accepted for publication in *Psychological Research*

Neural theory for the perception of causal actions

Falk Fleischer[1], Andrea Christensen[1], Vittorio Caggiano[2,3], Peter Thier[2], and Martin A. Giese[1*]

[1] *Section for Computational Sensomotorics*
Department of Cognitive Neurology
Hertie Institute for Clinical Brain Research and
Centre for Integrative Neuroscience
University Clinic Tübingen
Frondsbergstrasse 23, D-72070 Tübingen, Germany

[2] *Department of Cognitive Neurology*
Hertie Institute for Clinical Brain Research and
Centre for Integrative Neuroscience
University Clinic Tübingen
Frondsbergstrasse 23, D-72070 Tübingen, Germany

[3] *McGovern Institute for Brain Research*
Massachusetts Institute of Technology (MIT)
77 Massachussetts Avenue
Building 46, Room 6177
Cambridge, MA 02139 USA

* To whom correspondence should be addressed

Phone: (+49) 7071 2989124

Fax: (+49) 7071 29 4790

Email: martin.giese@uni-tuebingen.de

Abstract:

The efficient prediction of the behavior of others requires the recognition of their actions and an understanding of their action goals. In humans this process is fast and extremely robust, as demonstrated by classical experiments showing that human observers reliably judge causal relationships and attribute interactive social behavior to strongly simplified stimuli consisting of simple moving geometrical shapes. While psychophysical experiments have identified critical visual features that determine the perception of causality and agency from such stimuli, the underlying detailed neural mechanisms remain largely unclear, and it is an open question why humans developed this advanced visual capability at all. We created pairs of naturalistic and abstract stimuli of hand actions that were exactly matched in terms of their motion parameters. We show that varying critical stimulus parameters for both stimulus types leads to very similar modulations of the perception of causality. However, the additional form information about the hand shape and its relationship with the object supports more fine-grained distinctions for the naturalistic stimuli. Moreover, we show that a physiologically plausible model for the recognition of goal-directed hand actions reproduces the observed dependencies of causality perception on critical stimulus parameters. These results support the hypothesis that selectivity for abstract action stimuli might emerge from the same neural mechanisms that underlie the visual processing of natural goal-directed action stimuli. Furthermore, the model proposes specific detailed neural circuits underlying this visual function, which can be evaluated in future experiments.

Introduction

The prediction of others' behavior is a fundamental requirement for human interaction. It requires the recognition of the actions of others and an understanding of their action goals. This behavior is extremely important for survival and is accomplished quickly and robustly. Classical experiments demonstrate that human social interactions and causal relationships related to actions can be recognized with high reliability even from strongly impoverished stimuli consisting of simple moving geometrical shapes (Heider and Simmel 1944; Michotte 1946 / 1963). An example is a stimulus display consisting only of two moving disks, where one starts to move when the other one stops to move in the same direction. This stimulus induces the impression of causality ('launching effect'), i.e. participants perceive the movement of the second disk as caused by the first. However, when the spatial or temporal relationship between the two disks is disturbed this percept of causality can disappear (Scholl and Tremoulet 2000). The attribution of causality and intentions to such simple stimuli seems to be universal and consistent over different cultures (Leslie and Keeble 1987; Barrett et al. 2005).

It was hypothesized by Michotte that the capability to interpret such interactive movements might be innate and dependent on specific mechanisms. Work in developmental psychology shows that this capability is present already early during development, before the age of one year (Leslie and Keeble 1987; Rochat et al. 1997; Saxe and Carey 2006), and that it is modifiable by learning and experience (see Schlottmann et al. 2006 for a discussion). Many of Michottes' early findings on perceptual causality were replicated by other researchers (Scholl and Tremoulet 2000), and some work has extended the study of the perception of abstract motion stimuli to the study of inferences about intentions (e.g. Dasser et al. 1989; Schlottmann and Shanks 1992; Baker et al. 2009). Detailed psychophysical studies showed that the perception of causality in simple displays is critically dependent on the spatial and temporal contingency of the moving discs, and specifically on their direction and relative speed, in line with Michottes' original findings (Beasley 1968; Bassili 1976; Schlottmann and Anderson 1993; Dittrich and Lea 1994; White and Milne 1997; Blythe et al. 1999; Oakes and Kannass 1999; Schlottmann et al. 2006; Choi and Scholl 2006).

Knowledge about the neural mechanisms that might underlie the interpretation of such interactive motion displays is quite limited. Imaging studies have extensively studied cortical areas involved in the interpretation of such stimuli in terms of intentional actions, reporting selective activation specifically in the posterior superior temporal sulcus (pSTS) and the neighboring temporo-parietal junction (TPJ) (Frith and Frith 1999; Castelli et al. 2000, Allison et al. 2000; Frith and Frith 2003; Blakemore and Decety 2001; Saxe et al. 2004; Schultz et al. 2004; Brass et al. 2007; de Lange et al. 2008; Hamilton and Grafton 2008; Jastorff et al. 2011). For stimuli involving perceptual causality, selective activation in the intraparietal sulcus and the inferior parietal lobule as well as the medial frontal gyrus has been reported, in addition to the superior temporal regions (Blakemore and Decety 2001; Fonlupt 2003; Fugelsang et al. 2005). A lesion study with split-brain patients points to a lateralization of the associated neural processes, the perception of launching events being localized predominantly in the right hemisphere (Roser et al. 2005). These temporal, parietal and frontal regions form a densely connected network of areas known to be involved in the perception of natural action stimuli (see e.g. Van Overwalle and Baetens 2009).

At the level of single cells in macaque cortex, a similar interconnected network of areas has been shown to be activated during action perception (Rizzolatti and Sinigaglia 2010; Nelissen et al. 2011). In particular, in the macaque superior temporal sulcus neurons have been observed that are selective to the observation of movements of the body or body parts relative to objects in the surround (Perrett et al. 1989; Jellema and Perrett 2006; Barraclough et al. 2009). It seems possible that such neurons are also involved in the representation of interactive movements, potentially also for abstract stimuli. In functional imaging studies it has been observed that cortical regions involved in the observation of natural actions, such as the superior temporal sulcus, and parietal and premotor cortex, might also be recruited during the observation and interpretation of highly abstract action stimuli (Castelli et al. 2000; Martin and Weisberg 2003; Ohnishi et al. 2004; Schultz et al. 2004; Schubotz and von Cramon 2004; Reithler et al. 2007; Petroni et al. 2010). However, beyond a localization of potentially relevant cortical areas, knowledge about detailed neural circuits underlying the perception of causality from action stimuli is completely lacking.

While there are no detailed neural theories about the processing of causal interactions, a small amount of work exists on possibly underlying computational mechanisms. Blythe and colleagues (1999) demonstrated that a neural network model based on simple visual cues, such as the relative motion of the disks, reliably predicts participants' judgments about the intentionality of observed movements. This study shows that performance in this apparently highly cognitive task might be dependent on relatively elementary visual features that characterize the interaction between the moving elements. Another recent abstract model based on cognitive schemata theory has been proposed by Rips (2011). Other models have tried to account for related phenomena by Bayesian inference and inverse probabilistic planning (Baker at al. 2009). None of these models makes a direct link to physiological mechanisms, or even attempts to explain how the detection of causal events could be accomplished based on real video stimuli.

Based on previous theoretical work on the encoding of goal-directed hand movements (Fleischer et al. 2009; Fleischer and Giese 2010), we propose in this paper a neurally-inspired theory for the recognition of interactive movements from abstract motion displays. This theory is based on the hypothesis that the visual analysis of abstract motion displays can be explained by the neural mechanisms that are normally responsible for the processing of natural stimuli showing goal-directed movements, such as hand actions. We claim that some of the observed phenomenology for the perception of abstract movements can be derived from such mechanisms, when it is additionally assumed that the accuracy of form processing is reduced during the processing of abstract motion stimuli.

In the following, we will provide arguments in support of this hypothesis: 1) Exploiting a new set of video stimuli that present the same goal-directed hand actions in a natural and in an abstract way, we show that ratings of naturalness and the attribution of causality are very similar between those two stimulus classes. Observed differences indicate that the processing of abstract stimuli is less sensitive to spatial manipulations of the stimulus than the processing of naturalistic action stimuli. 2) We demonstrate that variations of causality and naturalness ratings with stimulus manipulations, which are known to affect the perception of causality, can be qualitatively reproduced with a physiologically-inspired model for the recognition of naturalistic goal-directed hand actions. The only manipulation

that was necessary to adapt this model for the processing of abstract stimuli was a reduction of the tuning accuracy.

Methods

Our psychophysical experiment compared ratings of manipulated action stimuli in terms of their naturalness and perceived causality. We used naturalistic stimuli of goal-directed hand actions (grasping and pushing), where we modified the spatial and temporal parameters of the hand and object movement along dimensions that were known to affect the perception of causality from simple displays. These stimuli were generated by video manipulation from two original movies in order to achieve precise control of the spatial and temporal parameters, keeping the shapes of effector and object exactly the same. In addition, we generated a set of abstract action stimuli that closely matched the naturalistic displays in terms of their motion parameters. The matched set of abstract action stimuli was derived from the naturalistic stimuli by tracking the positions of the hand and object and replacing them by two circular discs. Similar methods were recently proposed for the generation of abstract versions of intentional full-body movements (McAleer and Pollick 2008).

The model presented in this paper has been developed originally in order to account for the properties of action-selective single cells in monkey cortex. The available space in this article permits only to lay out the major concepts underlying the architecture of the neural model. With respect to the technical details about the implementation, the simulations of physiological data, and a more elaborate evaluation of the computational performance of the model with natural action videos we refer to previous publications (Fleischer et al. 2009; Fleischer and Giese 2010).

Participants

Eighteen volunteers from the University of Tübingen with normal or corrected-to-normal vision (12 male, 6 female; age 21 to 41 years) participated in the psychophysical study. All were naïve with respect to the purpose of this experiment and gave informed consent prior to testing. Participants received a financial compensation for taking part in the experiment. The study was in accordance with

the declaration of Helsinki and approved by the ethics committee of the Eberhard-Karls-University Tübingen.

Materials

Naturalistic video stimuli

Video stimuli of hand actions were recorded from a single perspective (side view) using a custom video camera (Sony PCR-5 Camcorder, 576x720 pixels, 25Hz). Two types of actions were recorded: (1) pushing a ball (diameter 8 cm) with the right hand, the hand moving from right to left, and the ball continuing to move to the left side after contact; and (2) grasping of the ball, lifting it, and displacing it to the right side. The first stimulus is similar to the classical 'launching stimulus' by Michotte (see Fig. 1 A and B). Hand movements started from a resting position at approximately 40 cm distance to the right of the ball.

----- please insert Figure 1 about here ----

We generated a set of video stimuli by varying critical parameters that were known from the literature to influence the perception of causality from abstract stimuli. For this purpose, we separated the hand and the object by segmenting them from the background using commercial software (Adobe™ AfterEffects). The resulting video streams were spatially resampled (500 x 1000 pixels, 25Hz) and recombined using custom-made software (implemented in Matlab 7.6, The MathWorks™). All stimuli were generated by overlaying the images containing the acting hand on top of the images of the object in order to generate normal occlusion patterns. The size of the hand and the object in the final stimulus corresponded to 3.8° respectively 1.7° visual angle. The whole action took about 1200 ms for grasping stimuli and 680 ms for pushing stimuli. The overall stimulus area subtended about 18° by 33° of visual angle.

Novel artificial video stimuli were generated by manipulating the distance between the hand and the object, the point of contact, and their relative timing on each individual video frame. In the *Shift* condition we varied the distance of hand and object by displacing the hand along the horizontal axis

(Fig. 2A). As a result, the hand did not touch the object and rather appeared to mimic the action at different distances from the object (50, 100, 150, 200, and 250 pixels). In the *Contact point* condition we rotated the center of gravity (CoG) of the hand stimulus about the CoG of the object clockwise by different angles (90°, 45°,0°, -45°, -90°), where the distance between the two CoGs was kept constant (Fig. 2B).

----- please insert Figure 2 about here ----

In a third set of conditions (*Pause*) the frame during which the hand first touched the object was repeated multiple times (resulting in presentation times of the initial contact event of 40 (no repetition), 200, 400, 600, and 800 ms). Longer pauses result in the perceptual impression that hand and the object stop briefly in the middle of the interaction (Fig. 3 A). The final set of conditions (*Time gap*) was created by introducing time delays with different durations (0, 40, 120, 200, 280, 360 ms) between the movement of the hand and the object. This causes the impression that the object responds to the action of the hand in a delayed fashion, somewhat like there was a rubber band between the hand and the object (Fig. 3B).

----- please insert Figure 3 about here ----

Abstract stimuli

For the generation of abstract motion stimuli from the naturalistic video stimuli the hand and the object were replaced by two circular discs with a diameter of 60 pixels (2° of visual angle) using custom-made software (implemented in Matlab 7.6, The MathWorks™). The hand was replaced by a green and the object by a blue disc, located at the corresponding CoGs in the naturalistic stimulus (Fig. 1 C, D). The green disc was slightly shifted along the line connecting the CoGs of object and hand in order to assure a tangential contact between the two discs at the same time the hand first touched the object in the corresponding naturalistic non-modified stimulus. The absolute and relative locations as

well as the motion patterns of the disc stimuli thus matched as closely as possible those of their realistic counterparts. Examples of the stimuli can be downloaded as supplementary material.

Procedure / experimental design

Participants rated all stimuli with respect to 1) their similarity to normal hand-object interactions (*Naturalness*), and 2) in how far they induced the impression that one stimulus element caused the movement of the other (*Causality*). The second task was chosen in accordance with the classical rating tasks used in many previous studies on causal interactions (Michotte 1946 / 1963; Scholl and Tremoulet 2000). The participants sat in front of a computer screen at a distance of approximately 50 cm. Stimuli were presented on a Dell Inspiron™ TFT monitor with a frame rate of 60 Hz. The video stimuli covered an area of about 18° x 33° visual angle on the screen.

The whole experiment consisted of three phases. Written instructions were given before each phase individually to each subject, and participants were asked whether they had understood the tasks. In each phase pushing and grasping stimuli were presented in random order. In the first phase, the abstract versions of the original actions as well as their most extreme manipulations were presented to the participants in random order (12 stimuli in total). Participants were first asked to rate their intuitive impression whether the green ball made the blue ball move. The purpose of this first phase was to assess the consistency of the participants' interpretations of the abstract stimuli, before their judgments were biased by the knowledge of the original natural action stimuli. Responses were given by adjusting a slider on a scale from 0 ('No, not at all') to 1 ('Yes, very much') in steps with a size of 0.1. Next, participants were asked to give a brief written explanation of the reasons for their judgments. Participants were allowed to watch the same stimulus multiple times, pressing a repetition key.

In the second phase of the experiment all artificial stimuli (including the non-manipulated ones; 10 stimuli with shifts, 8 stimuli with modified contact point, 8 conditions with pauses, and 10 conditions with time gaps) were presented in two subsequent blocks in random order. Participants had to rate, first, to which degree the presented stimulus corresponded to a normal hand-object interaction (*Naturalness*). Second, they had to rate the strength of their impression that the green disc made the blue one move (*Causality*). Responses were again given by adjusting sliders on a scales from 0 ('No,

not at all') to 1 ('Yes, very much') in steps with the size 0.1. In this phase stimuli were displayed only once. As we were interested in how far the abstract stimuli were judged as similar to real movies of grasping and pushing, we showed a single example of natural grasping and pushing in the instruction of this phase. In the third phase, participants were presented with all naturalistic stimuli, 40 in total, in two blocks with random order. The task was identical to the one in the second phase described above.

Model architecture

The proposed model was originally developed to account for electrophysiological data from action selective neurons in monkey cortex, addressing in particular the visual tuning properties of neurons in the STS and of mirror neurons in area F5. In contrast to other models for the mirror neuron system in the literature that focus on the influence of motor representations on action recognition (Oztop et al. 2006; Bonaiuto and Arbib 2010; Tessitore et al. 2010; Chersi et al. 2011), our model focuses specifically on the visual processing mechanisms for actions. The model is computationally powerful enough to recognize goal-directed hand actions from real video stimuli. Details about this work can be found in Fleischer et al. (2009). We demonstrate here that the same neural architecture can account for the perception of causality from abstract action stimuli. The major modification of the model was that we reduced the selectivity of the form-selective neurons in the model for the abstract stimuli. A task-dependent modulation, e.g. of the width or gain of tuning functions has been observed regularly, for example, in the context of attentional manipulations or perceptual learning (e.g. Treue & Maunsell, 2006; Kourtzi & Connor, 2011). An overview of the model architecture is given in Figure 4.

----- please insert Figure 4 about here ----

The model consists of three major modules: A) A form recognition hierarchy, modeling form-selective neurons in the ventral visual stream including the primary visual cortex, area V4, and IT, as well as form-selective neurons in the dorsal stream of the monkey cortex including the STS; B) an affordance module that computes information about the relationship between effector and object, i.e. the matching of the hand and object shape and their relative positions and speed. This module implements

computational functions which are likely realized by neurons in parietal cortex, such as the inferior parietal lobule (IPL) or the anterior intraparietal area (AIP); C) a third module that models neural representations of goal-directed actions in premotor and parietal cortex. The first level of this module represents the action in a time-resolved manner, with neurons that encode specific temporal phases (similar to grip phases), while the second level represents actions independent of their intrinsic time structure. The neurons on this second level are active when a particular goal-directed action (grasping or pushing) is perceived. Their activity makes it possible to predict the behavioral results from psychophysical experiments addressing the perception of causality.

The first module (Figure 4A), the *form recognition hierarchy*, is a physiologically-inspired model for the recognition of shapes, following the principles of many other established models for object recognition (e.g. Riesenhuber and Poggio 1999; Deco and Rolls 2005). It mimics the hierarchical structure of the ventral visual pathway, starting from primary visual cortex to higher form-selective structures such as area IT or equivalent structures in the dorsal stream. Simple cells in area V1 are modeled by Gabor filters with different orientations and spatial resolution levels. Complex cells are modeled by pooling of the outputs of simple cells with the same preferred orientation within a limited spatial receptive field using maximum operations. Mid-level shape detectors (*shape fragment detectors*) are modeled by combining the responses from the complex cells by radial basis functions. The selectivity of these detectors was optimized by unsupervised learning (using k-means clustering) from a training data set. These pattern detectors learned to represent a characteristic dictionary of mid-level form features, corresponding to parts of objects or hands. Such features are likely represented in area V4 and TEO of the monkey cortex. (For related modeling approaches for mid-level feature detectors cf. e.g. Serre et al. (2007) or Ullman (2007)). The highest level of the form recognition hierarchy is formed by model neurons that are selective for the shape of the whole goal object and the whole hand (*object and hand shape detectors*). These neurons are also modeled by radial basis functions whose selectivity is optimized by supervised learning (i.e. by training on a set of naturalistic stimuli that the model is supposed to recognize). An overview of the key properties of the model neurons along the visual recognition hierarchy is given in Table 1. Further details can be found in Fleischer et al. (2009).

----- please insert Table 1 about here ----

The form recognition pathway deviates from established object recognition models with respect to two properties: First, even at the highest level of the hierarchy the neural detectors are not completely position-invariant, as is the case in many other models for object recognition. Instead, they have receptive fields with a corresponding diameter of about 4°, compatible with electrophysiological data from area IT in the monkey (Op De Beeck and Vogels 2000; Di Carlo and Maunsell 2003). This allows to estimate the retinal positions of the goal object and the hand from the highest level of the recognition hierarchy. Second, the detectors for the hand shape are embedded in a recurrent neural network which makes the activity in the network dependent on the temporal order of the individual hand shapes in the stimulus movies. Following earlier work on motion recognition (Giese and Poggio 2003), we modeled temporal order selectivity by introducing asymmetric lateral couplings between the hand shape selective neurons (see inset in Fig. 4A). The outputs of the sequence-selective hand shape detectors are further analyzed in two different ways: First, they feed into the second module B) supporting the estimation of the retinal position of the hand. Second, the responses of all hand shape neurons that are selective for hand postures belonging to the same type of hand movement are summed up by *motion pattern neurons.* These neurons encode types of hand movements such as grasping or pushing, independent of the goal object.

The second module of the model (Figure 4B), the *affordance module,* recombines the following types of information about object and effector: 1) It determines the matching of the shapes of the goal object and the hand, 2) it computes their relative position and 3) their relative speeds (distinguishing approaching, moving apart, moving together). The core component of this module is a *relative position map* that represents the retinal position of the hand relative to the object as an activation peak in a two-dimensional neural activity map. This map is computed by a gain field mechanism (Salinas and Abbott 1995; Pouget and Sejnowski 1997). This is a feed-forward network that combines the outputs of shape-selective neurons form module A) in a multiplicative manner. (See Fleischer et al. (2009) for further details). One can learn a region in this relative position map that corresponds to

hand positions relative to the object that would arise during successful grips. We assume the existence of *affordance neurons* that sum the activity in the relative position map within this region. These neurons are activated only by spatial hand-object configurations that are typical for successful actions. The second useful information that can be extracted from the relative position map by simple neural mechanisms is the relative speed of hand and object, which corresponds to the speed of the activity peak in the map. Direction and speed of this peak are detected by *relative speed neurons*, which are modeled as simple correlative motion detectors (motion energy detectors), as extensively discussed as models for direction selective neurons in primary visual cortex (for review see Smith and Snowden (1994)). Finally, the output signals of subsets of the relative speed neurons are pooled by *relative motion neurons*. These neurons signal characteristic types of relative motion that are relevant for the analysis of goal-directed actions: approaching of hand and object, moving apart, or moving together. For example, the neuron detecting approaching pools the outputs of all relative motion neurons signaling motion of the hand towards the goal object, independent of the global motion direction or exact speed. In a similar way, detectors for the other motion events can be constructed. Similar circuits for the detection of complex motion patterns have been proposed as models for neurons in area MST (Koenderink et al. 1985; Beardsley and Vaina 2001). One class of relative motion detects essentially the absence of relative motion (moving together).

The third module (Figure 4C) contains model neurons with selectivity for goal-directed actions. These neurons combine the following information provided by the earlier modules: 1) Type of the hand action (grasping or pushing), as signaled by the motion pattern neurons; 2) matching of hand and object shape and their relative position, as signaled by the affordance neurons; 3) type of relative motion between hand and object, as signaled by the relative motion neurons. These different inputs are combined by *action state neurons*, which are again modeled by radial basis function units that are trained in a supervised manner from example actions. These units respond maximally during particular phases of individual goal–directed actions (e.g. the hand approaching the object or the object moving away after contact with the hand). Such behavior is typical for higher action selective neurons, e.g. in the superior temporal, parietal or premotor cortex. Finally, the highest level of our model is given by *action neurons* that sum the activity of the different action state neurons belonging to the same action

type. These neurons signal the presence of particular actions independent of particular phases in time. The activities of these neurons were compared to the psychophysical results.

Simulation procedure

For a fair comparison of the model performance to the experiment we fitted the response obtained at the level of the action neurons to the average experimental results. In order to simplify a quantitative comparison between the human ratings and the simulation results from the model, the model responses for the original, non-manipulated action stimuli of grasping and pushing were rescaled to match to the corresponding average causality ratings in the experiment. All other model responses for grasping and pushing stimuli were re-scaled by the same factor accordingly. Furthermore, we fitted the tuning parameters of the action state neurons, adjusting the tuning width parameters separately for the radial basis function inputs from the affordance neurons and the relative speed neurons.

A key assumption underlying our simulations was that the main difference between the processing of realistic and abstract action stimuli is the accuracy of the form tuning in the processing hierarchy. After training of the system with the naturalistic stimuli, for the processing of the abstract stimuli we reduced the accuracy of the form recognition hierarchy by lowering the firing thresholds of the neurons at the level of the shape detectors. This led to a strongly reduced selectivity of the shape detectors which then responded also to arbitrary shapes, such as the discs. As result, detectors for object shape as well as for hand shapes were equally activated at the location of the two discs. In situations where the two blobs overlapped within the receptive fields of the neurons computing the relative position map, the leftmost activity maximum was assigned to the hand and the rightmost to the object. This disambiguation seemed justified given that in the real experiment the blobs had different colors, and since participants were explicitly told which disc represented the hand and the object. As result, hand and object detectors were activated by artificial stimuli at very similar locations as for naturalistic grasping and pushing stimuli.

In addition, we increased the width of the Gaussian tuning functions of the action state neurons for the artificial stimuli in order to decrease their pattern selectivity in a similar way as for the shape detectors. Responses of the action state neurons for abstract disc stimuli were thus solely dependent on the patterns of relative position and motion. Gradual modulations of tuning properties of cortical

neurons have been observed, e.g. in the context of attentional modulation (e.g. Treue 2001; Deco and Rolls 2005), and it seems plausible that the cortex might be able modulate such properties in a task-dependent manner.

Results

In the following, we first present the psychophysical results comparing naturalistic and abstract stimuli in terms of the naturalness ratings (i.e. the similarity of the stimuli with natural hand actions) and the causality ratings. We then show that the neural model is able to reproduce the observed dependencies on the stimulus parameters.

Ratings for the non-manipulated movements

Figure 5 shows the normalized average ratings of naturalness and causality for the original, unmanipulated grasping and pushing actions as well as the corresponding abstract stimuli (cf. Figure 1). Normalization was necessary in order to make the ratings of different observers more comparable since not all participants used the full range of available ratings. Naturalness and causality ratings were normalized independently and for each participant by transforming the range of ratings linearly so that the minimum was 0 and the maximum 1.

All ratings of naturalness and causality for both stimulus types and both actions were consistently high and significantly above the midpoint (0.5) of the normalized rating scale (Wilcoxon signed rank test, all $p < 0.001$). This indicates that all stimuli were rated as quite similar to normally occurring hand object interactions. This likely makes them efficient as stimuli that induce the impression of causality in the sense of Michotte.

----- please insert Figure 5 about here ----

To test for differences between the stimuli types and actions we conducted two-factor repeated measures ANOVAs, separately for the two variables naturalness and causality with the factors

Stimulus type (naturalistic vs. abstract) and Action type (grasping vs. pushing). The ANOVA for the naturalness ratings revealed no significant main effect for the stimulus type ($F(1,17) = 1.928$, $p = 0.183$) but a trend towards significance for the factor Action type ($F(1,17) = 3.392$, $p = 0.083$). This reflects the higher naturalness ratings for naturalistic grasping than for pushing movements, potentially caused by differences in the familiarity of the two types of actions. The interaction between both factors was not significant ($F(1,17) = 1.845$, $p = 0.192$).

The corresponding ANOVA for the causality ratings revealed no significant main effects, but a significant interaction between Stimulus and Action type ($F(1, 17) = 8.858$, $p = 0.008$). This is consistent with the result from post hoc testing by comparing natural and abstract stimuli for the individual actions, which revealed significantly higher causality ratings for naturalistic than for abstract grasping stimuli (Wilcoxon signed rank test, $p = 0.005$), while the same test for the pushing actions failed to show significant differences.

In summary, these results show high naturalness and causality ratings in the range of 0.75 to one, with a slight tendency of artificial stimuli being perceived as less natural than the naturalistic hand action stimuli for grasping stimuli. Especially the ratings for pushing actions failed to show significant differences between abstract and natural stimuli, potentially indicating a higher influence of detailed form cues in the processing of grasping actions.

Ratings for the manipulated movements

To further analyze the similarity between the two stimulus classes, novel stimuli were generated that included spatial and temporal manipulations that were known to affect the perception of causality according to the classical literature.

The first manipulation was the *Shift* condition, where the hand was translated horizontally within the image plain against the ball, creating a spatial gap between effector and object. Figure 6 (panel A) shows the naturalness and Figure 7 (panel A for grasping, B for pushing actions) the causality ratings for different spatial displacements. For both, the naturalistic and abstract stimuli the average ratings of naturalness and causality were dependent on the size of the displacement. All ratings show similar trends and decay quickly for increasing shift sizes and particularly fast and to a larger degree for the

naturalistic stimuli than for the abstract ones. However, some quantitative differences exist in terms of the exact shapes of the decay.

This qualitative observation is confirmed by a dependent measures ANOVAs with the three factors Shift size, Stimulus type (naturalistic vs. artificial), and Action type (grasping vs. pushing). For the naturalness ratings the main effect of Shift size is highly significant ($F(1.72,29.247) = 68.55$, $p < 0.001$ with Greenhouse-Geisser correction). In addition, the naturalness rating for naturalistic grasping movements compared to abstract stimuli, and compared to pushing actions, drops abruptly even for very small spatial deviations (50 pixel, see Figure 6A). This results in highly significant two-way interactions between Shift size and Stimulus type ($F(5, 85) = 6.05$, $p < 0.001$), and between Stimulus type and Action type ($F(1,17) = 17.51$, $p = 0.001$), as well as in a significant three-way interaction ($F(5,85) = 14.189$, $p < 0.001$). For the causality ratings only the main effect of Shift size ($F(2.2,36.6) = 23.401$, $p < 0.001$, Greenhouse-Geisser corrected) and the three-way interaction, observable as a shallower decay of the causality ratings for the abstract grasping stimuli compared to the other conditions in Figure 7A/B, were statistical significant ($F(3.1,53.4) = 3.12$, $p = 0.03$, Greenhouse-Geisser corrected).

In contrast to the result patterns for grasping actions, for stimuli depicting pushing movements both ratings – for naturalistic and for abstract stimuli – show a highly comparable curve progression and no main effect for the Stimulus type was found. The observed interactions are consistent with the fact that the ratings for naturalistic stimuli decay somewhat faster with the shift size, potentially reflecting increased sensitivity for small spatial mismatches between hand shape and object for the naturalistic stimuli.

----- please insert Figure 6 about here ----

The second manipulation was the variation of the *Contact point*, rotating the hand position about the ball. The rating results from this condition are shown in Figure 6 (panel B) and 7 (panels C for grasping and D for pushing actions) for different rotation angles. Both, the naturalness and the causality rating, peak at very similar values for both stimulus types (naturalistic and abstract) without

manipulation (rotation angle zero). Both measures decay monotonically for increasing deviations of the rotation angle from zero, resulting in increasing deviations from the normal contact points of the fingers with the object (respectively of the corresponding discs). The resulting 'tuning curves' are clearly wider for the abstract than for the naturalistic stimuli. This is even more evident for grasping actions where the curves for abstract stimuli are nearly flat lines (solid blue lines in Figure 6B and 7C). This coincides with the observation that even relatively small deviations of the contact points of the fingers with the object from the normal ones makes this stimulus look rather unnatural while the perception of abstract forms is less affected by small deviations. For pushing actions the manipulation of the Contact point resulted in a shallower decay of the participants' ratings, thus exact finger configuration with respect to the object was less critical for the perception of an natural scene depicting a causal interaction.

These qualitative observations are confirmed by a statistical analysis, again performing a three-factor ANOVA with the factors Rotation angle, Stimulus type, and Action type. For the naturalness ratings we observed significant main effects for the Rotation angle ($F(2.7,46.7) = 54.642$, $p < 0.001$ with Greenhouse-Geisser correction) as well as for Stimulus type (naturalistic vs. artificial) ($F(1,17) = 41.2$, $p < 0.001$), but not for the Action type. All two-way interactions are significant (Rotation angle x Stimulus type: $F(2.1, 36.3) = 13.66$, $p < 0.001$ with Greenhouse-Geisser correction; Rotation angle x Action type: $F(4,14) = 6.07$, $p < 0.001$) and Stimulus x Action type: $F(1,17) = 63.0$, $p < 0.001$ and also the three-way interaction ($F(4,68) = 12.95$, $p< 0.001$). Results were similar for the causality ratings with significant main effects for the Rotation angle and the Stimulus type ($F(2.6, 44.1) = 17.98$, $p < 0.001$ Greenhouse-Geissser corrected, respectively $F(1,17) = 6.811$, $p < 0.02$), but not the Action type. All two-way interactions were significant (Rotation angle x Stimulus type: $F(2.2, 38) = 4.40$, $p = 0.016$; Rotation angle x Action type: $F(3.7, 59.1) = 3.21$, $p < 0.04$, both Greenhouse-Geisser corrected; Stimulus x Action type: $F(1,17) = 7.1$, $p < 0.01$) and also the three-way interaction ($F(4,68) = 7$, $p< 0.001$). The reduced width of the observed 'tuning curve' for the abstract stimuli may be interpreted as indication that such stimuli are processed with less accurate form tuning.

----- please insert Figure 7 about here ----

Our third manipulation was the *Pause* condition, where the frame of the first hand-object contact was repeated for time intervals with variable durations. The rating results from this condition are shown in Figure 6 (panel C) and 8 (grasping: panel A, pushing: panel B) for different durations of the pause. Notably, this manipulation resulted in the most obvious differences between grasping and pushing actions. While for grasping actions – independent of the Stimulus type – the length of the Pause at the contact point seems to have nearly no influence on the judgments of naturalness and causality (Figure 6C / 8A), both ratings decay quickly for the pushing actions (Figure 6C / 8B), again showing qualitatively very similar trends.

For more detailed quantitative analysis, we performed an independent-measures ANOVA with the three factors Duration, Stimulus type (naturalistic vs. artificial) and Action type (grasping vs. pushing). For the naturalness ratings the main effect of the Duration is highly significant ($F(2.2, 37.3) = 14.70$, $p < 0.001$ with Greenhouse-Geisser correction), although mainly driven by the pushing actions. In addition, the main effect of the Action type ($F(1,17) = 36.28$, $p < 0.001$) and the two-way interaction between the last two factors are significant ($F(4, 68) = 12.20$, $p < 0.001$). A similar picture arises for the causality ratings: The main effects of Duration and Action type are significant ($F(2.94, 50.1) = 16.17$, $p < 0.001$, Greenhouse-Geisser corrected respectively $F(1,17) = 15.16$, $p = 0.001$). So are also the two-way interactions between Duration and Action type ($F(3.1, 52.7) = 15.28$, $p < 0.001$ with Greenhouse-Geisser correction) and between Action type and Stimulus type ($F(1,17) = 11.13$, $p = 0.004$). All other effects were non-significant ($p > 0.05$). The lack of a main effect of Stimulus type is consistent with the similarity of the trends for the pushing stimuli. However, there is a difference between the ratings for the grasping stimuli that likely is responsible for the observed interaction effect.

----- please insert Figure 8 about here ----

The interactions with the factor Action type are consistent with the fundamentally different behavior for grasping and pushing stimuli. The ratings for the two actions are presented separately for the two actions in Figure 6C and 8A/B. The Pause manipulation basically did not affect the ratings for

grasping, while it had a strong influence on the ratings for pushing. Again ratings are similar for the two stimulus types. Two separate ANOVAs for the grasping and the pushing stimuli confirmed this observation. For grasping we found significant main effects of Stimulus type for the naturalness as well as for the causality ratings ($F(1, 17) = 6.88$, $p = 0.018$, respectively $F(1, 17) = 4.963$, $p = 0.04$). In addition, we found a significant interaction between Stimulus type and Duration for the causality ratings only $F(4, 68) = 3.20$, $p = 0.018$). For pushing, however, we found only a significant main effect for the Delay ($F(2.2, 37.3) = 15.54$, $p < 0.001$, respectively $F(2.3, 39.1) = 19.6$, $p < 0.001$, Greenhouse-Geisser corrected). The fact that the introduction of a pause did not affect naturalness and causality ratings for grasping seems plausible, since grasping with holding on the object for a while before lifting it is a valid and naturally occurring action, which, however, implies that the hand causes the movement of the ball.

The last manipulation tested was the *Time gap* condition, where a time delay was introduced between the movement of the object and the movement of the hand. The rating results from this condition are shown in Figure 6 (panel D) and 8 (grasping: panel C, pushing: panel D) for different durations of the delay. Both ratings decay with the duration of the delay and show qualitatively very similar differences between the two stimulus classes.

For a more detailed analysis we performed an independent-measures ANOVA with the three factors Duration, Stimulus type (naturalistic vs. artificial) and Action type (grasping vs. pushing). For the naturalness ratings the main effects of the Duration is highly significant ($F(1.6, 25.592) = 40.99$, $p < 0.001$ with Greenhouse-Geisser correction), and also the main effect of the Action type ($F(1,16) = 21.83$, $p < 0.001$). In addition, the two-way interaction between these two factors and between Duration and Stimulus type are significant ($F(3.3, 52.6) = 16.24$, $p < 0.001$ Greenhouse-Geisser corrected, respectively $F(5,12) = 3.8$, $p = 0.004$). Similar results were obtained for the causality ratings with significant main effects of Duration and Action type ($F(1.9, 30.68) = 43.11$, $p < 0.001$ Greenhouse-Geisser corrected, respectively $F(1,16) = 10.66$, $p = 0.005$) and significant two-way interactions between Duration and Action type ($F(5,80) = 3.19$, $p = 0.01$) and Duration and Stimulus type ($F(5,80) = 2.531$, $p = 0.035$). The interactions result from the fact that the ratings for grasping decay faster compared to the ratings for pushing (Figure 8C/D).

Summarizing, we found qualitatively quite similar trends for the two stimulus classes (naturalistic and abstract) for the tested stimulus manipulations. However, a detailed quantitative analysis revealed also some differences, especially in conditions where the exact localization of the fingers might be critical for the detection of successful grasping. In addition, for grasping the introduction of a pause interval at the frame of object contact did not have a substantial influence on naturalness and causality perception, opposed to the same manipulation applied to pushing stimuli.

Simulation results from the model

The simulation results of the model (Figure 4) compared to the causality ratings of the human participants are shown in Figure 7 for the spatial, and in Figure 8 for the temporal manipulations. The panels show the normalized activity of the action neurons at the highest level of the model hierarchy (cf. Figure 4C), averaged over time.

A comparison of model responses for the naturalistic stimuli with the human ratings for causality shows a close qualitative matching of the trends in dependence of the manipulation strength for grasping and pushing actions, with a very small number of exceptions. This good qualitative agreement is also supported by highly significant correlations (Table 2) between the model neurons' activities and the causality ratings in most cases, except for the ones where also the human data did not show significant variations with the manipulation strength (Contact point manipulation for abstract grasping stimuli, and Shift manipulation in grasping stimuli; indicated by diamonds in Table 2).

For the Contact Point manipulation the causality ratings for abstract grasping do not vary with the rotation angle, while they do so for the pushing action. This likely reflects the fact that a matching of the correct finger positions in grasping requires detailed shape information, which is not present in the abstract stimuli. Contrasting with the grasping stimuli, pushing stimuli result in less occlusions of the object by the hand, so that the detection of the correct contact points can still partially be accomplished based on relative position information. The model nicely reproduces this difference between the two stimulus classes.

For the Pause manipulation and grasping (naturalistic and abstract stimuli) the human ratings do not vary significantly with the pause duration while this is the case for pushing. Also this trend is reproduced by the model.

Like in the human data, the model shows often quite similar behavior for abstract and naturalistic stimuli. Also it reproduces many details of the patterns of human ratings. For most manipulations, the simulations reproduce accurately the decaying trends with the size of the manipulation, resulting in highly significant correlations between the human ratings and the activity of the action neurons. In many cases, the simulations reproduce also quite accurately the differences between the widths of the tuning curves for the Contact point manipulation between naturalistic and abstract stimuli.

Interestingly, even the fundamental difference in the trends between grasping and pushing actions for the time manipulations (cf. Fig. 7 B and C) is qualitatively reproduced: The dependence of the activity on the pause duration for grasping is rather flat while the curve for pushing decays. In the model this fundamentally different behavior emerges because the frozen frame of the grasping sequence activates adequately one of the action state neurons, which encodes the contact together with zero relative motion. For pushing, however, the contact frame is associated with non-zero relative motion between hand and object (first the hand approaches the resting object, then the object moves away from the resting hand). This implies that for this stimulus the replication of the contact frame results in an inadequate stimulus for the action state neurons, resulting in the observed decay of the activity with increasing duration of the delay.

The reproduction of the data at this level of detail seems quite astonishing, given that the model was originally developed for the processing of naturalistic action stimuli, and that no extra mechanisms were added for the processing of the abstract stimuli, except for a variation of the accuracy of the tuning.

Discussion

The recognition of actions of others requires the prediction of action consequences and goals, and classical experiments have demonstrated that humans can generate such predictions robustly even

from highly abstract stimuli, such as moving rigid geometrical shapes. This paper proposes a new neural theory for the perception of such abstract motion stimuli and the perception of causality assuming physiologically plausible simple neural mechanisms. Consistent with previous work in functional imaging (Castelli et al. 2000; Blakemore and Decety 2001; Fonlupt 2003; Martin and Weisberg 2003; Ohnishi et al. 2004; Schultz et al. 2004; Schubotz and von Cramon 2004; Fugelsang et al. 2005; Reithler et al. 2007), we hypothesized that the perception of abstract action stimuli might be explained by the same neural mechanisms as the perception of naturalistic goal-directed movements, such as object-directed hand actions. Going substantially beyond this previous work, our model proposes concrete neural circuits that are computationally sufficient for the processing of real action stimuli and which reproduce successfully, at least qualitatively, fundamental trends observed in psychophysical experiments on perceptual causality.

We provided two pieces of evidence in support of the hypothesis that real and abstract action stimuli might be processed by similar neural mechanisms. First, we compared the perception of naturalness and causality induced by naturalistic video stimuli showing grasping and pushing with the perception of the same measures from abstract motion stimuli, which consisted of two moving discs whose spatio-temporal parameters were exactly matched with the naturalistic stimuli. For both stimulus classes we found qualitatively very similar dependences on specific spatio-temporal manipulations that were known from previous work (Scholl and Tremoulet 2000) to affect the perception of causality. Apart from very similar trends in the parametric dependencies, we observed that the perception of naturalistic stimuli was more sensitive to spatial manipulations. This suggests that more fine-grained shape processing might play a critical role for the visual analysis of such stimuli, e.g. in order to verify the correct contact points of the fingers. As a second piece of evidence for our hypothesis we presented a physiologically-inspired model for the recognition of goal-directed hand actions that reproduces correctly the basic parametric dependencies observed in our psychophysical experiments, at least qualitatively. The only change compared to the original version of the model that was optimized for hand action recognition from real videos, was that we reduced the accuracy of the form tuning at several levels of the model. Such dynamic modulations of tuning properties have been shown to be present in visual cortex at earlier levels, e.g. in the context of attentional modulation (e.g.

Treue & Maunsell, 2006). The original model, at the same time reproduces a variety of results about the behavior of action-selective single cells in monkey cortex and has thus a direct link to detailed mechanisms in the cortex (Fleischer and Giese 2010). Given that this model was developed and optimized for the processing of naturalistic stimuli, we think that the observed generalization to abstract stimuli and the reproduction of parametric dependencies for this stimulus class is non-trivial and not necessarily expected.

Clearly the evidence provided is not sufficient as a complete proof of our hypothesis. For example, one might argue that there are many potential alternative mechanisms for the processing of causality, which operate in parallel to visual action processing and which work equally efficient for naturalistic and artificial stimuli. In addition, it seems likely that there are higher-level cognitive mechanisms, e.g. involving reasoning processes or inference about social intentions, which might be required to account for the attribution of more complex forms of causality (e.g. Rips, 2011; Baker et al. 2009). However, our theoretical model shows that plausible neural mechanisms for the visual processing of actions produce signatures very similar to the ones discussed in classical studies on perceptual causality. In this sense, our model provides sufficient explanation for some of the observed phenomena, but clearly lacks the proof of necessity. To our knowledge, there is so far no other work that gives an explicit implementation of mechanisms for the perception of abstract motion and causality that are applicable to real image sequences, nor are there any models that link such phenomena directly to the behavior of individual cortical neurons. Knowing that the model includes many strong simplifications and has serious shortcomings (such as the complete lack of top-down feedback, disparity cues, etc.), we think that it might be useful for experimentalists since it specifies exact computational mechanisms at a level that makes specific predictions at the level of individual neurons. This distinguishes the proposed model from a variety of more abstract models on causality perception in the literature (Blythe et al. 1999; Rips 2011). One of the most prominent predictions that follows from our theory is directly testable in physiological experiments: action-selective neurons at higher cortical levels, such as the parietal or the premotor cortex should show substantial generalization from naturalistic goal-directed action stimuli to abstract motion stimuli of the type discussed in this paper. Interestingly, this prediction could be recently confirmed in an electrophysiological experiment in monkeys assessing the

responses of mirror neurons in premotor area F5 using the same type of stimuli as in this study (Pomper et al., Abstracts of the Society for Neuroscience, 914.02 , 2011).

Finally, one might consider what the proposed theory might be able to contribute to central topics that are frequently discussed with respect to the perception of abstract motion and causality. One frequently discussed point is whether causality perception is based on innate mechanisms (Michotte 1946 / 1963; Scholl and Tremoulet 2000; Schlottmann et al. 2006; Rips 2011). While this question needs to be addressed thoroughly using methods from developmental psychology and potentially genetics, our computational model shows that in presence of an appropriate hierarchical architecture relatively elementary learning-based neural mechanisms are computationally sufficient to account for some of the observed phenomena in the context of the perception of causality. However, it seems likely that the basic structure of the underlying neural processing architecture is largely innate. A second issue is whether the perception of causality is a purely perceptual, or a higher cognitive phenomenon (Rips 2011). In our model the neurons reflecting the perception of causality emerge at the highest level (Figure 4C) of the processing hierarchy, corresponding to parietal and premotor levels of action processing. It is known that these levels of visual representations are linked to structures in the basal ganglia and the limbic system, e.g. the amygdala, known to be involved in processing non-visual aspects of causal interactions (e.g. Pessoa and Adolphs, 2010; Straube and Chatterjee, 2010). In addition, in some of these higher cortical regions visual and motor representations of actions clearly overlap at the level of individual neurons (e.g. Rizzolatti, Fogassi, and Gallese, 2001; Fogassi et al., 2005; Prinz, 1997). Such overlap might indicate a representation of actions at a relatively abstract level useful for the programming and control of reactive or interactive motor behavior. From a philosophical point of view, it seems to be a complex question to decide whether such high-level representations should be termed visual, motor, or cognitive.

Finally, it has to be mentioned that the present model addresses causal interactions only in a limited way, focusing on what has been called ‘physical causality’ (e.g. Schlottmann et al. 2006). We have not tested so far whether the same type of model can also be extended for the treatment of ‘social causality’, as studied in the classical displays by Heider and Simmel (1944) or Kanizsa and Vicario (1968). In this case the interaction of the two abstract objects is interpreted in terms of psychological

rather than of physical terms (for example as one disc 'chasing' another). Since the model structure that we propose has been originally derived from a neural model that accounts for the perception of biological motion (Giese and Poggio 2003) it has most ingredients for the recognition of movements of biological agents. 'Intentional' interactions would be characterized by the fact that the behavior of one agent specifies the goals for the other. The recognition of such interactive behavior seems again to essentially depend on the processing of the relationship between multiple agents, as accomplished by the neural circuitry illustrated in Figure 4B. However, the technical details of such a recognition circuit would have to be worked out and the solid testing of these ideas, using real-world and abstract interactive stimuli, defines an interesting topic for future research.

Acknowledgements

We thank H. Alhumsi for help with the data collection, D. Endres for stimulating discussions, M. Angelovska for help with the graphical illustrations, and K. Festl for help with the data analysis. This work was supported by the DFG, the EC FP7 projects SEARISE, AMARSI and TANGO and the Hermann and Lilly Schilling Foundation.

References:

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends Cogn Sci, 4*(7), 267-278.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329-349.

Barraclough, N. E., Keith, R. H., Xiao, D., Oram, M. W., & Perrett, D. I. (2009). Visual adaptation to goal-directed hand actions. *J Cogn Neurosci, 21*(9), 1806-1820.

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. (2005). Accurate judgments of intention from motion alone: A cross-cultural study. *Evolution and Human Behavior, 26*, 313-331.

Bassili, F. (1976). Temporal and Spatial Contingencies in the Perception of Social Events. *Journal of Personality and Social Psychology, 33*(6), 680-685.

Beardsley, S. A., & Vaina, L. M. (2001). A laterally interconnected neural architecture in MST accounts for psychophysical discrimination of complex motion patterns. *J Comput Neurosci, 10*(3), 255-280.

Beasley, N. A. (1968). The extent of individual differences in the perception of causality. *Can J Psychol, 22*(5), 399-407.

Blakemore, S. J., & Decety, J. (2001). From the perception of action to the understanding of intention. *Nat Rev Neurosci, 2*(8), 561-567.

Blythe, P. W., Todd, P. M., & Miller, G. F. (1999). How Motion Reveals Intention: Categorizing Social Interactions. In G. Gigerenzer, & P. M. Todd (Eds.), *Simple heuristics that make us smart* (pp. 257-285). Oxford: Oxford University Press.

Bonaiuto, J., & Arbib, M. A. (2010). Extending the mirror neuron system model, II: what did I just do? A new role for mirror neurons. *Biol Cybern, 102*(4), 341-359.

Brass, M., Schmitt, R. M., Spengler, S., & Gergely, G. (2007). Investigating action understanding: inferential processes versus action simulation. *Curr Biol, 17*(24), 2117-2121.

Castelli, F., Happe, F., Frith, U., & Frith, C. (2000). Movement and mind: a functional imaging study

of perception and interpretation of complex intentional movement patterns. *Neuroimage*, 12(3), 314-325.

Chersi, F. (2011). Neural mechanisms and models underlying joint action.*Exp Brain Res* 211(3-4): 643{53.

Choi, H., & Scholl, B. J. (2006). Measuring causal perception: connections to representational momentum? *Acta Psychol (Amst), 123*(1-2), 91-111.

Dasser, V., Ulbaek, I., & Premack, D. (1989). The perception of intention. *Science, 243*(4889), 365-367.

de Lange, F. P., Spronk, M., Willems, R. M., Toni, I., & Bekkering, H. (2008). Complementary systems for understanding action intentions. *Curr Biol, 18*(6), 454-457.

Deco, G., & Rolls, E. T. (2005). Attention, short-term memory, and action selection: a unifying theory. *Prog Neurobiol, 76*(4), 236-256.

Di Carlo, J. J., & Maunsell, J. H. R. (2003). Anterior inferotemporal neurons of monkeys engaged in object recognition can be highly sensitive to object *Neurophysiol, 89*, 3264-3278.

Dittrich, W. H., & Lea, S. E. (1994). Visual perception of intentional motion. *Perception, 23*(3), 253-268.

Fleischer, F., Casile, A., & Giese, M. A. (2009). Bio-inspired approach for the recognition of goal-directed hand actions. In X. Jiang, & N. Petkov (Eds.), *Conference on Computer Analysis of Images and Patterns (CAIP), LCNS* (Vol. 5702, pp. 714-722).

Fleischer, F., & Giese, M. A. (2010). Computational Mechanisms of the Visual Processing of Action Stimuli. In K. Johnson, & M. Shiffrar (Eds.), *Perception of the Human Body in Motion: Findings, Theory and Practice.* (Vol. in press): Oxford University Press.

Fonlupt, P. (2003). Perception and judgement of physical causality involve different brain structures. *Brain Res Cogn Brain Res, 17*(2), 248-254.

Frith, C. D., & Frith, U. (1999). Interacting minds - a biological basis. *Science, 286*(5445), 1692-1695.

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philos Trans R Soc Lond B Biol Sci, 358*(1431), 459-473.

Fugelsang, J. A., Roser, M. E., Corballis, P. M., Gazzaniga, M. S., & Dunbar, K. N. (2005). Brain mechanisms underlying perceptual causality. *Brain Res Cogn Brain Res, 24*(1), 41-47.

Gazzola, V., Rizzolatti, G., Wicker, B., & Keysers, C. (2007). The anthropomorphic brain: the mirror neuron system responds to human and robotic actions. *Neuroimage*, 35(4), 1674-1684.

Giese, M. A., & Poggio, T. (2003). Neural mechanisms for the recognition of biological movements. *Nat Rev Neurosci, 4*(3), 179-192.

Hamilton, A. F., & Grafton, S. T. (2008). Action outcomes are represented in human inferior frontoparietal cortex. *Cereb Cortex, 18*(5), 1160-1168.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychology, 57*, 243-249.

Jastorff, J., Clavagnier, S., Gergely, G., & Orban, G. A. (2011). Neural mechanisms of understanding rational actions: middle temporal gyrus activation by contextual violation. *Cereb Cortex, 21*(2), 318-329.

Jellema, T., & Perrett, D. I. (2006). Neural representations of perceived bodily actions using a categorical frame of reference. *Neuropsychologia, 44*(9), 1535-1546.

Kanizsa, G., & Vicario, G. (1968). The perception of intentional reaction. In G. Kanizsa, & G. Vicario (Eds.), *Experimental research on perception* (pp. 71-126). Trieste: University of Trieste.

Koenderink, J. J., van Doorn, A. J., & van de Grind, W. A. (1985). Spatial and temporal parameters of motion detection in the peripheral visual field. *J Opt Soc Am A, 2*(2), 252-259.

Kourtzi, Z. & Connor, C. E. (2011) Neural representations for object perception: structure, category, and adpative coding. *Annu Rev Neurosci* 34, 45-67.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition, 25*(3), 265-288.

Martin, A., & Weisberg, J. (2003). Neural foundations for understanding social and mechanical concepts. *Cogn Neuropsychol*, 20(3-6), 575-587.

McAleer, P., & Pollick, F. E. (2008). Understanding intention from minimal displays of human activity. *Behav Res Methods, 40*(3), 830-839.

Michotte, A. (1946 / 1963). *The Perception of Causality (Translated by T.R. Miles and E. Miles)*. London: Methuen: Basic Books.

Nelissen, K., E. Borra, M. Gerbella, S. Rozzi, G. Luppino, W. Vanduffel, G. Rizzolatti, and G. A. Orban (2011). Action observation circuits in the macaque monkey cortex. *J Neurosci* 31(10): 3743-3756.

Oakes, L. M., & Kannass, K. N. (1999). That's the way the ball bounces: Infants' and adults' perception of spatial and temporal contiguity in collisions involving bouncing balls. *Developmental Science, 2*(1), 86-101.

Op De Beeck, H., & Vogels, R. (2000). Spatial sensitivity of macaque inferior temporal neurons. *J Comp Neurol, 426*(4), 505-518.

Ohnishi, T., Moriguchi, Y., Matsuda, H., Mori, T., Hirakata, M., Imabayashi, E., et al. (2004). The neural network for the mirror system and mentalizing in normally developed children: an fMRI study. *Neuroreport*, 15(9), 1483-1487.

Oztop, E., Kawato, M., & Arbib, M. (2006). Mirror neurons and imitation: a computationally guided review. *Neural Netw, 19*(3), 254-271.

Pessoa, L. & Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat Rev Neurosci* 11(11): 773-783.

Petroni, A., Baguear, F., & Della-Maggiore, V. (2010). Motor resonance may originate from sensorimotor experience. *J Neurophysiol*, 104(4), 1867-1871.

Pouget, A., & Sejnowski, T. J. (1997). Spatial Transformations in the Parietal Cortex Using Basis Functions. *Journal of Cognitive Neuroscience, 9*(2), 222-237.

Prinz, W. (1997). Perception and Action Planning *European Journal of Cognitive Psychology, 9*(2), 129-154.

Reithler, J., van Mier, H. I., Peters, J. C., & Goebel, R. (2007). Nonvisual motor learning influences abstract action observation. *Curr Biol*, 17(14), 1201-1207.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nat Neurosci, 2*(11), 1019-1025.

Rips, L. J. (2011). Split identity: intransitive judgments of the identity of objects. *Cognition, 119*(3), 356-373.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat Rev Neurosci, 2*(9), 661-670.

Rizzolatti, G. & Sinigaglia, C. (2010). The functional role of the parietofrontal mirror circuit: interpretations and misinterpretations. *Nat Rev Neurosci* 11(4): 264-274.

Rochat, P., Morgan, R., & Carpenter, M. (1997). Young infants' sensitivity to movement information specifying social causality. *Cogn Develop, 12*, 537-561.

Roser, M. E., Fugelsang, J. A., Dunbar, K. N., Corballis, P. M., & Gazzaniga, M. S. (2005). Dissociating processes supporting causal perception and causal inference in the brain. *Neuropsychology, 19*(5), 591-602.

Salinas, E., & Abbott, L. F. (1995). Transfer of coded information from sensory to motor networks. *J Neurosci, 15*(10), 6461-6474.

Saxe, R., & Carey, S. (2006). The perception of causality in infancy. *Acta Psychol (Amst), 123*(1-2), 144-165.

Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia, 42*(11), 1435-1446.

Schlottmann, A., & Anderson, N. H. (1993). An information integration approach to phenomenal causality. *Mem Cognit, 21*(6), 785-801.

Schlottmann, A., Ray, E. D., Mitchell, A., & Demetriou, N. (2006). Perceived physical and social causality in animated motions: spontaneous reports and ratings. *Acta Psychol (Amst), 123*(1-2), 112-143.

Schlottmann, A., & Shanks, D. R. (1992). Evidence for a distinction between judged and perceived causality. *Q J Exp Psychol A, 44*(2), 321-342.

Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends Cogn Sci, 4*(8), 299-309.

Schubotz, R. I., & von Cramon, D. Y. (2004). Sequences of abstract nonbiological stimuli share ventral premotor cortex with action observation and imagery. *J Neurosci*, 24(24), 5467-5474.

Schultz, J., Imamizu, H., Kawato, M., & Frith, C. D. (2004). Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. *J Cogn Neurosci, 16*(10), 1695-1705.

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Trans Pattern Anal Mach Intell, 29*(3), 411-426.

Smith, A. T., & Snowden, R. J. (Eds.). (1994). *Visual Detection of Motion*. London: Academic Press Limited.

Straube, B., & Chatterjee, A. (2010). Space and time in perceptual causality. *Front Hum Neurosci, 4*, 28.

Tessitore, G., Prevete, R., Catanzariti, E., & Tamburrini, G. (2010). From motor to sensory processing in mirror neuron computational modelling. *Biol Cybern, 103*(6), 471-485.

Treue, S. (2001). Neural correlates of attention in primate visual cortex. *Trends Neurosci, 24*(5), 295-300.

Treue, S. & Maunsell, J. H. R. (2006) Feature-based attention in visual cortex. *Trends Neurosci*, 29(6), 317-322.

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. *Trends Cogn Sci, 11*(2), 58-64.

Van Overwalle, F. and Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage* 48(3): 564-84.

White, P. A., & Milne, A. (1997). Phenomenal causality: impressions of pulling in the visual perception of objects in motion. *Am J Psychol, 110*(4), 573-602.

Figure Captions

Fig. 1 Illustration of the stimuli. A) Naturalistic Grasping stimulus. B) Naturalistic Pushing stimulus. C) Abstract Grasping stimulus. D) Abstract Pushing stimulus. Discs were placed at the centers of gravity of hand and object and corrected for correct tangential contact.

Fig. 2 Illustration of the spatial manipulations of Grasping and Pushing stimuli. Frames generated from the original frame where the hand first touches the ball. A) Grasping and B) pushing action including a *Shift* manipulation, resulting in interactions without contact between hand and object. C) Grasping D) and pushing action with the *Contact point* manipulation, where the hand position was rotated by different amounts around the ball, defining incorrect contact points between fingers and object.

Fig. 3 Temporal manipulations. A) Modified stimulus with *Pause* manipulation. The contact frame is repeated (dashed line) for a variable time interval ranging from 40…200 ms. B) Stimulus with *Time Gap*. The movement of the ball in the video stream is delayed against the movement of the hand by various delays from 0…360 ms. For non-zero delay the hand moves back (yellow dashed arrow) before the ball starts to follow (green solid arrow).

Fig. 4 Model architecture. The model consists of three modules that reproduce specific properties of neurons in the visual pathway and in parietal and premotor cortex. A) Shape recognition pathway, mimicking the properties of neurons in primary visual cortex, area V4, and of shape and higher-level form and motion-selective areas, which recognize the shapes of goal object and the moving hand. B) Module that computes information about the relationship between the hand and the goal object. The relative position map encode the relative position of the hand relative to the goal object, and permits to compute the relative speed between them based on local motion detectors. C) Module containing neurons with selectivity for goal-directed movements. The action state neurons represent individual time phases of the action and link the information about the type of the hand movement, and about the spatial relationship and the relative speed of hand and object. Action neurons represent the type of the perceived action, integrating the activity over the whole time course. Their activity was compared to the obtained psychophysical data.

Fig. 5 Means of normalized ratings (N = 18) for the naturalistic and abstract stimuli without additional manipulations of the two actions. A) Ratings of the naturalness, i.e. of the fact whether the observed action represents a 'normal hand object interaction'. B) Ratings of causality i.e. whether the movement of one stimulus element (ball, disc) was caused by the other (hand or disc). Errorbars indicate standard errors (SE). Asterisks mark significant pairwise differences (uncorrected; * $p < 0.05$, *** $p < 0.001$). All ratings were significantly different ($p < 0.001$) from the midpoint value 0.5 of the normalized rating scale.

Fig. 6 Naturalness ratings for original and manipulated movements, comparing naturalistic (yellow) and abstract stimuli (blue) of grasping actions (filled circles and solid lines) and pushing movements (open circles and dashed lines). Errorbars indicate standard errors (N=18). A) Ratings for different levels of the *Shift* manipulation, where a spatial gap is present between hand and object. B) Ratings at different levels of the *Contact point* manipulation, where the hand position was rotated about the center of the ball. C) Ratings for different levels of the *Pause* manipulation, where the contact frame was repeated for different time intervals. D) Ratings for different levels of the *Time gap* manipulation, where a time delay of variable duration was introduced between the movement of the object and the hand.

Fig. 7 Causality ratings and simulation results for spatial manipulations. Left panels: Results comparing naturalistic (yellow) and abstract stimuli (blue) of grasping actions (filled circles and solid lines) and pushing movements (open circles and dashed lines). Errorbars indicate standard errors (N=18). Right panels: Normalized activation of the action neurons in the model, summed over time. This activity reproduces qualitatively many of the trends in the causality ratings. A) Ratings and for grasping actions for the *Shift* manipulation with a spatial gap between hand and object, compared with the action neuron for grasping. B) Causality ratings for pushing actions (Shift manipulation) compared with the activity of the action neurons for pushing. C) Causality ratings for different levels of the *Contact point* manipulation for grasping and corresponding simulation results. The hand position was rotated about the center of the ball. D) Results for pushing actions for the *Contact point* manipulation.

Fig. 8 Causality ratings and simulation results for temporal manipulations. Left panels: Results comparing naturalistic (yellow) and abstract stimuli (blue) of grasping actions (filled circles and solid lines) and pushing movements (opened circles and dashed lines). Errorbars indicate standard errors (N=18). Right panels: Normalized activation of the action neurons in the model, summed over time.

A) Causality ratings for grasping movements in the *Pause* manipulation, where the contact frame was repeated for different time intervals and corresponding activity of the action neurons. B) Corresponding results for the pushing action. C) Causality ratings for different levels of the *Time gap* manipulation and the related normalized responses of the model for grasping actions. D) Results for the pushing actions in the *Time gap* manipulation.

Table 1: Most important parameters of the model (Alternative numbers indicate neurons selective for grasping vs. pushing.). Further details see Fleischer et al. (2009).

Type of feature detector	Number of detectors	Receptive field size
Shape recognition hierarchy:		
Simple cells	> 3 millions	0.35° □ 0.99°
Complex cells	~ 100000	0.49° □ 1.38°
Fragment detectors	> 1.2 millions	1.5° □ 4.2°
Shape detectors	5500	4.5°
Affordance computation:		
Relative position map	~ 15000	
Affordance neurons	50	~ 4° (RPM)
Relative speed neurons	140.000	5° - 10°
Relative motion neurons	3	> 10°
Action-selective neurons:		
Action state neurons	17/30	> 10°
Action neurons	2	> 10°

Table 2: Comparisons between model predictions and human ratings. Pearson product-moment correlation coefficient (PCC) and corresponding p-values for the correlations between human ratings and the activity of the corresponding action neurons at the highest level of the model. Data is shown for the different stimulus types, action types, and manipulations. Diamonds (♦) indicate manipulations that did not significantly alter the human ratings of causality, resulting in flat curves in Figure 7 and Figure 8.

		Shift		Contact point			Pause			Time Gap	
		PCC	P	PCC	p		PCC	p		PCC	p
Naturalistic	Grasping	0.90	0.013	0.86	0.060		0.26	0.55	♦	0.95	0.004
	Pushing	0.94	0.005	0.97	0.006		0.98	0.003		0.98	<0.001
Abstract	Grasping	0.93	0.008	-0.522	0.367	♦	0	1	♦	0.96	0.003
	Pushing	0.71	0.11	0.95	0.013		0.94	0.019		0.99	<0.001

Figure 1

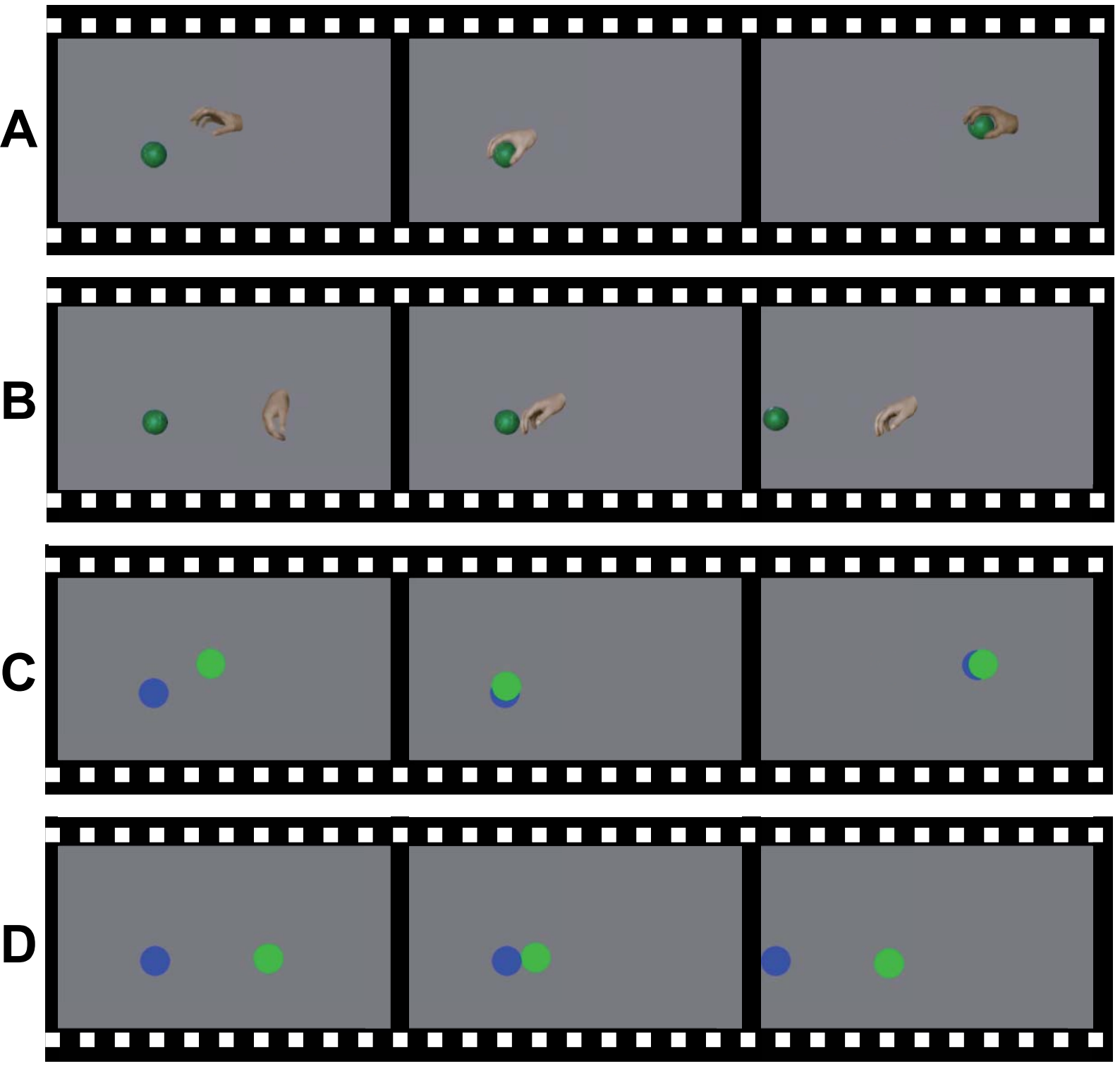

Figure 2

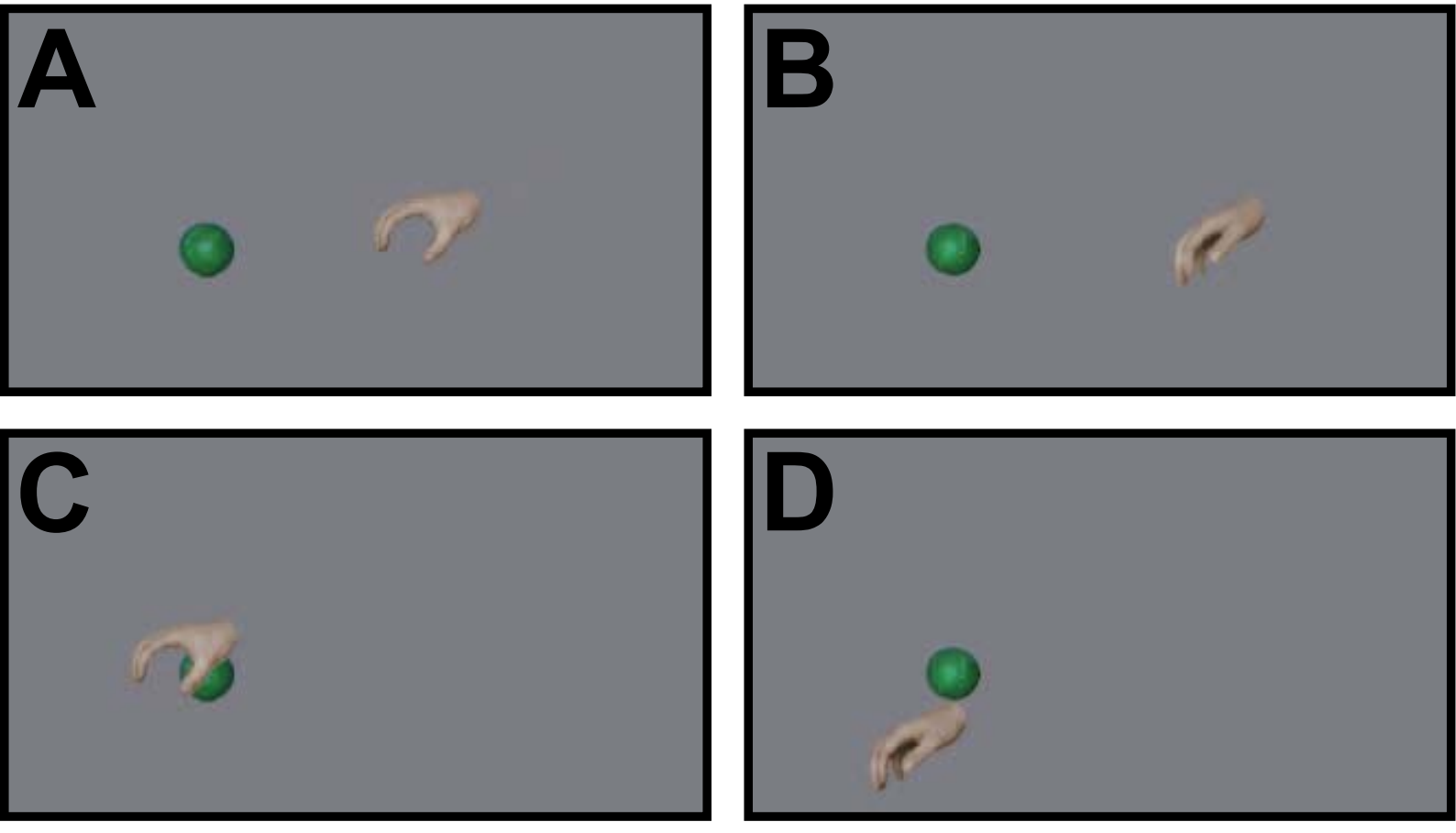

Figure 3

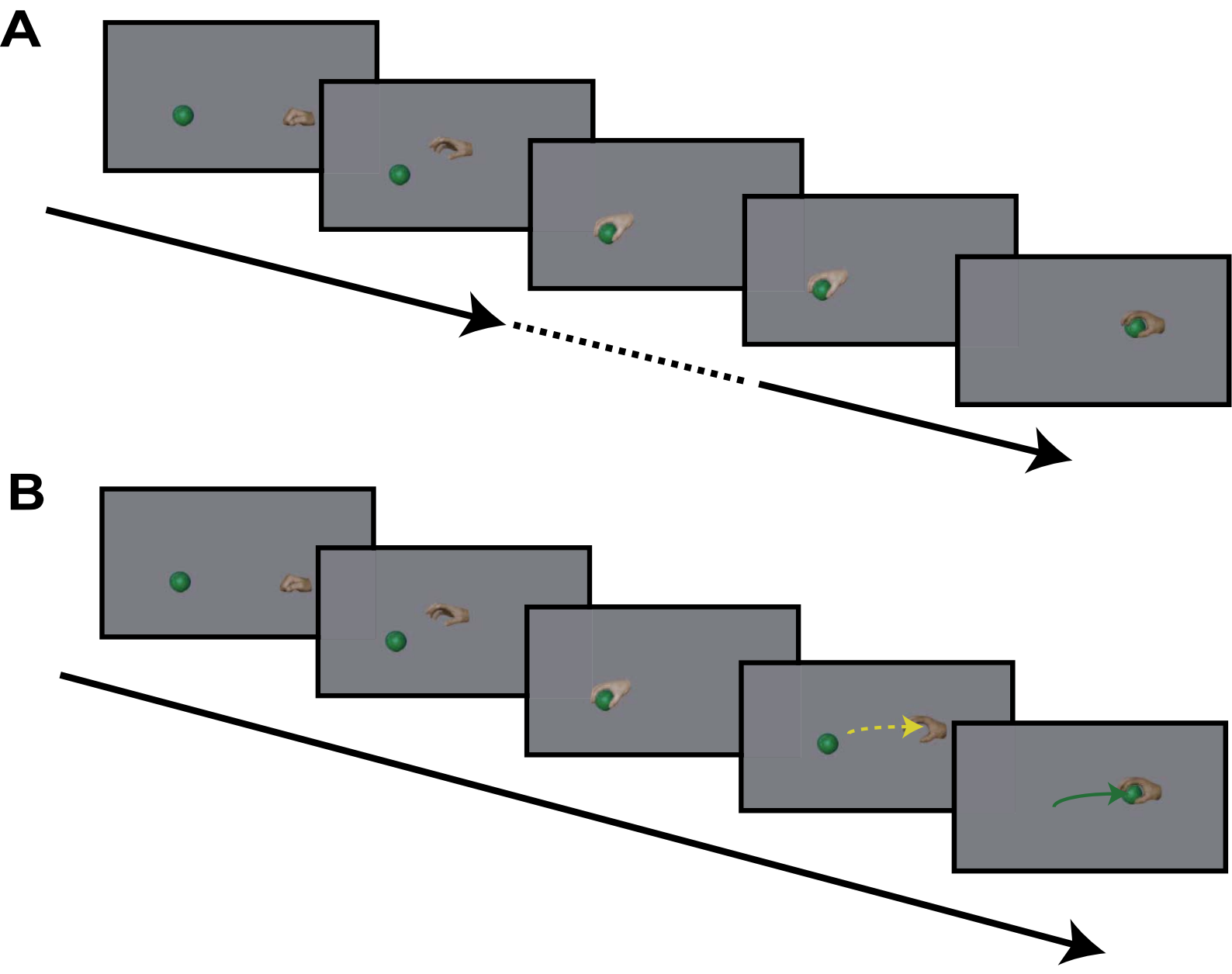

Figure 4

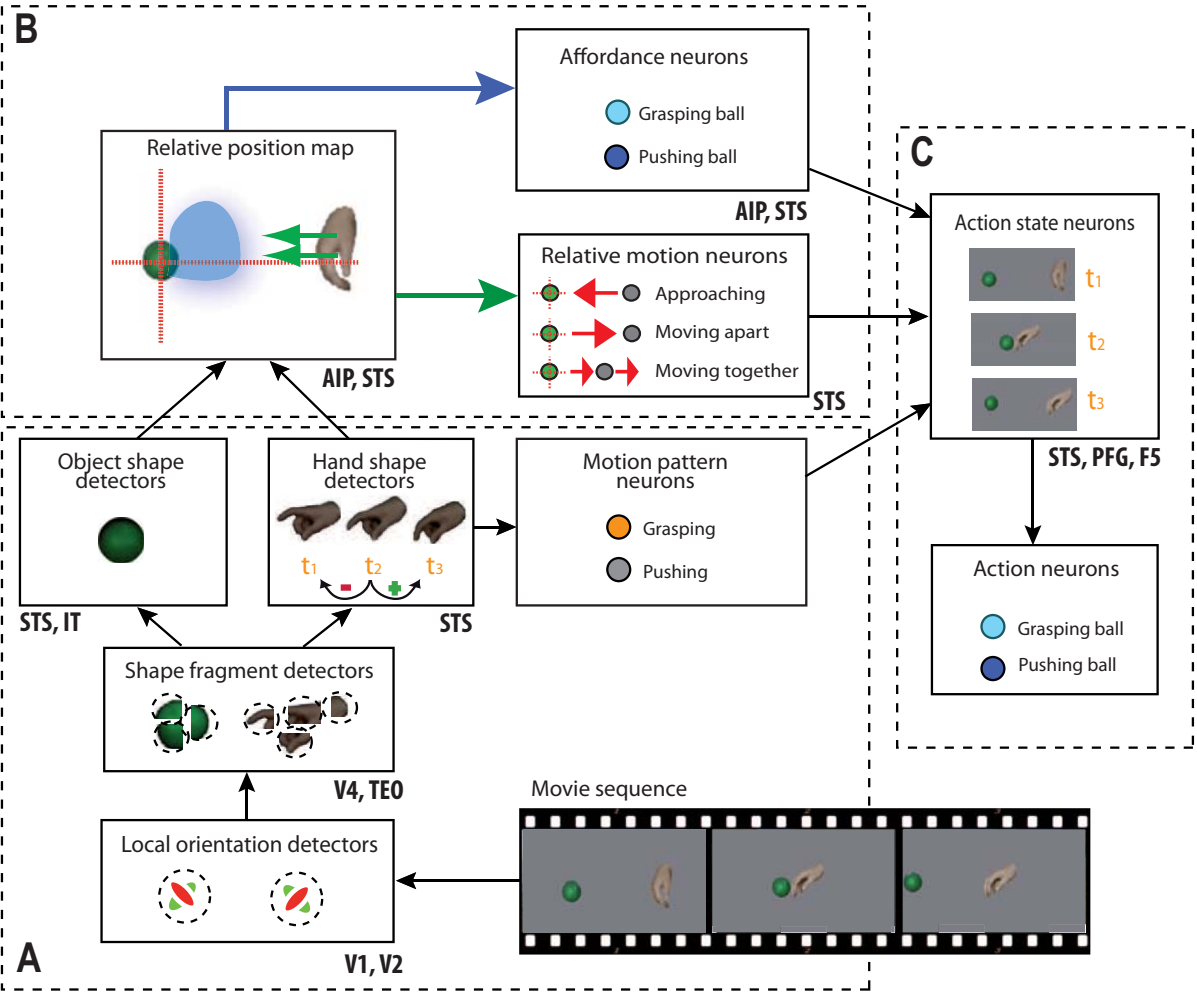
B
Affordance neurons
Grasping ball
Pushing ball
AIP, STS
Relative position map
AIP, STS
Relative motion neurons
Approaching
Moving apart
Moving together
STS
C
Action state neurons
t1
t2
t3
STS, PFG, F5
Action neurons
Grasping ball
Pushing ball
Object shape detectors
STS, IT
Hand shape detectors
t1
t2
t3
STS
Motion pattern neurons
Grasping
Pushing
Shape fragment detectors
V4, TEO
Movie sequence
Local orientation detectors
V1, V2
A

Figure 5

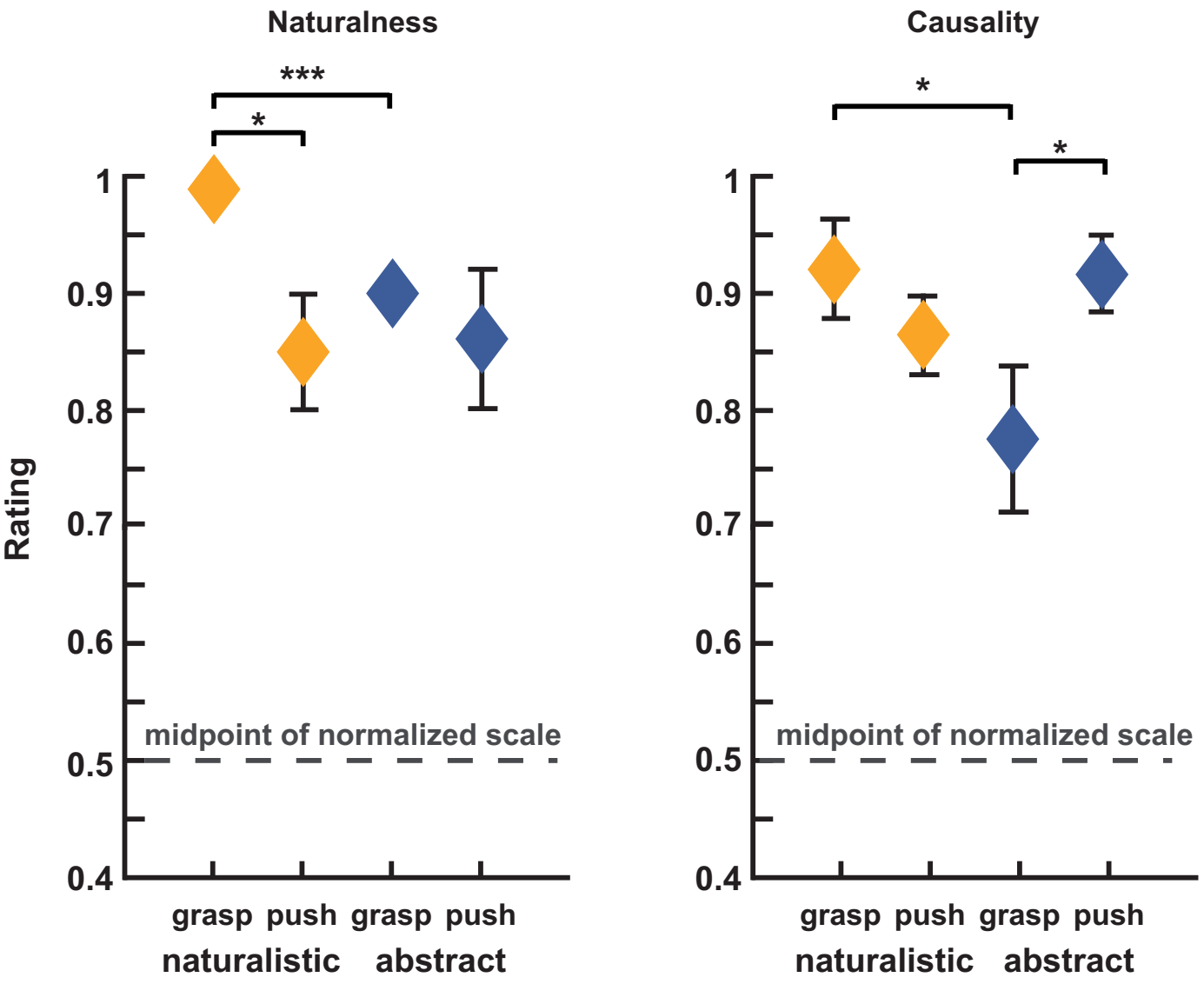
Naturalness
***
*
1
0.9
0.8
0.7
0.6
0.5
0.4
Rating
midpoint of normalized scale
grasp push grasp push
naturalistic abstract
Causality
*
*
1
0.9
0.8
0.7
0.6
0.5
0.4
midpoint of normalized scale
grasp push grasp push
naturalistic abstract

Figure 6

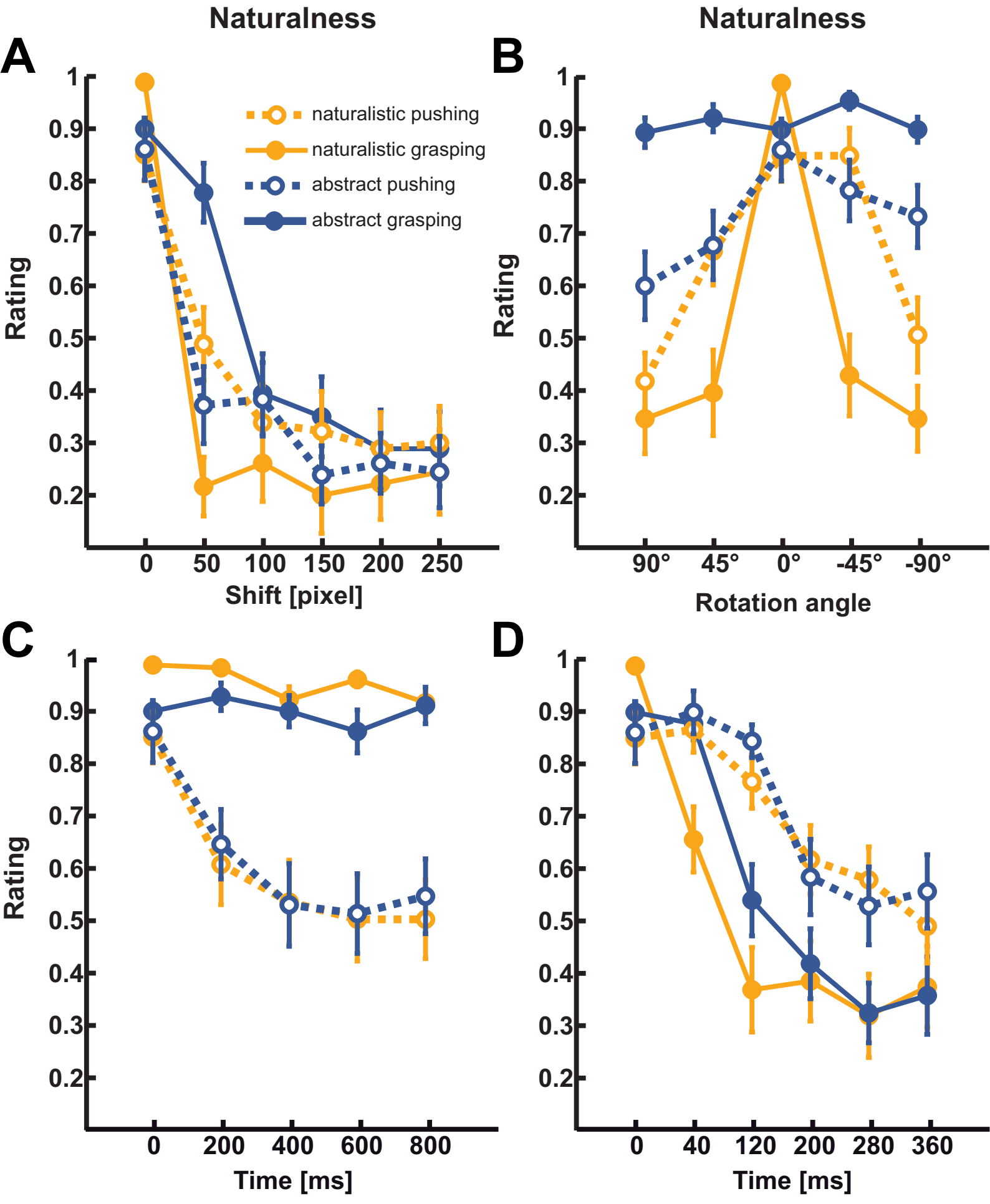
A
Naturalness
Rating
naturalistic pushing
naturalistic grasping
abstract pushing
abstract grasping
Shift [pixel]
B
Naturalness
Rating
Rotation angle
C
Rating
Time [ms]
D
Time [ms]

Figure 7

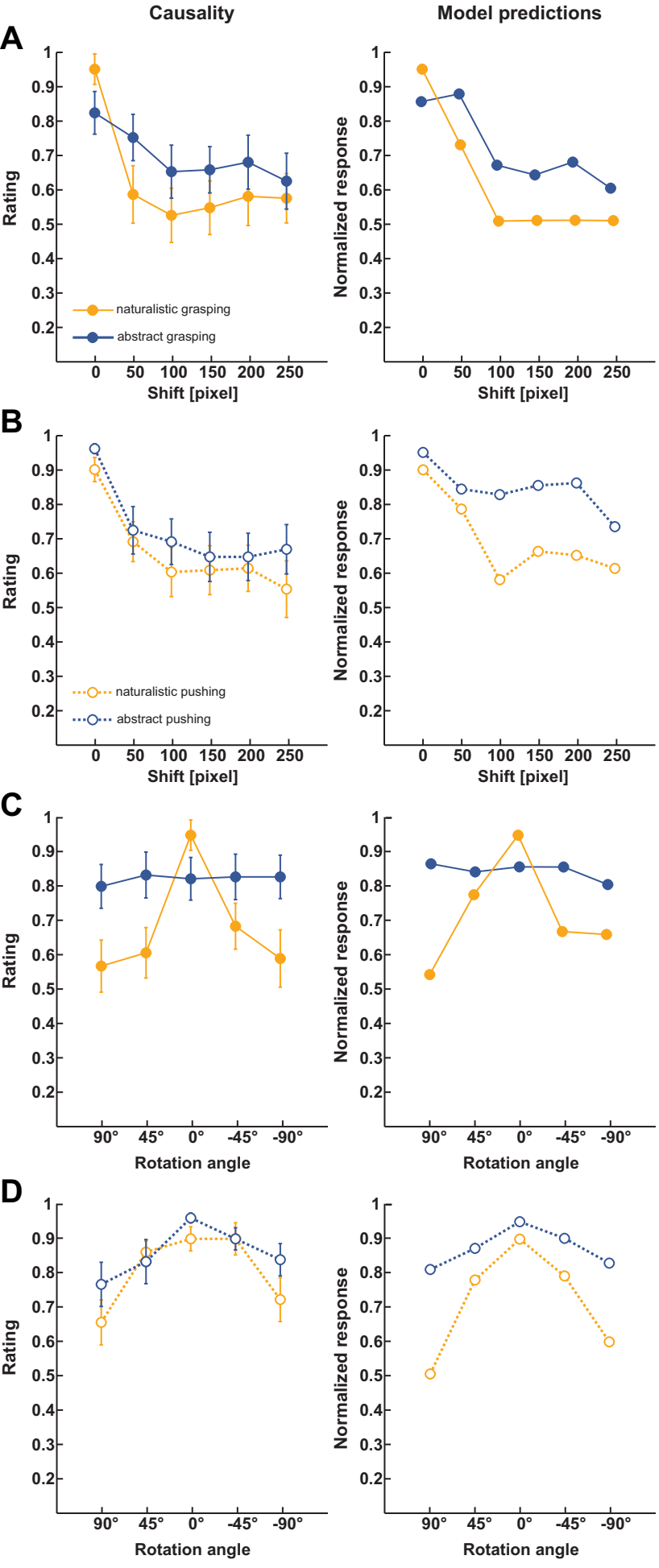
Causality
Model predictions
A
B
C
D
Rating
Normalized response
naturalistic grasping
abstract grasping
naturalistic pushing
abstract pushing
Shift [pixel]
Rotation angle
0
50
100
150
200
250
90°
45°
0°
-45°
-90°
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Figure 8

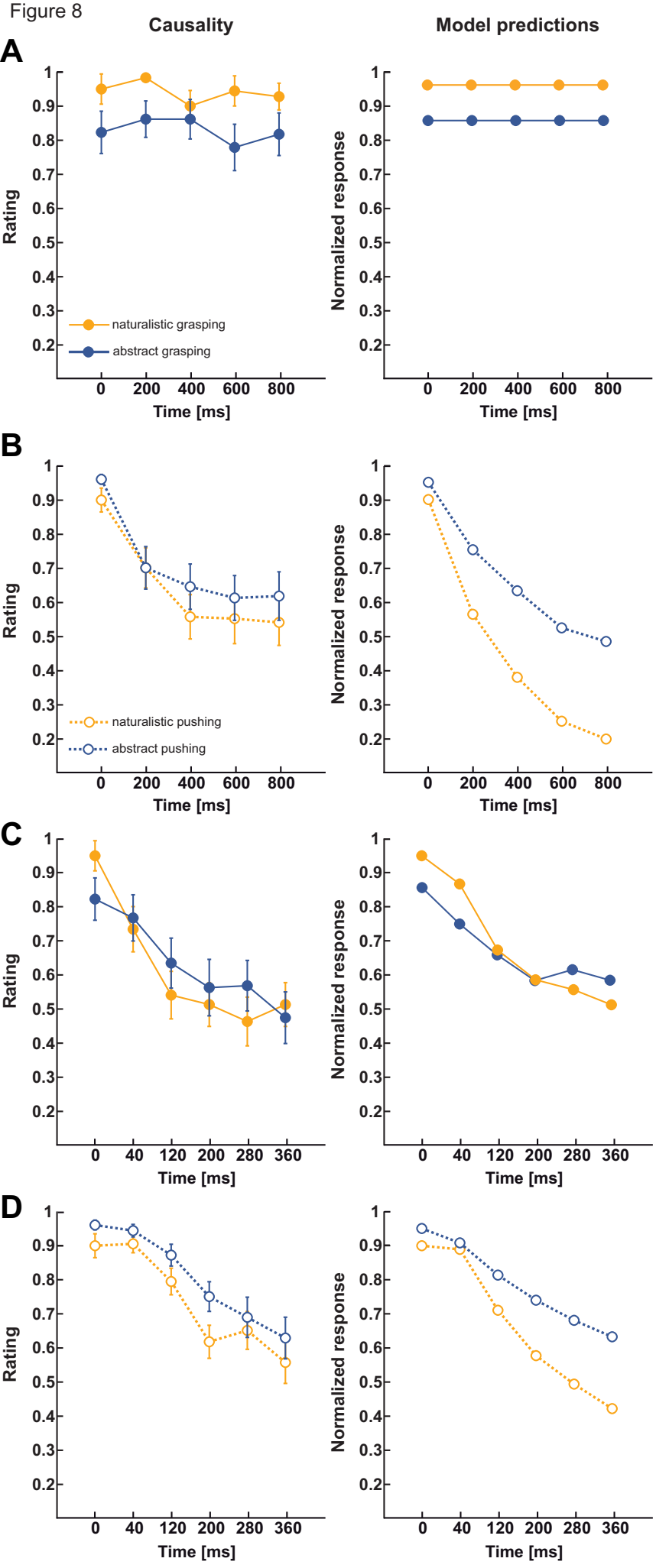
Causality
Model predictions
A
B
C
D
Rating
Normalized response
naturalistic grasping
abstract grasping
naturalistic pushing
abstract pushing
Time [ms]
0
200
400
600
800
40
120
280
360
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Table 2: Comparisons between model predictions and human ratings. Pearson product-moment correlation coefficient (PCC) and corresponding p-values for the correlations between human ratings and the activity of the corresponding action neurons at the highest level of the model. Data is shown for the different stimulus types, action types, and manipulations. Diamonds (♦) indicate manipulations that did not significantly alter the human ratings of causality, resulting in flat curves in Figure 7 and Figure 8.

		Shift		Contact point			Pause			Time Gap	
		PCC	P	PCC	p		PCC	p		PCC	p
Naturalistic	Grasping	0.90	0.013	0.86	0.060		0.26	0.55	♦	0.95	0.004
	Pushing	0.94	0.005	0.97	0.006		0.98	0.003		0.98	<0.001
Abstract	Grasping	0.93	0.008	-0.522	0.367	♦	0	1	♦	0.96	0.003
	Pushing	0.71	0.11	0.95	0.013		0.94	0.019		0.99	<0.001