Analyzing Gaze and Hand Movement Patterns in Leader-Follower Interactions During a Time-Continuous Cooperative Manipulation Task

Minghao Cheng[1*], Minija Tamosiunaite[1,2*], Anoushiravan Zahedi[3,4], Ricarda I. Schubotz[3,4] and Florentin Wörgötter[1$]

[1]Inst. Physics 3, Computational Neuroscience, Georg-August University Göttingen, Germany

[2]Faculty of Informatics, Vytautas Magnus University, Kaunas, Lithuania

[3]Department of Psychology, University of Münster, Münster, Germany

[4]Otto-Creutzfeldt-Center of Behavioral and Cognitive Neuroscience, University of Münster, Münster, Germany

*Equal contribution

$: Corresponding author (worgott@gwdg.de)

Abstract:

Humans often interact with each other during daily life and many times one finds that one person (at least for some time) takes the lead while the other follows. Different from usual experimental settings in a lab, this happens without external influences and in a time-continuous manner. While knowledge that gazes onto the object to be manipulated precede a then-following manipulation action emerged long ago, little is known how gaze- and hand movement patterns develop in complex interactive scenarios where one person needs to consider the other's action when planning their own.

To address this question in this study we investigate predictive, planning-related behavior during a two-player (Leader and Follower) table-top game. In this game – called "do-undo" – the Leader has to perform pick and place actions, following simple rules, to change the configuration of objects on the table. The Follower then has to use different objects to re-create the object-configuration back to the one that had existed before the Leader had acted. We track eyes, hand-movements and individual objects on the table and determine relations between eyes and hands of both players, with emphasis on the differences in behavior of Leader and Follower. Data is recorded using a setup where gaze tracking is combined with multicamera tracking of motion of the subjects and of the configuration on the table. The game proceeds without any external trigger signals and we address the problem of finding accurate time-points for phasing of the cooperative manipulation by using touch sensors on the hands of the players to record touching and untouching events by which we can unequivocally define the different temporal (inter-)action intervals for analysis.

While gaze clearly precedes manipulation in all cases, we find clear differences in eye-movement patterns of Leader and Follower. The Leader makes more and earlier gazes on the objects to be manipulated. In a substantial proportion of trials, the Leader is predicting by his gaze not only their own, but also the required potential Follower-actions before their own manipulation. Also, the gaze data shows that the Leader takes more interest in re-checking the outcome of the do-undo manipulation pair. Indications of trying to memorize a configuration are also found in the gaze pattern of both. In addition, we often find sequences of alternating gazes towards objects and location where to put, which are much more expressed by the Leader than by the Follower. Intriguingly some patterns show that players pre-plan over longer periods and not just for their next action. This clearly indicates different decision-making and planning characteristics for Leader as compared to Follower, which happen not only for own plan but also for potential action plans by the other shining a light on the complex cognitive processes that exist in daily human-human interaction.

Introduction

Collaborative human-human interaction happens all the time in our daily lives, for example, when two people assemble a piece of furniture, when they fix a bike, cook a meal together. In these everyday moments of shared action, one person many times takes the lead – at least for some duration – and the other follows suit performing those actions that the leader currently does not do, to arrive at their conjoint action goals. Such interactions involve an ongoing acquisition of information by both agents, often from receiving a verbal instruction from the other but typically by using vision to find out what

the other is currently up to. The latter is reflected in gaze behavior, as both closely observe the manipulation action of each other. However, one would also expect that eye movements not only follow an action, but that they will often also precede them, for example when looking at a target location even before touching the action's current object.

While we understand a fair amount about how people coordinate during joint tasks, there's still much to uncover about the distinct ways Leaders and Followers use their gaze to predict each other's movements, especially when it comes to manipulating objects. How do these anticipatory gazes differ between roles, and how can we accurately measure these differences to better understand the underlying cognitive processes?

Research on non-cooperative or individual tasks has consistently shown that the gaze tends to shift toward the object to be manipulated before the actual action takes place (Land et al 1999). Concerning pick and place manipulation actions, it is known that humans fixate the object until it is lifted and afterwards they start fixating the placing location and stay fixated there until the object is released (Lavoie et al 2018). Patterns of repeated fixations on the same object—where the gaze returns to an object multiple times—have also been observed and have been closely associated with the cognitive processes underlying action planning (Sullivan et al. 2021, Pelz and Canose, 2001). Clearly, overall gaze patterns are complex, where gaze can also be occasionally directed towards action-irrelevant objects or objects can be reached for without a preceding fixation on them (Hayhoe et al. 2003). The latter might be associated with different seeing modalities, like e.g. it is known that grasps can be performed using peripheral vision alone (Brown et al 2005). It is also known that fixation can be made onto an empty space where an object had been previously, reflecting memory processes (O'regan, 1992, Foerster 2019).

Not only the actors' fixations were investigated, but also the ones from an observer. It is known that – similar to actors – observers make predictive looks towards the actor's manipulated objects (Flanagan and Johansson, 2003) and such predictive fixations are more common than long-term tracking of the hand of the actor (Flanagan et al, 2013). Furthermore, when an observer has previously performed the action themselves, they can make even earlier predictions (Möller et al, 2015).

In spite of the existing knowledge on gaze behavior of actors and observers in such scenarios, much less is known about gaze behavior during joint manipulation action. For example, there were quite some studies on gaze behavior in cooperative settings when referencing to things (e.g. saying "look at the cup", or by pointing there, Gergle and Clark, 2011, Andrist et al, 2015). As the current study does not use verbal communication, gestures or other referencing pointers, we will not discuss this literature any further.

Here we investigate joint table-top manipulation actions, more specifically hand-object interactions without verbal communication. Complexity of the gaze behavior remains high in such interaction because one has to visually plan and attend one's own action, while also observing what the other is doing. Existing studies of similar kind often rely on artificial settings, including sparsely distributed objects with exaggerated size (Huang et al, 2015), virtual reality with large screens (Andrist et al 2017, Fuchs and Belardinelli, 2021) or even specifically designed robotic setups for slowing-down human motion (Stolzenwald and Mayol-Cuevas, 2018). All these settings were made to allow for improved resolution of the eye fixations. Many times, the question addressed in those studies concerns intention

prediction of the other, frequently in a black-box style, targeted for applied purposes, but less strongly focused on theory.

From eye fixation studies of joint actions with more complex table-top tasks performed in realistic settings (different from the studies mentioned above) it is known that much less mutual gaze coordination happens (i.e., looking at the same place in the scene) in case participants do not perform verbal communication (Hessels et al, 2023). It was also observed that in such a cognitively demanding table-top manipulation task people hardly ever looked into each other faces (less than 0.5% of the time). A follow up study (Hessels at al, 2024) more deeply investigated relations between coupling of gaze with own action vs. the coupling of gaze with the other's action. The main finding reported is that the coupling to the own action is stronger. In the same study, also speech episodes or gestures had been investigated with similar results. However, looking at individual objects is not disentangled in both aforementioned studies and only a resolution of looking at different rather large areas on the table is achieved.

In our study the dynamics of Leader-Follower interactions is analyzed at a resolution of 2x9 grid locations filled with small objects and their combinations in reaching space. The study investigates predictive, planning-related gaze behavior in a two-player table-top game, focusing on the Leader-Follower dynamics, with the aim of enhancing both the ecological validity of the task and the precision of gaze measurements.

We track eyes, hand movements and the individual objects on the table and determine relations between eyes and hands of both players. Data was recorded using a setup where gaze tracking was combined with multicamera tracking of motion of the subjects and of the configuration on the table, as well as by employing touch sensors on the hands of the players, which allows accurate phasing of the cooperative manipulation based on the moments of touching and untouching of manipulated objects by the hand.

The game consisted of changing a configuration of objects on the table by pick and place actions by the Leader, where the Follower then needs to undo this change, albeit not directly by inversely-manipulating what the Leader has done, but by using alternative objects to establish the initial configuration.

The complexity of this game leads to several hypotheses about the potentially resulting complex behavior. The resulting cognitive load should be reflected in the behavior of both players. For example, are non-informative eye-movement being avoided or reduced? Will the natural asymmetry between Leader and Follower lead to different behavioral patterns, which reflects this, too?

Naturally, the Leader is much more forced to perform proactive planning (for action selection) than the Follower and we were asking to what degree predictive eye-movements exist for the Leader and how far they are ahead of their own action, not only for the directly following action – which is expected according to many reports in the literature – but maybe also for predicting what the other is doing or planning to do. Specifically, search patterns for an appropriate object as well as patterns of repeated fixations between the object to grasp and the location of where to put it could also emerge for the Leader as an expression of planning activity more often than for the Follower.

Different from this, the Follower needs to tightly pay attention to the actions of the Leader, where this is not required for the Leader. As a consequence, one would expect the Follower to perform

anticipatory eye movements as they estimate the potential action goals of the Leader, potentially even considering strongly the movement patterns of the Leader's hand to arrive at early predictions.

Finally, we ask whether the complexity of this game might be reflected in effects that potentially represent "post-hoc cogitation" about the performed actions.

Methods

Experimental Paradigm and Task

For this study, we designed a paradigm consisting in an Action-Counteraction (ACA) game. The two participants assume different roles during the ACA-game: Leader and Follower. The Leader possesses the freedom to execute tasks freely, while the Follower must act based on the previous action of the Leader.

At start, a set of glasses and cubes are placed on the table in two rows with nine locations each. The different allowed object combinations are shown in Figure 1A, and Figure 3A shows the setup on the table. Examples of action-counteraction pairs are shown in Figure 1B. The setup was made so that there is always a counteraction possible for any of the allowed actions that the Leader can make. To this end, we have placed two single cubes and two glasses at 4 locations and filled additional ten locations with the allowed object combinations. As a consequence, the four remaining locations remained free. One of the two players is assigned the role of the Leader, while the other player acts as the Follower. When the game starts, the Leader performs an action, where they can freely choose what to do with the constraint that their action needs to lead to any of the allowed configurations including the singletons. Upon completion of the Leader's action, the Follower must execute an action to counteract the action performed by the Leader. The counteraction must satisfy the rule that, after the pair of action and counteraction, the configurations of the objects on the table must be the same as before. The only additional rule was that the same objects were not allowed for the counteraction. For example, if the Leader chooses to invert a glass as their action, the Follower must invert another glass back, as illustrated in Figure 1B. Or – more complex – if the Leader un-stacks a glass that is placed on top of a cube and stacks it onto another glass, the Follower must un-stack another glass placed on top of another glass and then place the glass onto another cube in the same orientation (Figure 1C). After each action pair, Leader and Follower perform the next pair, repeating this for 10 rounds. Then the two players are told to swap roles and perform 10 more rounds. After 20 rounds the experimental session ends.

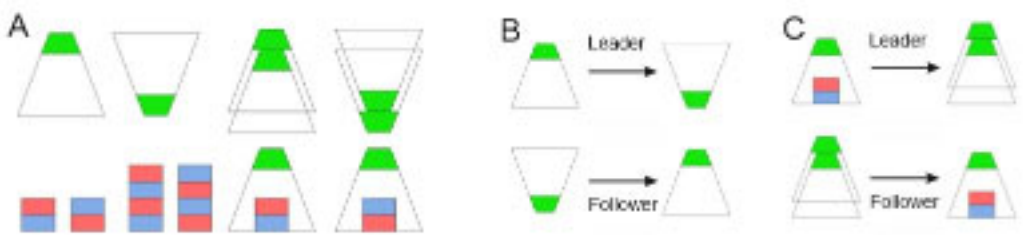

Figure 1) A) Objects used in the ACA-game and their allowed configurations. B) Simple example of an action-counteraction pair (glass inverting). C) More complex compound action example.

Subjects and Experimental procedure

We performed experiments with a total of 60 adult right-handed participants (39 male, 21 female, age 19-35) in pairs of two. Every participant had regular vision or was wearing contact lenses. All participants have been informed about the principles of this experiment and written informed consent has been obtained. Signed consent forms have been obtained from all participants and kept on record. The experiment is not harmful and no sensitive data had been recorded and experimental data has been treated anonymously and only the instructions explained below had been given to the participants. Participants were allowed to stop the experiment in case of discomfort or fatigue. The experiment was performed in accordance with the ethical standards laid down by the 1964 Declaration of Helsinki. We followed the relevant guidelines of the Germany Psychological Society according to which this experiment, given the conditions explained above, does not need explicit approval by an Ethics Committee (Document: 28.09.2004 DPG: "Revision der auf die Forschung bezogenen ethischen Richtlinien").

In detail: Participants were invited to sit at opposite places at a round table (see Figure 3A). Every session started with the experimenter introducing the game rules to the participants, allowing them to practice and become familiarized with the game. Once both participants were ready, they put on white gloves for better recognition of the hand by the computer vision system. In addition, gloves contained a microswitch under the index finger to accurately record touching events. Eye-trackers (Pupil Core eye trackers from Pupil Labs) were attached to the participants similar to wearing a regular eyeglass. This was followed by a calibration procedure where participants were asked to look at four different locations at the table. After that the actual recording of the session began. We recorded touching and untouching events using the touch sensors (switches in the gloves), hand movement patterns using a multi-camera system (see below) as well as eye-movements of both participants simultaneously. During each session, participants engaged in playing the game as described above. Every recording session consisted of four blocks and every block consisted of 20 pairs of action/counteractions where each of the players assumes both roles, Leader for 10 pairs of actions and Follower for the next 10 pairs. Once all four game blocks were finished, the session ended. Every block lasted about 3 to 5 minutes leading to a total time of less than 20min for every session. This leads to 120 data sets (30 participant pairs times 4 blocks) where each set contains the recordings of one block. From this, we extracted 110 valid data sets. Ten data sets were excluded due to recording failures.

Recording Setup

Overview

As depicted in Figure 2, the system consists of three input data sources, namely the five-view camera setup from FLIR (Teledyne FLIR LLC) and the two Pupil Core eye trackers from Pupil Labs (Kassner et al, 2014) and the touch sensors. They are synchronized using a universal clock, and jointly calibrated using Aruco makers (Apritag 36h11 family, Garrido-Jurado et al, 2014). The touch sensor was designed to capture the touching/untouching events between hands and objects (T/U events).

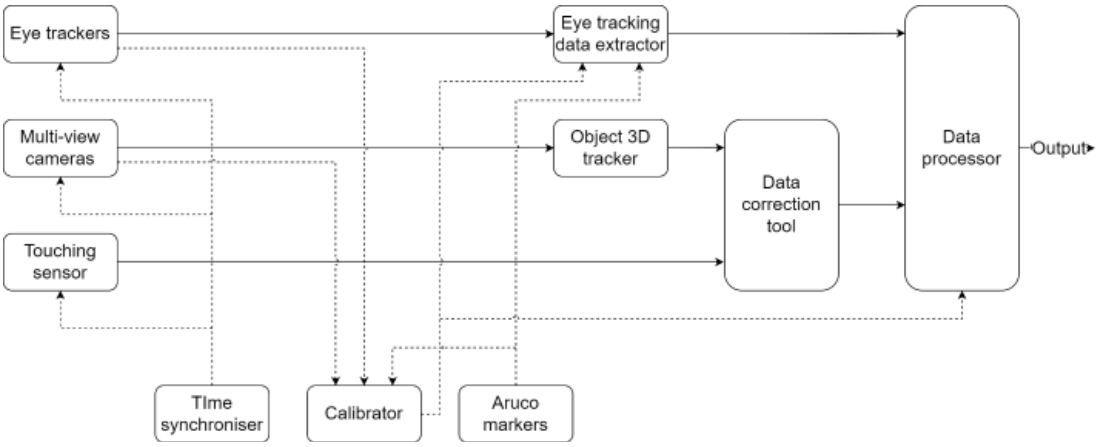

Figure 2) Block diagram of the experiment's technical setup.

The timing and data flow of the system are delineated as follows. Two calibration procedures are needed. First, a general calibration procedure is executed. During this, the experimenter generates a 3D representation of the tablecloth filled with Aruco markers utilizing the Pupil Labs eye tracker and associated recording software. Subsequently, a series of images of the same tablecloth is captured by the five-view camera setup, allowing for the generation of a corresponding 3D model. The calibration software is then utilized to compute the transformation matrix between the five-view camera configuration and the eye trackers. This calibration procedure only needs to be done once unless there are alterations in the physical positioning or orientation of table or cameras. Second, for each experiment, the experimenter initially performs calibration of the eye trackers for both participants as described above and, following this, the experimenter triggers the start of the recording. The incoming video stream from the five-view camera setup is subsequently encoded and written to the hard drive in real-time, alongside the data acquired from the eye trackers and the touch sensors. The recorded data is analyzed off-line. During this, an object tracker identifies and tracks the objects and hands, and it also extracts the head direction present in each video frame. Finally, the data of interest are extracted and subjected to analysis.

Hand tracking

The hands of the participants are tracked employing an Axis Aligned Bounding Box (AABB) pipeline. Using the recorded camera images of all five cameras, a custom trained detection Deep Neuron Network (DNN) - YoloV5 model (Jocher et al., 2020) detects the hands in the images and outputs 2D bounding boxes. Subsequently, a triangulation algorithm transforms these 2D bounding boxes into 3D AABBs. Next, the 3D AABBs are tracked using a modified Unscented Kalman Filter (UKF) and finally smoothed by a low pass Finite Impulse Response (FIR) filter. See Figure 3A for a view onto the scenery that includes a 3D AABB on one of the objects.

Object detection

Object detection and label assignment uses the same YoloV5 model. The labels of the different objects are being associated to the different possible object positions (see below) after the end of a manipulation action.

Eye-tracking data extraction

In this study, the eye tracking data of each player is defined as an array of eye fixations over time. Each array-entry contains: the starting time of the fixation; the duration of the fixation; the coordinates of

the target position or the instance ID of the target hand the participant had looked at. The details of this definition and the algorithm to calculate eye fixation data are stated in the following paragraphs.

Each eye tracker has three mounted cameras, one scene camera and two eye cameras. Using the recording software provided by Pupil Labs, the eye tracker outputs the gaze in the form of 2D coordinates on the scene camera, and videos from all the cameras. Two types of data are extracted offline after recordings using the software provided by Pupil Labs. One type is the fixation data, which is defined as sets of consecutive frames where the gaze remains at the same point for a sufficiently long period. To determine a fixation, a threshold of +/-2.45 degrees of visual angle had been set within which gazes were attributed to the same object. Furthermore, a resolution of 20ms for gaze shifts was chosen such that the initial sampling rate of the eye-movement data was set to 50Hz. A fixation is given when minimally 10 such samples hit the same object, hence we assume as minimal fixation duration 200ms (Johansson et al, 2001, Pannasch et al, 2008). Sometimes occurring glitches, where a sample from the eye-tracker had jumped off the actual target object, were filtered.

The second type of data extracted by the Pupil Player software is the head pose data. By detecting the Aruco markers on the table and employing the Perspective-n-Point (PnP) algorithm, one can determine the extrinsic parameters of the eye-tracker's scene camera. With the head pose data, the 2D gaze fixation was further processed into 3D, which is essential for integrating eye tracking data with location of the objects.

Determining hand and object location

To calculate which hand or object location the eye fixations hit, a ray tracing algorithm based on the principles outlined in (Williams, at al. 2005) is implemented in this study (see Appendix). Ray tracing is essentially a collision detection algorithm between a casted virtual ray originating at the eye and extended towards the object. By this the looked-at object is determined.

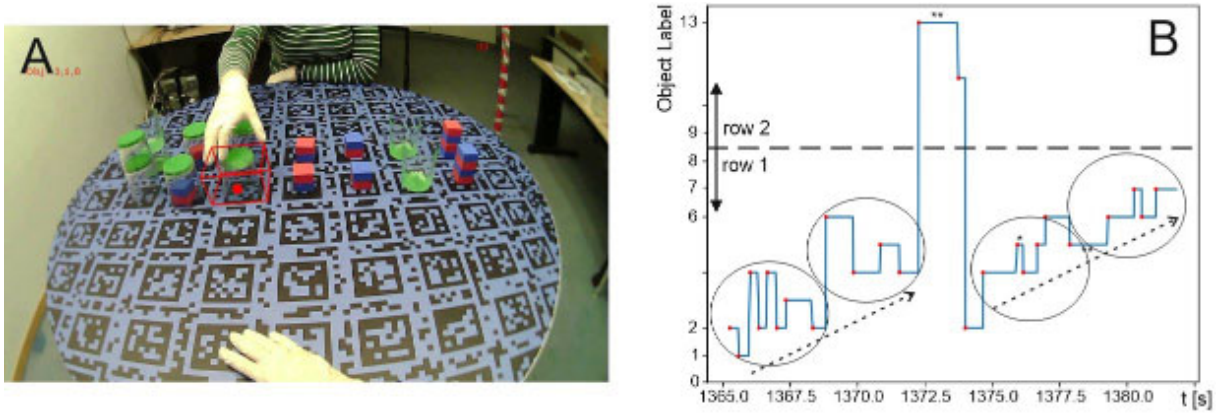

Figure 3. A) View as seen by the camera of the eye-tracker with object bounding box and gaze-indicator (red dot) of the person in front included. This view is essentially identical to the way participants see the scenery. B) Eye track over time. Start of fixations are marked by red dots.

To conveniently describe the positions of the objects as well as empty target positions, a discretized coordinate system of the table and a set of 18 corresponding virtual 3D AABBs were defined. This is possible, because no other object locations were allowed in this game. In Figure 3A the defined positions are based on the two rows of Aruco makers on the table. The top left corner of the marker corresponds to the defined location with a coordinate of (0,0), and the bottom right corner corresponds to (8,1). Notably, z-coordinates do not have to be used in this case, because they are not required for the here-performed analyzes. There are two reasons that virtual 3D AABBs are used in

this study. First, they are used to detect whether the participants look at empty positions on the table. Second, since the objects are relatively small in size, using the virtual 3D AABBs can more precisely determine which object the participants look at, as the objects can also only be put on these defined discretized positions. Note that object labels can be attached to the different locations following object detection as described above.

Figure 3B shows an exemplary eye-movement track from one of the sessions showing how the basic data is structured, which is used for all statistical analyzes. As mentioned, objects are arranged in rows 1 and 2 and numbered from 0 to 8 (left to right) in row 1 and from 9 to 17 in row 2, where this diagram is cut above 13 because no saccades have occurred to objects 14-17. For this participant row 1 was the one directly in front and row 2 was further away. The track shown here contains in total 25 saccades to different objects over a time of 17.5s. Minimal fixation duration was 220ms (marked by * near 1375.0 s in the figure) and maximal duration 1500ms (** at and after 1372.5 s). Clearly visible are two progressive sequences of saccades along row 1 (dashed arrows) interrupted by looking at row 2 for a short time. During these progressions alternating saccades to neighboring (or close-by) objects, highlighted by the ellipses, are found.

Data and Data Analysis:

As mentioned above, in total we obtained 110 valid data sets. The raw eye- and hand-tracking were analyzed using several methods. For hands we determined two data points: 1) the start of the hand movement, determined from the camera images as the moment when the hand leaves the "home" position for more than 5cm and 2) the moments when the hand touches (T-event) or releases an object (U-event) as well as the durations in-between, determined from the sensor data.

Chunking: This leads to a natural chunking of the data stream. In Figure 4, which is a schematic of how we will represent timed-data below, we show these intervals marked with green and orange bars. Note that, naturally, T-U-intervals alternate between Leader and Follower and there are intervals in-between, called object-static intervals, which represent those periods where no object is being moved (hands can move, though!). For example, note that the period where the Leader (or the Follower) does not move an object stretches from the end of interval 2 to the beginning of 6 (Follower: end of 4 to beginning of 8) always bridging 3 intervals. The green and orange arrows indicate this "bridge". Naturally, this repeats again for other groups of intervals in the same manner (e.g., 1 3, 7 9 and more).

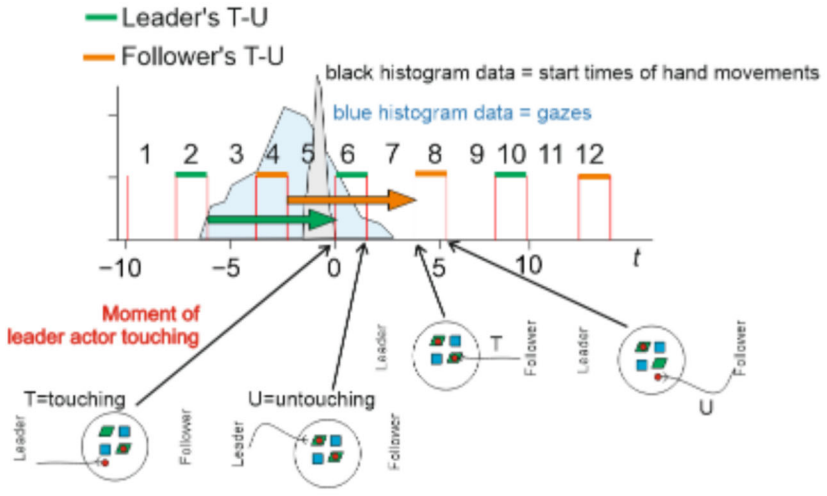

Figure 4) Schematic of the way temporal data will be presented.

A total of 12 intervals will be plotted in all Figures that show histograms. This number was chosen to include in total 3 T-U intervals each for Leader as well as Follower plus one more object-static-interval in front to create a full 3-interval bridge (for the Follower) at the start of each diagram. In the following we name the temporal duration from 1-12 an "episode".

The pictograms at the bottom show the course of action schematically, where the Leader has put a red disk onto a green parallelogram and the Follower subsequently performs the respective undo operation elsewhere on the table.

We use such a long stretch of an episode lasting from interval 1 to 12 to allow for an extended analysis of the viewing behavior but here care has to be taken that during an episode no object had been manipulated twice[1]. Episodes where this happened had to be excluded from the analysis, because, as soon as the same object, let us say O1, is manipulated twice in one episode, it is not clear to which of the two manipulations the gazes onto O1 shall be attributed. As a consequence, the number of episodes in different histograms is not the same. To account for this, and to make the data comparable, we normalized all histograms to 1000 episodes.

Histogram data representation and its time-origin: In addition to the chunking 1-12, Figure 4 shows also other aspects. Here zero of the time-axis refers to the moment of the Leader's touch of the to-be-manipulated object (start of interval 6). This centering is done, because most gazes were expected (and found as shown below) before this moment in time. Below this will be plotted as histograms and the blue schematic shown here is characteristic of this fact. Numerical values at the abscissa represent, accordingly, the average times before or after in seconds. The same zero-centering has also been used for the Follower and, naturally, there time-point zero refers to the start of interval 8 (see Figure 7 and Figure 8). In black we encode histograms for the starting moments of the hand movements (see panels A and C in Figure 7 and panels A and E in Figure 8).

Temporal Normalization: Given that the duration of the intervals is not constant and varies from episode to episode and also between different players, we normalized all interval durations to their respective averages found to be for Leader as well as Follower 1.5 s for T-U and 2.3 s for object-static intervals, and we performed time-warping of all the temporal gaze data within each interval. In a free-running (no triggers, no time-limits) experiment such as the presented one, sometimes players were inattentive or distracted or for some other reasons overly long interval durations would sometimes happen. Those were removed (above 2.9s for T-U and above 5.6s for inter-activity intervals). We found a total of about 5% of too long intervals. This temporal normalization applies to all results (e.g. histograms) that show timed data.

Results

All analyses used as minimal fixation duration 200ms, which is realistic for minimal fixation duration in such settings (Johansson et al, 2001, Pannasch et al, 2008). Note that we annotate with "object" the place on the game-grid (!) where an object stands that will now be manipulated by the actor. We use the same annotation "object" also for game-grid places where the manipulated object has been moved away from a moment ago. With "location" we annotate the place where an object will be (or has been) put down. The latter includes all possible places on the game-grid (not only empty ones).

[1] Note: only direct repetitions were forbidden by the rules of the game, but a repetition by manipulating an object in interval 2 and then again at 10 would be permitted.

Central versus Peripheral Viewing

For this we analyzed intervals 3 to 6 for the Leader and 5 to 8 for the Follower as numbered in Figure 4, where – as can be quickly seen from the histograms below (Figure 7, Figure 8, main blue histogram peaks) – most gazes happen.

In these intervals and for all experiments we count all incidences where actors (or observers) looked at least once at the object which is being manipulated or the location where it is being put down. The here-plotted bars (Figure 5) show the proportion of those gazes over the count of all performed manipulations. Thus, they take values between 0 and 1. For the number of gazes, we use a 95% confidence interval calculated for proportions, where the Clopper-Pearson (exact) method for the Binomial distribution was used. Confidence intervals are for all bars small (in the range 0.01 to 0.04, left out in the plot as they are too tiny) and are, thus, omitted in the figure. All below-discussed differences are significant.

Here we plot with columns "n" (narrow view field, right columns) only the fraction of gazes that had been directly at the object (or location), with columns "w" (wide view field, left columns) we include gazes at the immediate neighbors of the targeted object (or location). This is done to see to what degree in some cases participants were using peripheral (peri-foveal) vision. Naturally columns "w" are larger than the corresponding columns "n" because "n" is a subset of "w".

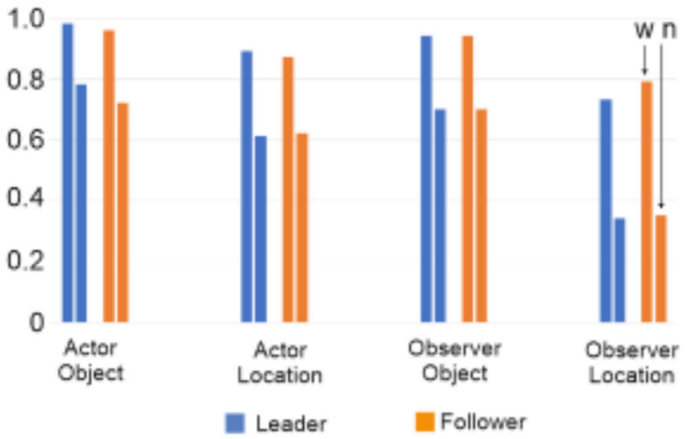

Figure 5) Central versus Peripheral Viewing. Annotations are, for example, "Actor Object" meaning that the actor directs its gaze at the object they will manipulate and similarly for "Actor Location" for the places to put down the object. Accordingly, being the observer, the analyzed gazes are those that are directed at the object/location that the actor chooses. Direct looks are given by "n" (narrow) and indirect ones by "w" (wide) as defined in the text.

The fact that bars "w" for the actor as well as observer-object are close to 1.0 demonstrates that computer vision analysis is highly reliable where a drop out of less than 10% is acceptable. The larger drop for "Observer-Location" will be explained later (see Discussion).

Notably all obtained numbers for "n" (narrow view field) are below 1.0, which means that peri-foveal vision has been used quite often. The viewing angle difference between a direct gaze and a nearby gaze is 6-8 degrees and, thus, relatively small[2]. Hence, participants can use peri-foveal vision, which is accurate enough to target the object/location.

In general, the obtained numbers for Leader and Follower are quite similar (blue versus orange), but several additional specific aspects can be seen. Gazing at objects happens always more often than gazing at locations for the observer (right half of the diagram). For the actor (left half) this is not seen and the very small drop does not reach significance. For the observer, however, this drop is quite pronounced.

[2] Objects are arranged linearly. This leads to the fact that the angular difference for two objects away from the center is larger than for center objects.

The above discussed observations are consistent for narrow as well as wide view fields ("n" versus "w"). To allow for highest accuracy for the following analyzes we will, thus, use only data from "n".

Single versus multi-gaze

In a similar way, Figure 6 demonstrates how often gazes onto the same entities are repeated from one (no repetition) up to five repetitions analyzed for intervals 3 to 6 for the Leader and for the corresponding intervals 5 to 8 for the Follower. Confidence intervals for proportions (95%) were calculated in the same way as for Figure 5. The overall pattern looks similar to that in Figure 5 and, note, that by summing up individual groups from 1 to 5 you will indeed receive this figure. As expected, higher numbers of repeated gazes were less common than lower ones. For gazes at the object, the Leader produces many more multi-views with 3 or more gazes than the Follower. This is significant for 3 repetitions for the Leader being actor and for 3 to 5 repetitions for the Leader being observer. Reversely for the Follower being observer of the object (the Leader will handle), single gazes dominate. No other differences are found.

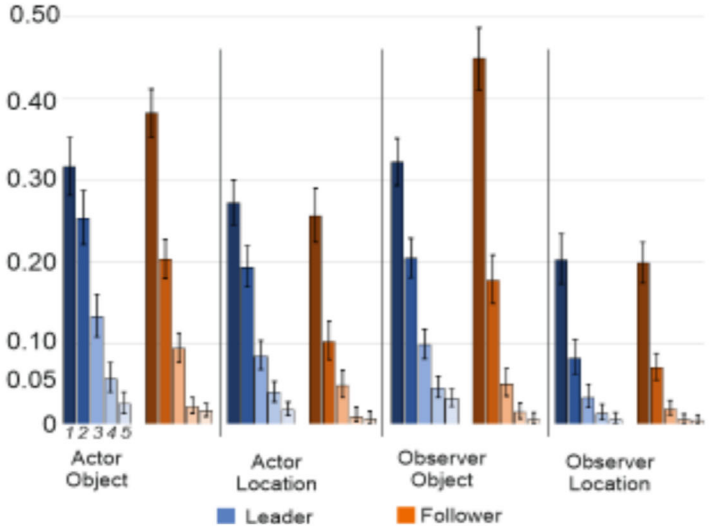

Figure 6) Single versus multiple gazes, Bars 1 to 5 represents how often a certain look has happened.

Temporal characteristics of the viewing behavior

Figure 7 shows the viewing behavior as blue histograms as well as when the hand starts to move (panels A, C, black histogram). Histograms are centered ("zero") at the moment where the Leader (A,B) or the Follower (C,D) touched the object, which is to be manipulated. This is done, because most gazes were expected (and found) before this moment as shown by the main blue peaks. Statistical evaluations for this are provided below (Figure 8). The dashed vertical line allows for visually aligning top and bottom panels.

When the Leader is the actor (A), they start their hand movements during interval 5 (black histogram) and the hand then eventually touches the object. The blue histogram shows the distribution when the Leader gazes at the object he is manipulating at this stage. The bulk of this happens before interval 6. Hence, as expected, eye movements predict hand movements. In addition, there is a tail that covers intervals 7-9, which will be discussed later. Panel (B) shows a baseline. It contains all gazes of the participants at objects not having been manipulated during the length of the episodes contained in panel (A). Below we will use the baseline to calculate statistical significances.

Panel (C) shows the corresponding situation when the Follower is the actor and baseline given in (D).

Hence, in summary, both histograms A and C represent the gazing behavior at the to-be-manipulated objects showing part of the sequence of do-undo: It shows the object-centered viewing behavior of the Leader who prepares and performs an action and the (also object-centered) viewing behavior of the Follower who then acts, too, to perform the undo action.

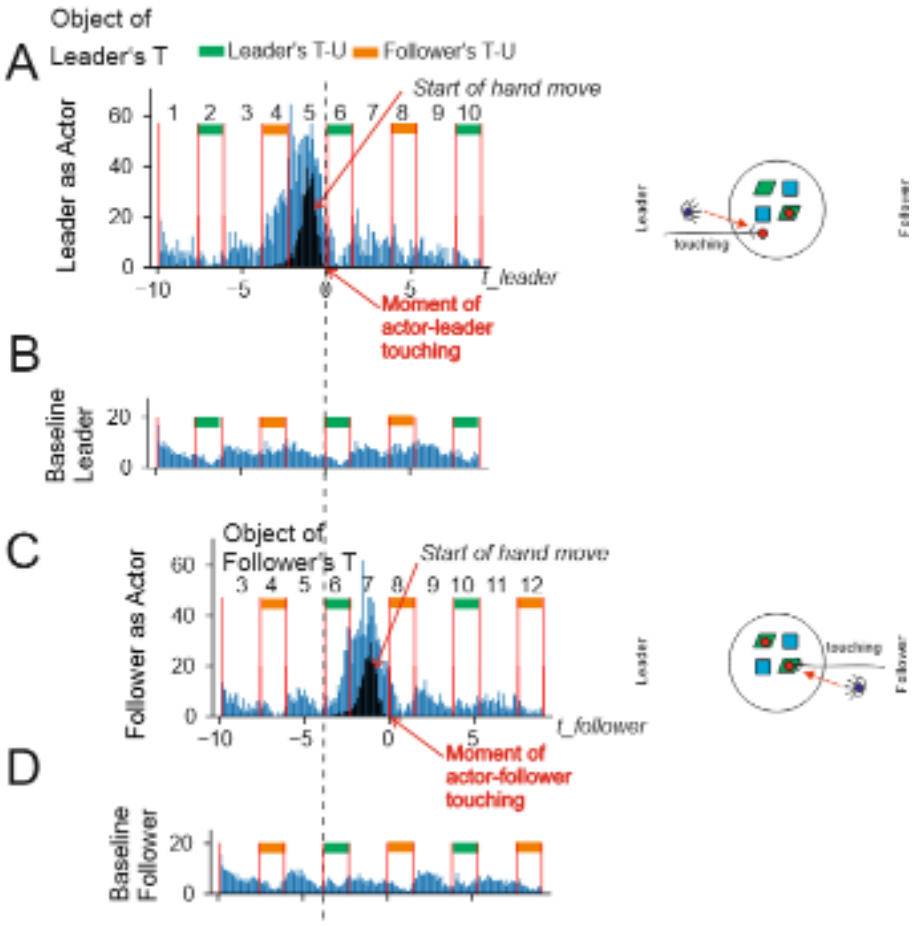

Figure 7) Histograms for two situaƟons and belonging base lines. Here and in the following: when a diagram is labeled "Object" (panels A,C) it represents the gazes at that parƟcular object that is being manipulated during this episode and similarly for "LocaƟon" (see next Figure) of where that object is being put down. Histograms are calculated to represent the number of fixaƟons onto object or locaƟon over all parƟcipants and all objects/locaƟons, where the ordinate is normalized by the number of episodes analyzed (see Method secƟon). On the right side again, a schemaƟc is shown how the configuraƟon of the table looks like at the given moment in Ɵme and also what happens for Leader and Follower (touching an object, untouching the object aÕer placing it at a new locaƟon). Hence, panel (A) leÕ side belongs to the situaƟon where the Leader touches an object (pictogram of the small red disk) and the histogram(s) show here the start of the Leader's hand movement (black) and – as pictogrammaƟcally indicated by the eye – the staƟsƟcs of the Leader's gazes at this object (blue). Diagrams are labeled by different acƟon- as well as inacƟvity-intervals (numbers and orange, green markers).

In Figure 8, we will now discuss these and more cases in more detail also showing which intervals display a significant deviation from baseline and which not. We have, thus, subtracted the baselines from their corresponding original plots, which can lead to negative numbers, too. Red asterisks represent intervals which are significantly greater than baseline level ($p<0.01$ by a one-sided Wilcoxson signed rank test).

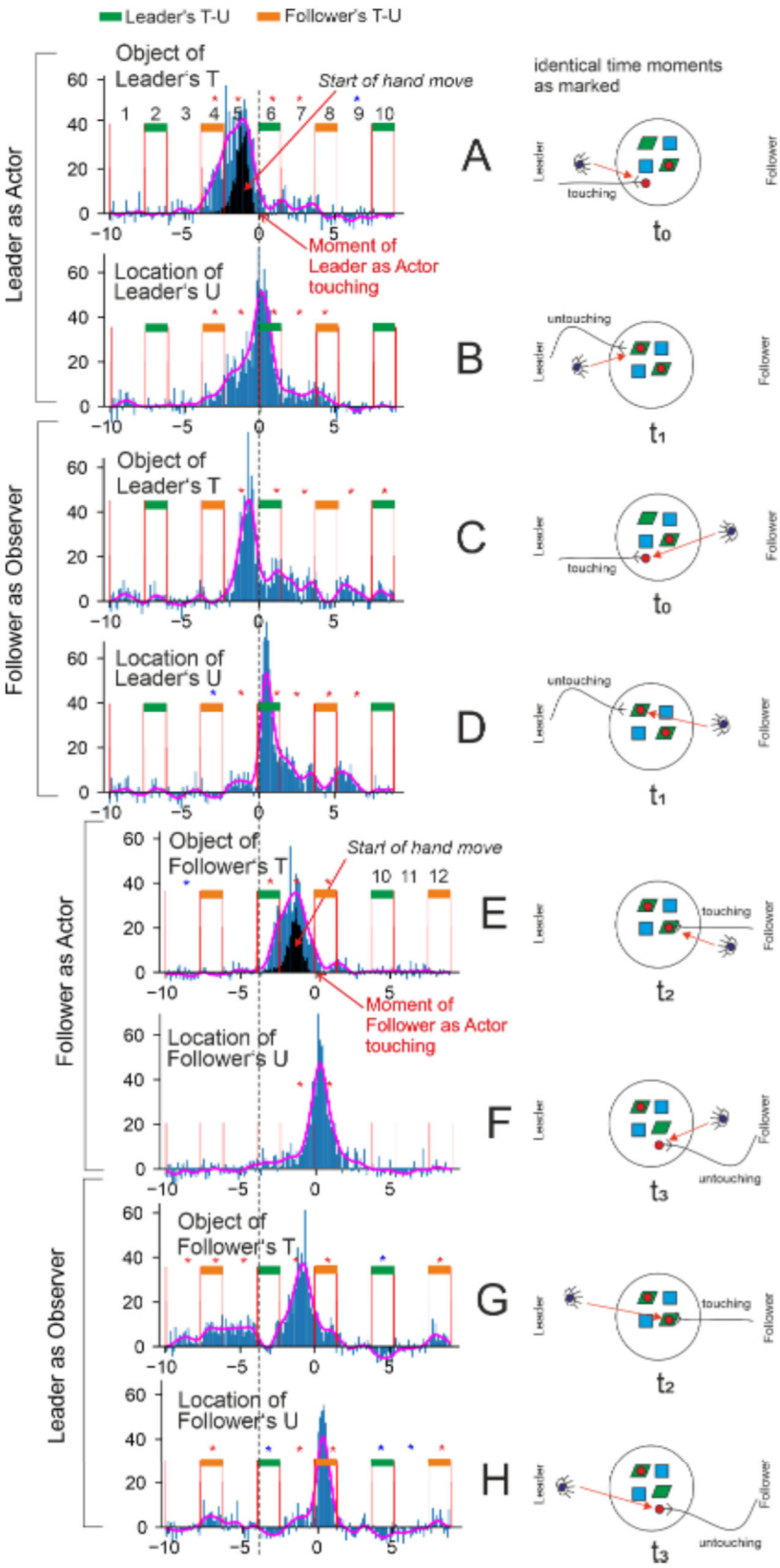

Figure 8) Detailed temporal analysis of hand and gaze paΣerns. A-H) Actor-Observer, Leader-Follower, and Object-LocaΘon combinaΘons as indicated.

Pink curves represent a low-pass filtered version of the blue histograms, using a Gaussian filter with STD=0.3s. Note that diagrams come in pairs which happen at the same time (t_0, t_1, t_2,t_3, indicated under the pictograms). The vertical dashed line marks the start of interval 6, which is the one where the Leader actually starts their action.

All four top diagrams (A-D) show a tail behind the main peak. These diagrams are all about the Leader's object (taking and placing). These tails represent looks "into the past", either at the place where the object had been before it was manipulated and/or at the location which had been covered by the object a while ago.

We will now provide results in more detail for the different figure panels in Figure 8.

Panel (A): This panel shows the behavior of the Leader when they consider the to-be-manipulated object. It also shows the start of the Leader's hand movement, which happens in interval 5 before the touching of the object. The Leader's gaze shifts to the object they are touching just before interval 4 (but mostly in interval 4 and 5). Meanwhile, in interval 4, the Follower continues their task, undoing the previous game event – an action that does not require the Leader's attention. Intervals 6 and 7 show a tail. Intriguingly, at/after their untouch, some Leaders look again at the location where the object had been taken away from.

Panel (B): This panel shows the behavior of the Leader when they consider the location of where to put the object. The Leader looks at the location very often before they have even touched their object (earliest start of looks at end of interval 3). The whole plot is shifted by one interval relative to panel A, which is expected, because targeting a location must come after targeting an object. Also here exists a tail lasting until interval 8.

Panel (C): This panel depicts the behavior of the Follower who observes in this case the object that the Leader will take. Naturally, the Follower looks at the Leader's object only when or after the Leader actually started to move the hand (compare peak in (C) to the black histogram in (A)). There is only a quite small delay of the peak in (C) relative to the black peak in (A) of approx. 400ms. Again, there is a tail of looks at the location of where the Leader's object has been placed.

Panel (D): This panel depicts the behavior of the Follower who observes in this case the location that the Leader uses to put down their object. This is an observation which rather strictly follows interval 6 – the interval where the Leader acts – with not much of a predictive component. The bulk of the Follower's gazes happens only mildly before the untouch by the Leader and some gazes occur also afterwards. Again, there is a clear longer-lasting after-look tail existing.

Intriguingly, for diagrams E-H (Follower as actor and Leader as observer) there is no tail after the main peak. Note that at time t3 the initial situation on the table has been recovered (compare panel (A) with (F).

Panel (E): This panel shows the behavior of the Follower when they consider the object to be used for the undo action. This diagram looks like a shifted copy of (A) with very much the same characteristics but without the tail. This general shape is expected as here the Follower prepares for touching their object. Remarkably, when cutting out the smoothed curve of (E) and pasting it over panel (A) one can show an almost perfect fit to the main peak in (A).

Panel (F): This panel shows the behavior of the Follower when they consider the location of where to put the object. The same observation as for (E) also hold for (F): It's a shifted copy of (B) without tail. Also here, the smoothed curve matches the one from above exceedingly well.

Panel (G): This panel depicts the behavior of the Leader who observes in this case the object that the Follower will take. The Leader produces a little bit wider peak when looking at the object of the Follower as compared to the Follower in (C) who looks at the object of the Leader. The main peak happens before the object is touched by the Follower. By the alignment with the hand movement plot in (E) it can be seen that the Leader uses the hand movement of the Follower to predict which object the Follower will touch. This is again similar to the observation in (C). The main peak in (G) is, however, slightly wider in front of its maximum than the peak in (C) because in (C) the Follower has no information about the next action of the Leader and cannot predict early. They can only use the hand-movement of the Leader for prediction. Different from this, the Leader knows that the Follower has only a limited number of options, which object to take and this sometimes allows for an earlier guess. Furthermore, in intervals 4 and 5 there is a small, but significant peak indicating that – while the Leader plans their own action (and at the same time observing what the Follower is currently doing) – they also look at potential objects the Follower could use in the future to undo their currently planned actions.

Panel (H): This panel show the behavior of the Leader who observes in this case the location that the Follower uses to put down their object. The histogram is narrow and, quite similar to (D), it starts mildly

in a predictive manner, alongside the movement of the Follower (interval 6, orange) and ends with the Follower releasing the object. A narrow histogram is expected, because, similar to (D), also here this is mostly a reactive (non-predictive) observation, thus, this happens rather sharply alongside the movement interval (orange) of the Follower.

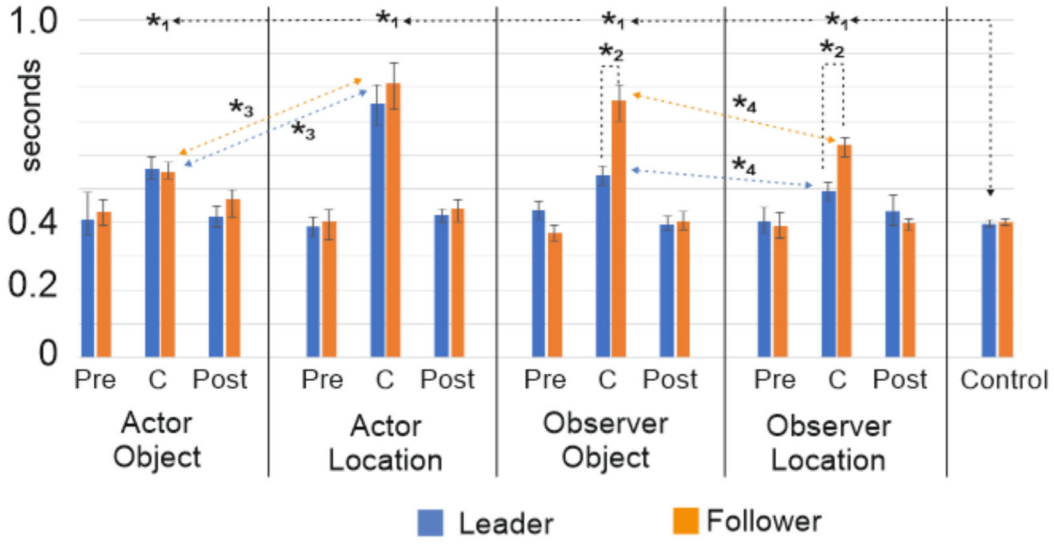

Figure 9) Viewing duraƟon before (Pre), at (C), and aŌer (Post) the main peak in the histograms. For this the two intervals centered at the peak define Cand two each before and aŌer define Pre and Post, respecƟvely.

Viewing Durations

Figure 9 shows the viewing durations for actor and observer in their roles as Leader or Follower. Error bars represent the 95% confidence intervals of the median calculated with the Hettmasperger-Sheather method with alpha=0.05. Significant effects are marked with an asterisk and here the strongest effect is that gazes in columns “C”, that belong to the main peak in the histograms in Figure 8, are significantly longer (*1) than gazes at random objects (Control) or gazes that were taken from outside the histogram peak. This holds for Leader as well as Follower. Furthermore, the Follower as observer looks longer at object as well as location than the Leader does as observer (*2). When being the actor, both Leader and Follower consider the location longer than the object, which they will manipulate (*3). The opposite is found when being an observer. Here the object that is being observed appears to be of more interest than the location (*4), where this is just above significance level for the Leader (blue) and much more pronounced for the Follower (orange).

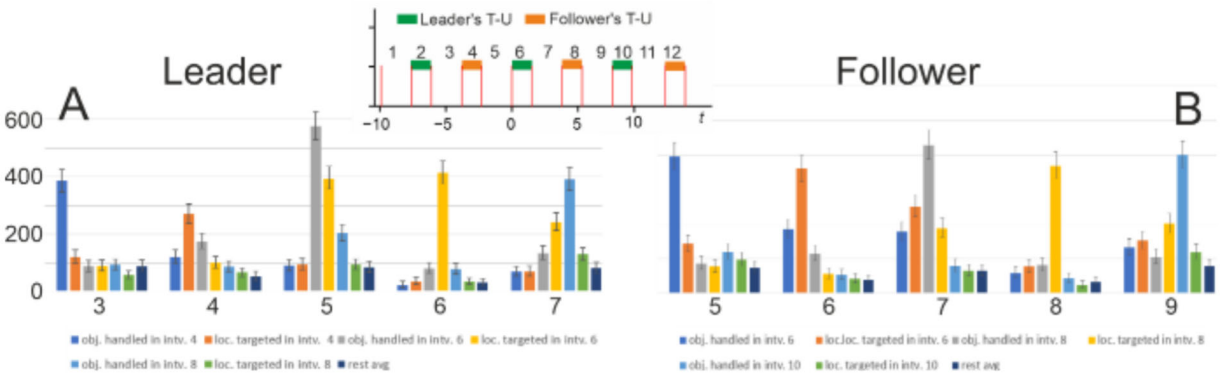

Figure 10) Objects and locaƟons per interval. Intervals (abscissa) are numbered as shown in the inset above (see Figure 4).

Objects and locations per interval

In Figure 10 we analyze at which objects/locations Leader and Follower have looked in the different intervals. A total of 500 intervals has been analyzed for each block. Histograms show the number of gazes per 500 intervals. Error bars represent 95% confidence intervals calculated for Poisson distributions.

The Leader has finished their action in interval 2 and begins considering their next action but before they look quite often at what the Follower does. Hence, gazes at the Follower's object (blue bar, left) are common in interval 3 (compare also to Figure 8G). Note that these gazes occur before the Follower has actually touched their object, which happens only at the start of 4. During this time the Leader shifts gaze to the targeted location of the Follower (orange bar). However, the Leader now also begins to consider, which object to manipulate next (gray bar). As expected, this becomes prevalent in interval 5, but here the Leader also starts to search for the placing location (yellow bar). During execution of the Leader's action (intv. 6) the search for the location is dominant. Note that interval 7 is equivalent to 3 but not identical. This is due to the fact that intervals follow each other to define an episode and the later coming intervals are conditional on the earlier ones, leading to similar but not identical results for equivalent intervals. Hence for 7 the light blue bar in 7 corresponds to the dark blue one in 3 (same for green and orange). Indeed, in interval 7 the looks at the next object of the Follower again begin to dominate (as in 3), but here the Leader still keeps gazing at their chosen location (yellow), too.

For the Follower a quite similar pattern emerges. However, in interval 6 and 7 the dark blue and orange bar are larger for the Follower than the equivalent counterparts for the Leader (which are found in intervals 4 and 5). For both of them, these two bars represent "what the other does" and, evidently, this is more relevant for the Follower than for the Leader.

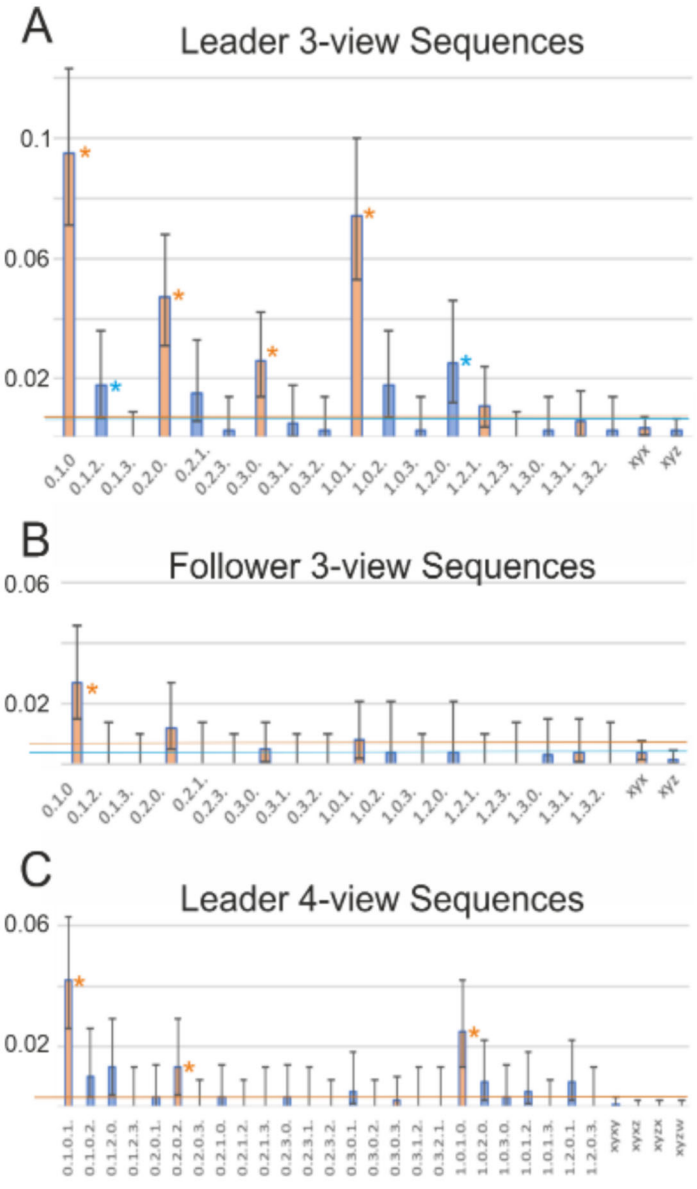

Figure 11) Sequences, A,B) 3-view sequences, C) 4-view sequences only for the Leader. Numerals mean: 0=own object, 1=own location, 2=the other's object, 3=the other's location.

Sequences

Figure 10 shows that in many cases there are multiple looks at an object (or location). We ask now: Does this entail specific sequences? This is of particular interest when considering possible

decision processes that follow after a participant has looked at that object which they will indeed a moment later take – encoded at the abscissa in Figure 11 by a leading “0”, or that location where they will actually place it – encoded by “1”. Hence, to address this, we analyzed potentially relevant 3-view and 4-view sequences starting with 0 or 1. For example, a sequence is coded here at the abscissa as “0.1.0” and the belonging bar shows the fraction of gazes by the participant where they have looked at own object (“0”) followed by a look at own location (“1”) and then again by a look at own object (“0”). Indices “2” and “3” refer to the objects manipulated in the next step by the other. Figure 11 plots the occurrence frequencies of several such sequences, where the last columns show the respective baselines by computing sequencies of gazes at objects not involved in the manipulation at the given episode (e.g. the bars marked with “xyx” and “xyz” for the 3-view sequences in panels A and B). Error bars represent binomial confidence intervals, computed using the Clopper-Pearson (exact) method. The horizontal lines, colored according to the relevant sequence structure, transfer the confidence intervals of the baseline across the diagram. Hence, a result is significant for a bar with its lower confidence interval which does not cross its belonging line. Leaders (panels A,C) do indeed perform several sequences above chance expectation. Here all combinations of 010, 101 as well as 1010 and 0101 are significantly over-represented. (As a sidenote, evidently these 3-view sequences are contained in the corresponding 4-view sequences). Here also 020 and 0202 are prevalent. Hence, Leaders perform alternating gazes at own object and location (0,1) but also at own object and the other’s object (0,2). In addition to this, there is also a small over-representation of 030, where 3 stand for the other’s location, but 0303 does not happen often. Mixed combinations of three different entities (x,y,z) looked at in various 3-view sequences are only found for 0=own object, 1=own location and 2=the other’s object. These are 012, 021, 102, and 120, two of which (012 and 120) are just above chance and the two others just below. This show at least a certain trend that Leaders consider those three entities also often in a 3-view sequence. In summary, Leaders indeed – and as expected – frequently alternate between considering their own object and location. However, though less often, they also check their own object or location again, as well as also checking the potential object and location of the other.

For the Follower (B) the first four orange bars look like a compressed version of the corresponding bar of the Leader (A), but only the 3-view sequence of 010 was above chance. No significant 4-view sequences were found for the Follower at all and this plot is omitted.

Discussion

In the current study we have investigated joint table-top manipulation actions considering hand-object interactions, without verbal communication. Unlike many previous studies (Huang et al, 2015, Andrist et al, 2017, Fuchs and Belardinelli, 2021, Stolzenwald and Mayol-Cuevas, 2018), which focus on scenarios with discrete external action cues, we explored cognitive and behavioral processes as they unfold in a time-continuous setting without external prompting. This setup imposes a relatively high cognitive load, requiring sustained attention, visual planning, and self-guided coordination. The need to plan and monitor one’s own actions while simultaneously tracking those of another gives rise to intricate eye-movement patterns, reflecting a rich and diverse range of cognitive processes. Several relevant findings have already been reported in the literature concerning human-human non-verbal interaction, for example that during interaction, people hardly ever look into each other’s faces and that gaze is strongly coupled to the to-be-manipulated object (Hessels et al, 2024). The same we

observed here, but – in addition – our scenario strongly encouraged participants to consider the other's actions. We raised a set of hypotheses which we could test with this setup.

As expected, actors and observers showed a different gaze behavior. When being an observer there is a strong difference concerning looking at objects or locations: Gazing at objects happens always more often than gazing at locations (Figure 5). Once you decided on an object, there are, on the one hand, less choices for locations, which might underly this effect. In addition, on the other hand, observers do not really have to consider location much at all. Only knowledge about "what needs to be un-done" is required (for the Follower) and locations where the other put the object mostly does not matter. Above, we had speculated that the high cognitive load of this experiment should be reflected in the behavior of the players and that non-informative eye-movement are being avoided or reduced. The above-mentioned observations indeed support this. Observers move their eyes in a conservative way, as has been found in previous studies (Itti & Koch, 2001, Wurtz et al, 2010. Oliva et al, 2003). That actors do not do the same, was the first indication that there is an asymmetry between actor and observer.

Asymmetries were also found for Leader versus Follower, as revealed by different behavioral patterns. Thus, Leaders plan their actions well ahead of time (see Figure 8A) which was not the case for Followers who, as such, of course had to wait for the action of the Leader. Our finding confirms that the Leader engages in more proactive planning for action selection than the Follower.

The fact that Leaders perform sequences of gazes at object-location-object, etc. (Figure 11) points in the same direction of a planning-related search pattern. Such look-ahead fixations were reported before (Sullivan et al. 2021, Mennie et al, 2007, Pelz and Canose, 2001), but sequencies of repeated looking back and forth between objects and locations had not yet been considered so far.

Hence, the above results showed that predictive planning-related eye movements manifest for the Leader and in addition, we had asked, whether this happens – as expected from the literature – only for the immediate next action, or whether even more pronounced proactive planning exists. As seen in Figure 8G, there is a significant peak in intervals 4 and 5 indicating that – while the observing Leader plans their own action during the time they do not act – they also look at potential objects the Follower could touch. This indicates that the Leader in some cases apparently plans not only their own action but also mentally anticipates potential plans of the Follower to counteract the Leader.

As expected, we found that Followers seemed to spot the planned target object of the Leader very early after the Leader's hand movement onset (Figure 8). This did not take much longer than approx. 400ms. Thus, despite having the option to simply follow the Leader's hand movements, Followers still engaged in anticipatory cognition, pointing to a reciprocal prediction process between Leader and Follower.

An intriguing finding concerns the "tails" behind the main peak in the histograms in Figure 8 for the acting Leader and – slightly more pronounced – for the observing Follower. Note that these gazes were looks at objects that did not exist there anymore as they have been moved a moment ago. Hence in some sense they represent gazes into the past. We speculate that this looking back indicates a memorization process (O'regan, 1992, Johansson&Johannson, 2014, Foerster, 2019). After all, the Follower needs to remember how the scene looked before the Leader changed it. However, also the Leader might want to remember this in order to see whether or not the Follower then performs the correct undo action.

In sum, this study highlights distinct decision-making and planning traits in Leaders compared to Followers, in a complex time-continuous scenario not only in their own plans but also in anticipating potential actions of others, shedding light on the intricate cognitive processes involved in daily human interactions.

Acknowledgements:

This publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 454648639 - SFB 1528, "Cognition of Interaction", Project B01 and by Lower Saxony Ministerium für Wissenschaft und Kultur (MWK), Project: Kognitiv und Emphatisch Intelligente Kollaborierende Roboter (KEIKO), TP6.

References

Andrist, S., Collier, W., Gleicher, M., Mutlu, B., & Shaffer, D. (2015). Look together: Analyzing gaze coordination with epistemic network analysis. Frontiers in psychology, 6, 1016.

Andrist, S., Gleicher, M., & Mutlu, B. (2017, May). Looking coordinated: Bidirectional gaze mechanisms for collaborative interaction with virtual characters. In Proceedings of the 2017 CHI conference on human factors in computing systems (pp. 2571-2582).

Brown, L. E., Halpert, B. A., & Goodale, M. A. (2005). Peripheral vision for perception and action. Experimental Brain Research, 165, 97-106.

Flanagan, J. R., & Johansson, R. S. (2003). Action plans used in action observation. Nature, 424(6950), 769-771.

Flanagan, J. R., Rotman, G., Reichelt, A. F., & Johansson, R. S. (2013). The role of observers' gaze behaviour when watching object manipulation tasks: predicting and evaluating the consequences of action. Philosophical Transactions of the Royal Society B: Biological Sciences, 368(1628), 20130063.

Foerster, R. M. (2019). The function of “looking-at-nothing” for sequential sensorimotor tasks: Eye movements to remembered action-target locations. Journal of Eye Movement Research, 12(2).

Fuchs, S., & Belardinelli, A. (2021). Gaze-based intention estimation for shared autonomy in pick-and-place tasks. Frontiers in Neurorobotics, 15, 647930.

Garrido-Jurado, S., Muñoz-Salinas, R., Madrid-Cuevas, F. J., & Marín-Jiménez, M. J. (2014). Automatic generation and detection of highly reliable fiducial markers under occlusion. Pattern Recognition, 47(6), 2280-2292.

Gergle, D., & Clark, A. T. (2011, March). See what I'm saying? Using dyadic mobile eye tracking to study collaborative reference. In Proceedings of the ACM 2011 conference on Computer supported cooperative work (pp. 435-444).

Jocher, G., Stoken, A., Borovec, J., Changyu, L., Hogan, A., Diaconu, L., ... & Dave, P. (2020). ultralytics/yolov5: v3. 0. Zenodo

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. Journal of vision, 3(1), 6-6.

Hessels, R. S., Teunisse, M. K., Niehorster, D. C., Nyström, M., Benjamins, J. S., Senju, A., & Hooge, I. T. (2023). Task-related gaze behaviour in face-to-face dyadic collaboration: Toward an interactive theory?. Visual Cognition, 31(4), 291-313.

Hessels, R. S., Li, P., Balali, S., Teunisse, M. K., Poppe, R., Niehorster, D. C., ... & Hooge, I. T. (2024). Gaze-action coupling, gaze-gesture coupling, and exogenous attraction of gaze in dyadic interactions. Attention, Perception, & Psychophysics, 1-17.

Huang, C. M., Andrist, S., Sauppé, A., & Mutlu, B. (2015). Using gaze patterns to predict task intent in collaboration. Frontiers in psychology, 6, 1049.

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. Nature reviews neuroscience, 2(3), 194-203

Johansson, R. S., Westling, G., Bäckström, A., & Flanagan, J. R. (2001). Eye–hand coordination in object manipulation. Journal of neuroscience, 21(17), 6917-6932.

Johansson, R., & Johansson, M. (2014). Look here, eye movements play a functional role in memory retrieval. Psychological science, 25(1), 236-242.

Kassner, W. Patera, and A. Bulling, (2014) "Pupil: An open source platform for pervasive eye tracking and mobile gaze-based interaction," in Adjunct Proceedings of the 149 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '14 Adjunct, (New York, NY, USA), pp. 1151–1160, ACM, 2014.

Land, M., Mennie, N., & Rusted, J. (1999). The roles of vision and eye movements in the control of activities of daily living. Perception, 28(11), 1311-1328.

Lavoie, E. B., Valevicius, A. M., Boser, Q. A., Kovic, O., Vette, A. H., Pilarski, P. M., ... & Chapman, C. S. (2018) Using synchronized eye and motion tracking to determine high-precision eye-movement patterns during object-interaction tasks. Journal of vision, 18(6), 18-18.) .

Mennie, N., Hayhoe, M., & Sullivan, B. (2007). Look-ahead fixations: anticipatory eye movements in natural tasks. Experimental brain research, 179, 427-442.

Möller, C., Zimmer, H. D., & Aschersleben, G. (2015). Effects of short-term experience on anticipatory eye movements during action observation. Experimental Brain Research, 233, 69-77.

Oliva, A., Torralba, A., Castelhano, M. S., & Henderson, J. M. (2003, September). Top-down control of visual attention in object detection. In Proceedings 2003 international conference on image processing (Cat. No. 03CH37429) (Vol. 1, pp. I-253). IEEE.

O'regan, J. K. (1992). Solving the" real" mysteries of visual perception: the world as an outside memory. Canadian Journal of Psychology/Revue Canadienne de Psychologie, 46(3), 461.

Pannasch, S., Helmert, J. R., Roth, K., Herbold, A. K., & Walter, H. (2008). Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. Journal of Eye Movement Research, 2(2).

Pelz, J. B., & Canosa, R. (2001). Oculomotor behavior and perceptual strategies in complex tasks. Vision research, 41(25-26), 3587-3596.

Stolzenwald, J., & Mayol-Cuevas, W. W. (2018, October). I can see your aim: Estimating user attention from gaze for handheld robot collaboration. In 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3897-3904). IEEE.

Sullivan, B., Ludwig, C. J., Damen, D., Mayol-Cuevas, W., & Gilchrist, I. D. (2021). Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. Journal of vision, 21(3), 13-13.

Williams, S. Barrus, R. Morley, and P. Shirley, (2005). An efficient and robust ray-box intersection algorithm. J. Graphics Tools, vol. 10, pp. 49–54, 01 2005.

Wurtz, R. H., McAlonan, K., Cavanaugh, J., & Berman, R. A. (2011). Thalamic pathways for active vision. Trends in cognitive sciences, 15(4), 177-184.