of the Shapiro-Wilk test are given in the last row. Total number of tests performed: 12 for curvature proxies and 6 for variance proxies. For space reasons, the degrees of freedom for each F-value are not printed in the table but would be $F(1,22)$ for all entries.

Due to the relaxation of behavioral constraints in our experimental setup as compared to the established paradigm, we could not assume that these established methods of measuring motor interference would automatically transfer. For this reason, we calculated both curvature and variance and subsequently performed the inferential analysis on both datasets separately. In terms of curvature, there is an important difference between our approach and the one used by Chaminade et al. [12] and Oztop et al. [34], which is described in Section A.1.5 of the supporting material.

2.6.2 Outlier Detection and Removal (Step 6). Prior to conducting inferential statistical tests, it is common to remove outliers. Yet, in the motor resonance literature, the issue of outliers is frequently not discussed. Thus, it is unclear whether the authors of the respective publications removed outliers at all, and if they did, on what level. We analyzed the data under three different levels of outlier removal:

- Removal of the entire participant's data plus removal of single data points (*P's removed*)
- Removal of single data points only—curvature or variance measurements (*O's removed*)
- No removal of outliers.

Outliers were detected and removed separately for variance and curvature data, and details are provided in the supporting material. As a result, participants P08 and P16 qualified as outliers within the curvature and participant P24 in the variance dataset. Outliers were excluded in the analysis where the attribute *Participants removed* (*P's removed*) holds (Table 1). In the variant where only single outliers on the level of the baseline proxies were removed, the same outlier detection criterion in terms of distance from the quartiles as explained in the supporting material was applied. However, the outlier detection was performed for each participant and movement direction separately but on otherwise pooled data.

2.6.3 Baseline Descriptive Statistics (Step 7). Descriptive statistics of the proxies typically form the basis for the inferential statistical procedures such as *t*-tests or analyses of variance (ANOVAs). In other words, the inferential statistics do not operate on the proxies—variance or curvature—directly but on derivatives or aggregates of these. The outcome of aggregation is one statistic per participant and run, which then may or may not be normalized prior to forming the basis for the inferential analysis. We will refer to these two practices of calculating one data point per participant and run as *mean* and *variance aggregation*, respectively. For our analyses, we employed both established methods to be able to compare the outcomes and establish whether all methods lead to identical outcomes, or whether the outcome is dependent on the chosen analytical route. This means that we used only mean aggregation for variance proxy data, and both mean and variance aggregation for curvature proxies.

2.6.4 Normalization of Motion Segments (Step 8). Normalization only applies if curvature is the proxy of choice and is applied to the statistics from step 7. Oztop et al. [34] appear to normalize the variance of the curvature per trial by the mean of the curvature in that particular trial. We used both the method of Oztop et al. [34] and a new normalization method described in the supporting material. The two normalization methods will be referred to as *N1* and *N2*, respectively.

2.6.5 Inferential Statistical Analysis (Step 9). The statistical analyses were performed in six separate blocks that differ from each other by the particular combination of the size of the linear model ($2{\times}2 \times 2{\times}2$ (or 2^4) or $2{\times}2{\times}2$ (or 2^3)), the type of proxy (variance or curvature), and the type of aggregation (variance or mean). Variance proxies are only aggregated via mean such that there are only six instead of eight ensuing analytical blocks. In this main article, we will mostly discuss the results from the 2^4 variance proxy models, as well as the 2^4 and 2^3 mean-aggregated curvature models. As a reason for this preferred exposure of mean-aggregated models, we would like to argue that mean aggregation, as used by the majority of authors, is the more sensible way of aggregating the baseline data. When investigating motor interference, we are mostly interested in the average magnitude of variation in the direction orthogonal to the main axis of movement rather than the variance of this variation. Where mean aggregation seems to yield a spatially interpretable measure, variance aggregation does not, but rather tells us about the "volatility" of the spatial variation. Analyses of the "dispreferred" analytical blocks, as well as the tables listing the results based on the 2^3 models, are provided in the supplementary materials. We will refer to parts of these results in the following where necessary.

In the 2^3 models, curvature and variance data were both split into two parts along the factor *movement direction* (*mDir*) such that *left-right* and *forth-back* data was analyzed separately. The split was performed as we did not consider movement direction a genuine experimental factor: *forth-back* movements were only added tentatively because they seemed plausible in this setup but have no equivalent in previous motor interference research.

On the 2^3 models, we executed both parametric tests (mixed ANOVAs) and rank-based non-parametric tests (`nparLD` [31]) because some test configurations violated the normality assumption required by the parametric tests. Particularly in the full models, the normality of residuals was frequently violated despite a Box-Cox transformation of the data (cf. Table 1). In these cases, we need to be careful about the results of the parametric tests and should optimally perform non-parametric tests. Unfortunately, we could not obtain non-parametric tests that were applicable or working on the full models, and we therefore use the analytical results of the smaller models for additional support. However, the statistical results based on the smaller models are only valid under the assumption that the variation measurements of a run are independent of the movement direction of the previous run.

On all of those datasets where both participants and outliers on the run level were removed, as well as those where only run-level outliers had been removed, both parametric and non-parametric tests were performed. On complete datasets without outlier removal, only non-parametric tests were performed due to the volatility of parametric tests on these types of data. Our tables list only those tests and test configurations that yielded effects or interactions that were statistically significant ($p < .05$). The total number of statistical tests performed in each analytical block is stated in the captions of the respective tables (Table 1 in the following, and Tables 6, 8, and 10 in the appendix).

2.7 Research Hypotheses

Given the research questions from Section 1.3 and the described experimental setup, our hypotheses are the following (relevant research questions are referenced in parentheses, and curvature and variance as dependent variables are abbreviated with *DVs*):

(H1) *DVs* will be larger in *incongruent* conditions as compared to *congruent* conditions. This should be the case at least in the conditions that most closely resemble the established paradigm—that is, in *simultaneous* × *left-right* conditions. (~*Q1, Q2*)
(H2) The difference in *DVs* between *congruent* and *incongruent* conditions will be larger for *primed* participants. (~*Q2*)
(H3) *DVs* will be larger in *incongruent* than *congruent forth-back* movements. (~*Q2*)
(H4) There will be a difference in *DVs* in at least one of the *consecutive* conditions as compared to the respective simultaneous condition. (~*Q3*)
(H5) Participants will rate their *engagement* higher in *congruent* as compared to *incongruent* conditions. (~*Q4*)
(H6) Participants will rate their *perceived behavioral interdependence* as higher in *incongruent* than in *congruent* conditions. (~*Q4*)

3 RESULTS

3.1 Motion Data

We did not take a baseline measure of participants' movements and will therefore use the data from the *congruent-consecutive* condition to describe participants' movement lengths. *Left-right* movements were on average approximately 31 cm long ($M = 30.94, SD = 8.18$). At approximately 27 cm, participants' *forth-back* movements were on average slightly shorter but also exhibited less variation along the main axis of movement ($M = 27.09, SD = 4.81$).

Table 1 shows those test and data configurations for which the statistical tests yielded significant ($p < .05$) effects and interactions.[4] The table contains the results for tests of both the 2^4 model of the variance proxies as well as the 2^4 model of the mean-aggregated curvature proxies. We did not include results approaching significance ($.05 \leq p < .1$) but note that some tests results missed the $p = .05$ threshold only by a small margin. Thus, the absolute count of significant results should be regarded with a certain caution, as the particular choice of outlier removal, normalization, or data pruning may turn significant results into non-significant ones or vice versa.

3.1.1 Full Model, Mean-Aggregated Curvature Proxies. Half of all tests, 6 of 12, which are all tests issued on the *No Gaps* dataset, yield a statistically significant interaction between *movement direction* and *congruency*. The effect size of all of these interactions is small ($.01 \leq \eta^2_G < .06$). Post hoc tests performed to tease apart these interactions (Table 2) show that the difference in the means of curvatures (*MoCs*) between *congruent* and *incongruent* runs are only significant for participants'

[4]Note that we do not state the degrees of freedom for the *F*-statistic in the tables but only in the captions.

Table 2. Post Hoc Test Results for 2^4 Tests with Mean Aggregation

Conf. #	*Effect/Interaction*	*Fixed Factor*	*Variable Factor*	*Result*	*Mean/SD of Aggregated Proxy Data at Stated Factor Level* (**bold**)	
					congruent	**incongruent**
1	move.-dir. x cong	move.-dir.='f b'	congruency	$\chi^2(1) = 8.0$; $p = .009$	1.05/0.24	0.9/0.35
2	move.-dir. x cong	move.-dir.='fb'	congruency	$\chi^2(1) = 6.0$; $p = .029$	1.04/0.25	0.91/0.35
3	move.-dir. x cong	move.-dir.='fb'	congruency	$\chi^2(1) = 6.6$; $p = .02$	1.05/0.32	0.92/0.41
					left-right	**forth-back**
4	move.-dir.	n/a	n/a	n/a	48.68/33.06	12.34/7.87
5	move.-dir.	n/a	n/a	n/a	51.95/36.30	12.95/8.39
6	move.-dir.	n/a	n/a	n/a	53.09/36.67	13.76/9.18
					congruent	**incongruent**
7	move.-dir. x cong	move.-dir.='fb'	congruency	$\chi^2(1) = 8.9$; $p = .006$	20.4/11.4	17.33/11.43
8	move.-dir. x cong	move.-dir.='fb'	congruency	$\chi^2(1) = 7.27$; $p = .014$	20.56/11.29	18.13/11.84
9	move.-dir. x cong	move.-dir.='fb'	congruency	$\chi^2(1) = 8.2$; $p = .009$	21.33/11.42	18.58/12.18
					left-right	**forth-back**
9	cong x move.-dir.	cong='cong'	movm.-dir.	$\chi^2(1) = 47.26$; $p < .001$	62.92/35.68	21.33/11.41
9	cong x move.-dir.	cong='incong'	movm.-dir.	$\chi^2(1) = 75.26$; $p < .001$	69.94/40.56	18.58/12.18

Note: *Conf. #*: Index of relevant test configuration for which the stated post hoc test was performed (cf. Table 1). *Effect/Interaction*: The main effect or interaction under consideration. *Fixed Factor*: Fixed factor of the test and level to which it was fixed. *Variable Factor*: Variable factor whose levels were compared to each other in the test. *Result*: *p*-Value and χ^2 statistic of the post hoc test. *Mean/SD*: Means and standard deviations of the mean-aggregated curvature values for the specified levels. For significant main effects that are independent of any interaction, only the descriptive statistics are stated. cong: congruency; move.-dir.: movement-direction.

forth-back movements. Rather surprising, however, is the direction of this difference: the average *MoCs* are higher in runs where participants and the robot act in a *congruent* manner as compared to them performing *incongruent* movements, something akin to a negative or inverted motor interference effect. In 3 of the 6 relevant tests, the models' residuals are not normally distributed, and we should therefore lower our trust in the results. The circumstance that their results are no different from those models where the normality criterion is not violated does lend them support. We will additionally supplement this analysis with the analytical results from the 2^3 models for further reassurance in the following.

In 9 of 12 tests, *movement direction* is a significant main effect. In 6 of these the effect size is large, and it is small in the remaining 3. The average *MoCs* of *left-right* movements are considerably higher than the ones of *forth-back* movements.

3.1.2 Analytical Results Based on Split 2^3 Models. The analysis of the 2^3 models based on variance proxies and *forth-back* movement data (appendix, Section A.2.1) yields a small indication (1 of 10 tests) that among *primed* participants, mean variances of incongruent runs were higher than those of congruent ones, indicating that motor interference had occurred.

The test results of the curvature-based 2^3 models based on *forth-back* data and mean-aggregated curvature proxies are consistent with the curvature-based results from the 2^4 model (Section 3.1.1). All tests based on the *No Gaps* dataset (10 of 20) indicate that some form of inverse motor interference has occurred. In other words, on average, the *MoCs* of *congruent* runs are higher than those of *incongruent* ones.

In addition to the support of the results based on the 2^4 models, the non-parametric tests yield one additional result. On the *left-right* dataset, the five non-parametric tests launched on the *No Gaps* dataset indicate that *congruency* is a main effect with small effect size. Two more

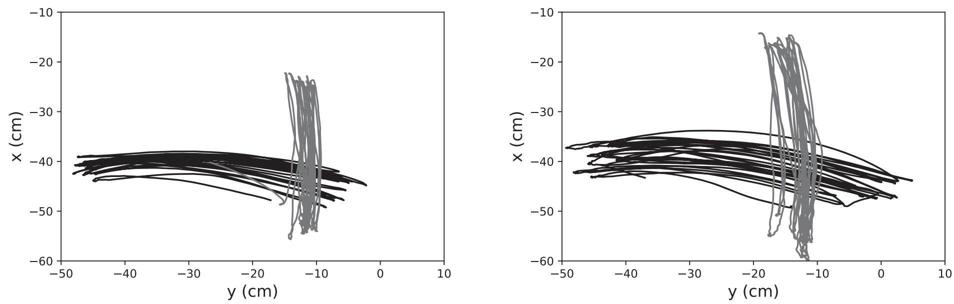

Fig. 6. Left: Participant P20's arm movements during two *congruent* runs (left) and during two *incongruent* runs (right). *Left-right-left* movements are the horizontal movements plotted in black, and *forth-back-forth* movements are the vertical movements plotted in dark gray. The units are displayed as measured by the Polhemus Liberty sensor and relative to its coordinate system (cf. Figure 4).

non-parametric tests launched on the *Gaps* dataset just missed the $p < .05$ threshold, and a third one was just below the $p < .1$ threshold. The direction of difference in this movement direction, opposite to what was indicated for *forth-back* movements, is the one we would expect: on average, the *MoCs* of *incongruent* runs are higher than in *congruent* ones. Thus, if our analysis were based solely on the *No Gaps* dataset, and if we decided to use non-parametric tests exclusively, our conclusion would be clear: motor interference has taken place in the way it is documented in the literature (cf. the visualisations in Figure 6).

A single test based on the *left-right* models additionally flags up the interaction *congruency* × *interaction mode* as significant. The post hoc analysis of this interaction (see Table 7) yields the unexpected result that the differences in curvature caused by differences in congruency were only significant in consecutive interactions. However, given that the significance of this post hoc result is slightly below the .05 margin ($p = .07$), given that it is the only significant result based on the *Gaps* dataset, and given that all other tests with significant results do not indicate a dependency of *congruency* with respect to *interaction mode*, we deem this result an outlier. The effect may be a potential random side effect of the pruning of the boundary data.

Due to space constraints, we will only summarize the results of the tests based on variance-aggregated curvature proxies as we deem this aggregation method inappropriate. The results, however, are largely comparable with the ones originating from the mean-aggregated curvature proxies. Tests operating on the 2^4 models identify *movement direction* as a significant main effect and/or part of a significant interaction (11/18, cf. Table 8). Among these interactions, we find one with the factor *prime* that was not flagged up in the analysis based on mean aggregation. *Congruency* is identified twice as significant main effect and once as part of a significant interaction with *movement direction*. In the context of this interaction, differences in *VoCs* between *congruent* and *incongruent* runs are only significant for *forth-back* movements—the same result as indicated by the analysis of the mean-aggregated curvature models. Again the direction of difference is contrary to what we would expect: on average, *VoCs* in *congruent* runs are higher than in *incongruent* runs. The analyses of the 2^3 models (see Table 10) strengthen the results of the 2^4 models: in the *forth-back* direction, *congruency* is a significant main effect in 21 of 30 tests. In all cases, the mean *VoCs* are higher in *congruent* runs as compared to *incongruent* runs. See the supplementary materials for a more detailed analysis of the variance-aggregated curvature models.

3.1.3 Evaluation of Hypotheses 1 through 4.

(H1) We hypothesized that the *DVs* would be larger in *incongruent* conditions as compared to *congruent* conditions. The analyses of the 2^3 model for *left-right* movements give some

Table 3. Internal Reliability of Constructs from Questionnaire in Terms of Cronbach's α and Results of Normality Tests Before and After Box-Cox Transformation

Construct	*Cronbach's* α	*Normality Test (Original)*	*Lambda*	*Normality Test (After Transform.)*
TPI-R	0.94	S-W pass	n/a	n/a
TPI-SPA	0.82	$p = .030$	0	S-W pass
TPI-SPP	0.34	n/a	n/a	n/a
TPI-Eng	0.8	S-W pass	n/a	n/a
NMI-PBI	0.74	S-W pass	n/a	n/a
NMI-PAE	0.42	n/a	n/a	n/a
NMI-CoP	0.57	n/a	n/a	n/a

Note: Constructs with at least acceptable reliability and better are set in bold. The entries in the *Normality* columns state whether the residuals of the respective linear model passed the Shapiro-Wilk test ($p > .05$) or not, and if not, with which p-value they failed. In case of failure, the data was Box-Cox transformed and the lambda value used for the transformation is stated in the respective column. Constructs with insufficient reliability were dropped from the analysis, hence the n/a values.

support for this hypothesis—that is, that motor interference has indeed occurred. This support is contingent, however, upon the use of *curvature* as dependent variable and the use of the *No Gaps* dataset as an underlying basis. *Forth-back* movements exhibited the opposite effect. The *curvature* of participants' *forth-back* movements was larger in *congruent* runs as compared to *incongruent* runs.

(H2) We hypothesized that the differences in *DVs* between *congruent* and *incongruent* conditions would be larger for *primed* participants. In the split 2^3 models, we found one test configuration using variance proxies that supported this hypothesis for *left-right* movements. However, the other nine test configurations did not flag up *prime* as a significant main effect, or as part of a significant interaction, nor did the tests based on curvature proxies. Hence, there is only very limited support for this hypothesis.

(H3) We expected that the *DVs* would be larger in *incongruent* runs as compared to *congruent* runs also for *forth-back* movements. We found precisely the opposite effect. Hence, *H3* is not supported by our experimental data.

(H4) We hypothesized that there would be at least one *consecutive* condition where there was a difference in *DVs* compared to the respective *simultaneous* conditions. Only one test based on the 2^3 model of *left-right* movements flags up the *interaction mode* as part of a significant interaction with *congruency*. Counter-intuitively, the post hoc test indicates that the difference in curvature between *congruent* and *incogruent* runs were only significant in *consecutive* interactions but not in simultaneous ones. Only a single test yields this results, and it is based on the *Gaps* dataset, which otherwise does not yield any significant effect or interactions. We are therefore skeptical about this result and do not consider this sufficient support for this hypothesis.

3.2 Questionnaires

3.2.1 Internal Reliability. Four of the seven constructs used in the questionnaire had acceptable ($\alpha > 0.7$), good ($\alpha > 0.8$), or excellent ($\alpha > 0.9$) internal reliability in terms of their Cronbach alpha values (Table 3). These four constructs and, tentatively, *co-presence* (*NMI-CoP*) were subsequently used as dependent variables in the statistical analyses of the questionnaire data.

Table 4. Statistical Results for Questionnaire Data

Construct	*Effect/Interaction*	*Result (Parametric Test)*	*Result (Non-Parametric Test)*
TPI-R	iMode	$F = 11.40, p = .003, \eta^2_G = .019$	$F = 9.25, p = .002$
TPI-SPA	none	N/A	
TPI-Eng	iMode	$F = 3.73, p = .066, \eta^2_G = .009$	$F = 3.22, p = .072$
NMI-PBI	cong.	$F = 11.58, p = .003, \eta^2_G = .050$	$F = 14.06, p < .001$
	iMode	$F = 6.68, p = .017, \eta^2_G = .035$	$F = 6.12, p = .013$
NMI-CoP	prime x cong.	*no significant effect*	$F = 4.51, p = .034, \eta^2_G = .019$

Note: Listed are statistically significant effects and interactions ($p < .05$) and effects and interactions approaching significance ($p < .1$). Effect sizes (η^2_G) are reported for parametric tests only unless the significant results are exclusively non-parametric. *Parametric*: Results of ANOVA. *Non-Parametric*: Results of the rank-based non-parametric test implemented by `nparLD`, cong.: congruency. Degrees of freedom are not printed in the table and would be $F(1,22)$ in all cases.

3.2.2 Factorial Analysis. As all items were evaluated on a Likert scale, the ratings were considered interval data. We performed separate tests for each factor: *Social Richness*, *Social Presence–Active Interpersonal*, *Engagement*, *Perceived Behavioral Interdependence*, and *Co-Presence*. Following Shapiro-Wilk tests on the respective models, only the residuals of *Social Presence–Active Interpersonal* were not normally distributed ($W = .97, p = .030$). The measurements of this factor were subsequently Box-Cox transformed. A Shapiro-Wilk test on the residuals of the resulting model indicated that the transformed measurements were not violating the normality constraint any longer. We nevertheless also performed non-parametric tests for comparison.

One participant failed to complete the items associated with *Social Richness*. For this participant, we used multiple imputation to replace the missing values.

Table 4 summarizes the results of the factorial analysis. If not stated otherwise, we will by default use the results of the parametric tests as points of reference in the following discussion, as the results of both parametric and non-parametric tests are largely identical in terms of the factors being marked as significant. The only exception are the analyses of *Co-Presence*.

The ANOVA reveals that the interaction mode had a significant effect on participants' ratings of *social richness* (*TPI-R*) ($F(1, 22) = 11.40, p = .003$). Participants in the simultaneous condition rated the *social richness* of the interaction as higher ($M = 4.02, SD = 1.41$) than in the consecutive condition ($M = 3.62, SD = 1.52$, cf Table 5). In other words, they perceived the simultaneous interaction with the robot as somewhat more sociable, more immediate, or more sensitive than when acting consecutively relative to the robot. The effect size of this perceived difference is comparatively small ($\eta^2_G = .019$). A similar result, although just below the significance threshold ($F(1, 22) = 3.73$, $p = .066$), can be observed for the rating of *engagement* or *mental immersion* (*TPI-Eng*).

Ratings of *perceived behavioral interdependence* (*NMI-PBI*) were significantly impacted by both the particular *interaction mode* ($F(1, 22) = 6.68, p = .017$) and whether participants' movements were *congruent* ($M = 3.61, SD = 1.33$) or *incongruent* ($M = 3.01, SD = 1.37$) ($F(1, 22) = 11.58$, $p = .003$). Although the effect size of this perceived difference was still small for the *interaction mode* ($\eta^2_G = .035$), it just missed medium size for *congruency* ($\eta^2_G = .05$).

For judgments of *co-presence* (*NMI-CoP*), the ANOVA indicated that the interaction *prime x congruency* was below the significance level ($F(1, 22) = 2.79, p = .109, \eta^2_G = .013$), yet it was well above the significance threshold in the non-parametric test ($F(1, 22) = 4.51, p = .034, \eta^2_G = .019$). The parametric post hoc tests of this interaction yields a significant contrast between *congruent* ($M = 3.92, SD = 0.72$) and *incongruent* ($M = 4.20, SD = 0.91$) forms of interaction ($\chi^2(1) = 4.62$, $p = .032$) but only for participants who had not been *primed* prior to the interaction. The equivalent non-parametric post hoc test for these two factors does not corroborate the parametric results and, more generally, does not yield any significant result.

Table 5. Descriptive Statistics for Levels of Main Effects or Significant Interactions

Construct	*Main Effect*		*Levels* *Mean/SD*	*Mean/SD*
			simultaneous	**consecutive**
Social Richness (TPI)	interaction mode		4.03/1.41	3.62/1.52
Engagement (TPI)	interaction mode		4.79/1.24	4.53/1.48
Perceived Behavioral Interdependence (NMI)	interaction mode		3.56/1.36	3.06/1.37
			congruent	**incongruent**
	congruency		3.61/1.33	3.01/1.37
Construct	*Interaction*	*Fixed Factor*	*Levels Var. Factor*	*Mean/SD*
Co-Presence (NMI)	[P] prime x congruency	prime = *not watching*	congruent	3.92/0.72
			incongruent	4.20/0.81

Note: Co-Presence interaction: both parametric and non-parametric post hoc tests were performed. Only the parametric post hoc indicated that the stated interaction was significant ($\chi^2(1) = 4.62$, $p = .032$). [P], parametric post hoc test.

3.3 Evaluation of Hypotheses H5 and H6

(H5) We did not find support for this hypothesis.

(H6) This hypothesis was not supported by participants' self-reports. Contrary to our expectation, participants felt more inter-dependent on the robot's action when engaging in congruent movements. Thus, motor interference, if anything, correlates with participants feeling less inter-dependent from the machine rather than vice versa. Additionally, participants perceived a higher level of interdependence during *simultaneous* interaction than during *consecutive* interaction.

4 DISCUSSION

The average movement lengths of our participants' transitive movements, at approximately 30 cm (*left-right*) and 27 cm (*forth-back*), were somewhat shorter than the intransitive ones depicted or reported in the literature.

Both Kupferberg et al. [23] and Gowen et al. [19] instructed participants to execute 50 cm long movements in both directions. Chaminade et al. [12] do not present descriptive statistics for participants' movement lengths, but judging by the published exemplary trajectory plot, participants' movements were also approximately 50 cm long in the vertical direction. Movements in the horizontal direction appear to be a few centimeters longer in the depicted plot. Kilner et al. [22] do not provide descriptive statistics on movement lengths either, but judging by their published example plots, their participants' vertical arm movements were just under 70 cm, and their horizontal movements were just above 60 cm long.

Hence, comparatively speaking, our participants' transitive movements were at least 40% shorter than the intransitive movements documented in the literature. This may be one factor that makes it more difficult for motor interference to be detected in transitive movements: less long movements mean less opportunity to accumulate variance or curvature on the level of a single movement.

4.1 Answers to Research Questions

(*Q1*) In terms of the suitability of the established interference detection techniques toward more naturalistic scenarios, we are only moderately optimistic. For *left-right* movements, we obtained some confirmation that motor interference had occurred. The effect sizes of the effect, however, were small throughout such that it is questionable whether the number of repetitions could be reduced considerably.

Furthermore, the effect only materialized when the motion segments were not pruned, which indicates a strong sensitivity of the effect to such treatment. Whereas some authors use (semi-) automatic segmentation methods combined with manual checks, others employ fully automatic methods that may induce a significant amount of pruning as a side effect. Our results suggest that the amount of pruning should be documented and made explicit in the literature.

For *forth-back* movements, we observed an unexpected "inverse motor interference effect," which we will discuss in greater detail in Section 4.3. If our guess with respect to the cause of this effect is correct, then great care has to be exercised with regard to the spatial design of the target setup. As a result of our decision to place the participant and robot face to face and align them in the most natural way, namely directly opposite each other, we may have unknowingly opened the door to a distractor interference effect.

Independent of this complication in terms of the spatial setup, we also need to acknowledge the methodological complications that we encountered. We found the two established methods for quantifying spatial variation, variance with respect to the orthogonal plane versus curvature, to be not entirely equivalent (see Section 4.2).

(*Q2*) As already discussed under *Q1*, *movement direction* was the dominant factor in our experiments. Apart from *movement direction* and *congruency*, very few factors were consistently flagged as main effects or formed part of significant interactions. We have a weak indication that the *prime* may have impacted upon motor interference in such a way that *primed participants* exhibited motor interference when performing *left-right* movements. This indication, however, is based on a single test.

(*Q3*) If the incurred motor interference were to differ significantly between simultaneous and turn-based interactions, we would expect to see statistically significant interactions between *congruency* and *interaction mode*. Only a single test based on a 2^3 model of *left-right* movements flags up such an *interaction mode* interaction. Given the fact that the test is based on the *Gaps* dataset, which under other test configurations does not show any indications of motor interference having occurred, we do not take this result at face value. However, those tests that indicate significant differences between *congruent* and *incongruent* movements do not mark *interaction mode* as a significant factor. This seems to indicate that motor interference extends to turn-based interaction, as we would have expected the effect to be limited to *simultaneous* interactions otherwise.

(*Q4*) The only factor where *congruency* had a significant impact is *perceived behavioral interdependence*. Assuming that motor interference, in its regular form as well as in its "inverse" form, has indeed occurred, this result indicates that participants do have some awareness of its impact. The "direction of impact", however, is opposite to what we expected. Rather than reporting higher levels of perceived interdependence under the influence of motor interference, participants reported higher levels of such interdependence in the absence of the effect. A possible explanation may have to do with the fact that congruent joint actions arguably exhibit a higher degree of behavioral sychronization as compared to incongruent ones.

Independent of *congruency*, participants rated *simultaneous* interactions as more *engaging*, more *socially rich*, and perceived a higher level of *behavioral interdependence* while engaged in these.

4.2 Differences Between Variance and Curvature Proxies

Methodologically noteworthy is the observation that our tests operating on *variance* proxies marked out less factors and combinations of factors as statistically significant than did tests that operate on *curvature* proxies. This means that our results are somewhat dependent on our choice of proxy, and the two proxies are not entirely commensurate in terms of the ensuing statistical outcomes. Kupferberg et al. [24] decomposed the standard variance measure (*SA*) into three distinct measures: tilt with respect to the error plane (*TA*), deviation of single movements with respect to the "global" line fit of movements of that run (*DA*), and a curvature measure where an interpolated line was used as line of reference (*CA*). The authors found that *TA* and *DA* correlated more strongly with *SA* and hence contributed more to the *SA* measure than did *CA*.

Kupferberg et al.'s curvature measure differs from the one employed by us and Chaminade et al. [12], Gowen et al. [18], and Oztop et al. [34] in the way the "line of reference" is set. Whereas our measure uses the line connecting the start and end points of the respective movement, Kupferberg et al. use the line that interpolates the respective movement best. The difference in the two ways of quantifying curvature is presumably small. Variance or standard analysis and curvature analysis are not commensurate in that the standard analysis is majorly impacted by movement being tilted with respect to the reference plane, whereas curvature analysis factors out tilt by means of adapting the line of reference. The difference between the two proxy measures diminishes the comparability of existing research based on these two methodological strands.

We found it unhelpful that previous researchers in the field omitted to report effect sizes for the detected significant effects. Although it is possible to assess the significance of modeled factors upon a dependent variable such as spatial variation under ignorance of effect sizes, the non-reporting of these has been criticized in the adjacent fields of psychology [14] and human-computer interaction [38]. In our particular case, the omission of effect sizes makes it exceedingly difficult to assess how much the constraints of the methodology can be relaxed to render it embedable in naturalistic interaction scenarios.

4.3 On "Inverse" Motor Interference

We are not aware of any other authors having observed something akin to inverse motor interference, where participants' movement variability is higher when the movement is performed *congruently* as compared to it being performed *incongruently* with respect to the model's movement. Our experimental setup differs from the established paradigm measuring motor interference in three important ways:

(1) The employed movements are transitive in the sense that participants are moving an object to a self-chosen target location. These motor actions are not perfectly transitive, however, as the grip on the object is never released and the "pick up" part of the motor action is only performed once at the start of the run.
(2) The *forth-back* movements, where inverse motor interference occurred, are orthogonal to the two main directions of movements in the established paradigm. *Forth-back* movements differ from vertical and horizontal waving motions in that the interactors' hands approach each other in the *forth-back* dimension when performing the movement in a roughly synchronous manner. That means both actors' arms and hands approach the imagined middle line of the table that separates them.
(3) We did not instruct participants with respect to the details of the movements. Instead, we told them to move roughly forth-and-back, or left-and-right, and demonstrated these two movements to them. The movements on demonstration were approximately straight (i.e., parallel to the respective coordinate axis). However, we did not emphasize that participants should move precisely this way. Yet, during the experiment, we did not observe participants deliberately moving in a diagonal manner—that is, toward the robot's right hand's factual but unmarked target location.

Bouquet et al. [5] observed that transitive (i.e., goal-directed) motor actions typically incur stronger effects of motor interference than do intransitive, or meaningless, movements. Rizzolatti et al. [37] go so far as to distinguish two types of motor resonance: *low-level* and *high-level* resonance. This distinction is based solely on whether the movement in question is goal-directed and is therefore a genuine motor action. If it is goal-directed, they would speak of *high-level* motor resonance. If the movement is goal-less, the motor resonance invoked is considered to be *low level*. This distinction is at least partially based on the authors' observation that certain mirror neurons appear to code exclusively for complete motor actions, whereas others code for certain types of movements, irrespective of the larger action of which the movement forms part.

For our purposes, this means that the move from intransitive to transitive movements may have also brought about a transition from low-level to high-level resonance. Moving to high-level resonance, in turn, may bring about sensitivities and side effects that are not relevant to low-level resonance.

One side effect of actions having spatial targets is the potential for competing or distracting target locations. Welsh and Elliott [47] demonstrated in a setup employing touch actions on visual targets that the presence of a visual distractor can impact the movement trajectories. In their case, the trajectories deviated toward the location of the distractor.

It is conceivable that our participants consciously or subconsciously perceived either the unmarked target location of the robot's movements, or the robot's hand itself as distractor location or distractor object. If so, this could have caused their movements to veer toward that location. In this scenario, we might see a competition between spatial effects caused through motor resonance, *imitative compatibility effects* in terms of Catmur and Heyes [9], and spatial effects induced by the mere location of the distractor, *spatial compatibility effects*.

In our setup, the directions of the two types of effects are opposite each other: both veering toward the distractor and veering to and from the distractor would increase the spatial variation, whereas the imitative compatibility should render the movement less variable.

Catmur and Heyes' study [9] tested and confirmed the independence of *imitative compatibility effects* from *spatial compatibility effects*. Utilizing a response-time paradigm, the authors observed that the effect size of the spatial compatibility effects was larger than that of the effects based on motor resonance. Although our dependent variable is of spatial rather than temporal nature, it is conceivable that spatial compatibility effects may also dominate imitative compatibility effects if measured in spatial terms rather than temporal terms. Thus, the effect that we identified earlier as "inverse motor interference" may be indeed a spatial compatibility effect that works in the opposite direction of the imitative compatibility effect. As a consequence, it may dominate the combined interference effect, thereby rendering any potentially occurring imitative compatibility effect invisible.

5 CONCLUSION

Our experiment provides some support that motor interference measurements based on spatial variation do extend to transitive *left-right* movements. The detection, however, is very sensitive to the pruning of boundary data between motion segments, and the detected effect sizes are small. It is therefore questionable whether a further "relaxation of conditions" such as a reduction of repetitions toward more naturalistic scenarios is possible without losing the effect. For *forth-back* movements, we found an effect that is inverse to what has been reported in the literature. The spatial variation of participants' arm movements was smaller in incongruent interaction as compared to congruent interaction. We hypothesize that the potentially occurring motor resonance effect has been overridden by a simultaneously occurring spatial compatibility effect.

Based on these observations, it does not seem likely that this method of measuring motor resonance can be embedded in real-world scenarios to detect the presence of motor resonance and, by extension, the real-time quality of interaction. Alternative indices for motor resonance based on motor synchronization such as motor contagion may be better candidates if the intended quantitative assessment of an interaction ought to happen in near real time. On a more methodological level, we found the two established methods for quantifying motor interference to be non-equivalent. Although they may measure conceptually overlapping variants of "spatial variation," they do not appear to measure the same spatial concept.

6 FUTURE WORK

Given the likely presence of the distractor effect mentioned previously, future experiments involving spatial variation measures will need to be adjusted accordingly. Aligning the actors such that the relevant arms are located in one line along the *forth-back* axis would probably eliminate this effect, yet such alignment is also somewhat unnatural. Instead of aligning the actors faces such that the line connecting them is parallel to the *forth-back* axis of the table, it would be their right arms that would be aligned along this axis. This effectively produces a relative offset or shift of their faces along the *left-right* axis of the table. In other words, the connecting line between the two faces would be somewhat diagonal and not parallel to the *forth-back* axis any more. This may have undesired side effects in terms of potentially ensuing social interaction. Only an experimental implementation can shed light on this issue.

A less well-explored path for detecting motor resonance has been taken by Bisio et al. [4]. Instead of using spatial variation or reaction time as indices for motor resonance, Bisio et al. used temporal characteristics of participants' movements and their degree of adaptation to the ones exhibited by the model as dependent variables ("motor contagion"). Although motor contagion has been used far less often in the context of motor resonance research, we deem it worthwhile to explore whether motor contagion is triggered and modulated to a similar or even equal degree by motor resonance as is motor interference. Compared to the complications we encountered measuring spatial interference effects, motor contagion most likely involves fewer researchers' degrees of freedom and even promises to be less sensitive to tacit features of the spatial arrangement within which the agents interact.

REFERENCES

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48.

[2] Frank J. Bernieri and Robert Rosenthal. 1991. Interpersonal coordination: Behavior matching and interactional synchrony. In *Fundamentals of Nonverbal Behavior*, R. S. Feldman and B. Rimé (Eds.). Cambridge University Press, New York, NY, 401–432.

[3] Frank Biocca, Chad Harms, and Judee K. Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence: Teleoperators and Virtual Environments* 12, 5 (2003), 456–480.

[4] Ambra Bisio, Alessandra Sciutti, Francesco Nori, Giorgio Metta, Luciano Fadiga, Giulio Sandini, and Thierry Pozzo. 2014. Motor contagion during human-human and human-robot interaction. *PLoS ONE* 9 (2014), 1–10.

[5] Cédric A. Bouquet, Thomas F. Shipley, Rémi L. Capa, and Peter J. Marshall. 2011. Motor contagion. *Experimental Psychology* 58, 1 (2011), 71–78.

[6] Edgar Brunner, Sebastian Domhof, and Frank Langer. 2002. *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. J. Wiley.

[7] S. van Buuren and K. Groothuis-Oudshoorn. 2011. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45, 3 (2011), 1–68.

[8] Justine Cassell and Kristinn R. Thorisson. 1999. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence* 13, 4–5 (1999), 519–538.

[9] Caroline Catmur and Cecilia Heyes. 2011. Time course analyses confirm independence of imitative and spatial compatibility. *Journal of Experimental Psychology: Human Perception and Performance* 37, 2 (2011), 409–421.

[10] Thierry Chaminade. 2011. A social cognitive neuroscience stance on human-robot interactions. In *BIO Web of Conferences*, Vol. 1. EDP Sciences, 00014-p.1–00014-p.4.
[11] Thierry Chaminade and Gordon Cheng. 2009. Social cognitive neuroscience and humanoid robotics. *Journal of Physiology-Paris* 103, 3–5 (2009), 286–295.
[12] Thierry Chaminade, David W. Franklin, Erhan Oztop, and Gordon Cheng. 2005. Motor interference between humans and humanoid robots: Effect of biological and artificial motion. In *Proceedings of the 2005 4th International Conference on Development and Learning*. IEEE, Los Alamitos, CA, 96–101.
[13] Crystal Chao and Andrea L. Thomaz. 2013. Controlling social dynamics with a parametrized model of floor regulation. *Journal of Human-Robot Interaction* 2, 1 (2013), 4–29.
[14] Geoff Cumming. 2014. The new statistics: Why and how. *Psychological Science* 25, 1 (2014), 7–29.
[15] Helios De Rosario-Martinez. 2015. *phia: Post-Hoc Interaction Analysis*. R package version 0.2-1. Retrieved November 6, 2019 from https://CRAN.R-project.org/package=phia.
[16] Eric Jones, Travis Oliphant, and Pearu Peterson. 2001. SciPy: Open Scientific Tools for Python. Available at http://www.scipy.org/.
[17] Andy Field, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. SAGE Publications.
[18] David W. Franklin, Rieko Osu, Etienne Burdet, Mitsuo Kawato, and Theodore E. Milner. 2003. Adaptation to stable and unstable dynamics achieved by combined impedance control and inverse dynamics model. *Journal of Neurophysiology* 90, 5 (2003), 3270–3282.
[19] E. Gowen, J. Stanley, and R. C. Miall. 2008. Movement interference in autism-spectrum disorder. *Neuropsychologia* 46, 4 (2008), 1060–1068.
[20] David C. Howell. 2012. *Statistical Methods for Psychology* (8th ed.). Cengage Learning, Belmont, CA.
[21] James Kilner, Antonia F. de C. Hamilton, and Sarah-Jayne Blakemore. 2007. Interference effect of observed human movement on action is due to velocity profile of biological motion. *Social Neuroscience* 2, 3–4 (2007), 158–166.
[22] James M. Kilner, Yves Paulignan, and Sarah-Jayne Blakemore. 2003. An interference effect of observed biological movement on action. *Current Biology* 13, 6 (2003), 522–525.
[23] Aleksandra Kupferberg, Stefan Glasauer, Markus Huber, Markus Rickert, Alois Knoll, and Thomas Brandt. 2011. Biological movement increases acceptance of humanoid robots as human partners in motor interaction. *AI & Society* 26, 4 (2011), 339–345.
[24] Aleksandra Kupferberg, Markus Huber, Bartosz Helfer, Claus Lenz, Alois Knoll, and Stefan Glasauer. 2012. Moving just like you: Motor interference depends on similar motility of agent and observer. *PLoS ONE* 7, 6 (2012), e39637.
[25] Michael A. Lawrence. 2016. *ez: Easy Analysis and Visualization of Factorial Experiments*. R package version 4.4-0. Retrieved November 6, 2019 from https://CRAN.R-project.org/package=ez.
[26] Roman Liepelt and Marcel Brass. 2010. Top-down modulation of motor priming by belief about animacy. *Experimental Psychology* 57, 3 (2010), 221–227.
[27] Matthew Lombard, Theresa B. Ditton, and Lisa Weinstein. 2009. Measuring presence: The Temple Presence Inventory. In *Proceedings of the 12th Annual International Workshop on Presence*. 1–15.
[28] Matthew R. Longo and Bennett I. Bertenthal. 2009. Attention modulates the specificity of automatic imitation to human actors. *Experimental Brain Research* 192, 4 (2009), 739–744.
[29] Ludovic Marin, Johann Issartel, and Thierry Chaminade. 2009. Interpersonal motor coordination: From human–human to human–robot interactions. *Interaction Studies* 10, 3 (2009), 479–504.
[30] Giorgio Metta, Giulio Sandini, David Vernon, Lorenzo Natale, and Francesco Nori. 2008. The iCub humanoid robot: An open platform for research in embodied cognition. In *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*. ACM, New York, NY, 50–56.
[31] Kimihiro Noguchi, Yulia R. Gel, Edgar Brunner, and Frank Konietschke. 2012. nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software* 50, 12 (2012), 1–23.
[32] Lindsay M. Oberman, Joseph P. McCleery, Vilayanur S. Ramachandran, and Jaime A. Pineda. 2007. EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing* 70, 13–15 (2007), 2194–2203.
[33] Stephen Olejnik and James Algina. 2003. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods* 8, 4 (2003), 434–447.
[34] Erhan Oztop, David W. Franklin, Thierry Chaminade, and Gordon Cheng. 2005. Human–humanoid interaction: Is a humanoid robot perceived as a human?*International Journal of Humanoid Robotics* 2, 4 (2005), 537–559.
[35] Ugo Pattacini. 2010. *Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub*. Ph.D. Dissertation. RBCS, Istituto Italiano di Tecnologia, Genoa, Italy.
[36] Miriam Rennung and Anja S. Göritz. 2016. Prosocial consequences of interpersonal synchrony. *Zeitschrift für Psychologie* 224, 3 (2016), 168–189.

[37] Giacomo Rizzolatti, Luciano Fadiga, Leonardo Fogassi, and Vittorio Gallese. 2002. From mirror neurons to imitation: Facts and speculations. In *The Imitative Mind: Development, Evolution, and Brain Bases*, A. N. Meltzoff and W. Prinz (Eds.). Cambridge University Press, New York, NY, 247–266.

[38] Judy Robertson and Maurits Kaptein (Eds.). 2016. *Modern Statistical Methods for HCI*. Springer.

[39] Alessandro Roncone, Ugo Pattacini, Giorgio Metta, and Lorenzo Natale. 2014. Gaze stabilization for humanoid robots: A comprehensive framework. In *Proceedings of the 2014 14th IEEE-RAS International Conference on Humanoid Robots (Humanoids'14)*. IEEE, Los Alamitos, CA, 259–264.

[40] Alessandro Roncone, Ugo Pattacini, Giorgio Metta, and Lorenzo Natale. 2016. A Cartesian 6-DoF gaze controller for humanoid robots. In *Proceedings of Robotics: Science and Systems*.

[41] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 4 (1974), 696–735.

[42] Qiming Shen, Hatice Kose-Bagci, Joe Saunders, and Kerstin Dautenhahn. 2009. An experimental investigation of interference effects in human-humanoid interaction games. In *Proceedings of the 2009 18th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'09)*. IEEE, Los Alamitos, CA, 291–298.

[43] Qiming Shen, Hatice Kose-Bagci, Joe Saunders, and Kerstin Dautenhahn. 2011. The impact of participants' beliefs on motor interference and motor coordination in human–humanoid interactions. *IEEE Transactions on Autonomous Mental Development* 3, 1 (2011), 6–16.

[44] James Stanley, Emma Gowen, and R. Chris Miall. 2007. Effects of agency on movement interference during observation of a moving dot stimulus.*Journal of Experimental Psychology: Human Perception and Performance* 33, 4 (2007), 915–926.

[45] Andrea L. Thomaz and Crystal Chao. 2011. Turn taking based on information flow for fluent human-robot interaction. *AI Magazine* 32, 4 (2011), 53–63.

[46] Ishabel M. Vicaria and Leah Dickens. 2016. Meta-analyses of the intra-and interpersonal outcomes of interpersonal coordination. *Journal of Nonverbal Behavior* 40, 4 (2016), 335–361.

[47] Timothy N. Welsh and Digby Elliott. 2005. The effects of response priming on the planning and execution of goal-directed movements in the presence of a distracting stimulus. *Acta Psychologica* 119, 2 (2005), 123–142.

[48] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for non-parametric factorial analyses using only ANOVA procedures. In *Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'11)*. ACM, New York, NY, 143–146. http://depts.washington.edu/aimgroup/proj/art/.

Received November 2018; revised June 2019; accepted July 2019