Decoding the physics of observed actions in the human brain

Moritz F. Wurm[1*] & Doruk Yiğit Erigüç[1,2]

[1] CIMeC – Center for Mind/Brain Sciences, University of Trento, Rovereto, Italy
[2] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

*Corresponding author: Moritz Wurm, CIMeC – Center for Mind/Brain Sciences, University of Trento, Corso Bettini, 31, 38068 Rovereto, Italy. Phone: +39 0461-28 8632. Email: moritz.wurm@unitn.it

---

Abstract

Recognizing goal-directed actions is a computationally challenging task, requiring not only the visual analysis of body movements, but also analysis of how these movements causally impact, and thereby induce a change in, those objects targeted by an action. We tested the hypothesis that the analysis of body movements and the effects they induce relies on distinct neural representations in superior and anterior inferior parietal lobe (SPL and aIPL). In four fMRI sessions, participants observed videos of actions (e.g. breaking stick, squashing plastic bottle) along with corresponding point-light-display stick figures, pantomimes, and abstract animations of agent-object interactions (e.g. dividing or compressing a circle). Cross-decoding between actions and animations revealed that aIPL encodes abstract representations of action effect structures independent of motion and object identity. By contrast, cross-decoding between actions and point-light-displays revealed that SPL is disproportionally tuned to body movements independent of visible Interactions with objects. Lateral occipitotemporal cortex (LOTC) was sensitive to both action effects and body movements. Moreover, cross-decoding between pantomimes and animations revealed that right aIPL and LOTC represent action effects even in response to implied object interactions. These results demonstrate that parietal cortex and LOTC are tuned to physical action features, such as how body parts move in space relative to each other and how body parts interact with objects to induce a change (e.g. in position or shape/configuration). The high level of abstraction revealed by cross-decoding suggests a general neural code supporting mechanical reasoning about how entities interact with, and have effects on, each other.

Introduction

Action recognition is central for navigating social environments, as it provides the basis for understanding others' intentions, predicting future events, and social interaction. Many actions aim to induce a change in the world, often targeting inanimate objects (e.g. opening or closing a door) or persons (e.g. kissing or hitting someone). Recognizing such goal-directed actions is a computationally challenging task, as it requires not only the temporospatial processing of body movements, but also processing of how the body interacts with, and thereby induces an effect on, the object targeted by the action, e.g. a change in location, shape, or state. While a large body of

work has investigated the neural processing of observed body movements as such (Grossman et al., 2000; Giese and Poggio, 2003; Puce and Perrett, 2003; Peuskens et al., 2005; Peelen et al., 2006), the neural mechanisms underlying the analysis of action effects, and how the representations of body movements and action effects differ from each other, remain unexplored.

The recognition of action effects builds on a complex analysis of spatial and temporal relations between entities. For example, recognizing a given action as "opening a door" requires the analysis of how different objects or object parts (e.g. door and doorframe) spatially relate to each other and how these spatial relations change over time. The specific interplay of temporospatial relations is usually characteristic for an action type (e.g. opening, as opposed to closing), independent of the concrete target object (e.g. door or trash bin), and is referred to here as *action effect structure* (Fig. 1A). In addition, action effects are often independent of specific body movements – for example, we can open a door by pushing or by pulling the handle, depending on which side of the door we are standing on. This suggests that body movements and the effects they induce are at least partially processed independently from each other. Moreover, we argue that representations of action effect structures are distinct of conceptual action representations: The former capture the temporospatial structure of an object change (e.g. the separation of a closing object element), the latter capture the meaning of an action (e.g. bringing an object into an opened state to make something accessible) and can also be activated via language (e.g. by reading "she opens the box"). Previous research suggests that conceptual action knowledge is represented in left anterior LOTC (Watson et al., 2013; Lingnau and Downing, 2015; Wurm and Caramazza, 2022) whereas structural representations of action effects have not been investigated yet.

We argue that object- and movement-general representations of action effect structures are necessary for the recognition of goal-directed actions as they allow for inferring the induced effect (e.g. that something is opened) independently of specific, including novel, objects. Here we test for the existence of action effect representations that are neuroanatomically distinct from representations of body movements. We argue that both the recognition of body movements and the effects they induce rely critically on distinct but complementary subregions in parietal cortex, which is associated with visuospatial processing (Goodale and Milner, 1992; Kravitz et al., 2011), action recognition (Caspers et al., 2010), and mechanical reasoning about manipulable objects (Binkofski and Buxbaum, 2013; Leshinskaya et al., 2020) and physical events (Fischer et al., 2016; Fischer and Mahon, 2021). Specifically, we hypothesize that the neural analysis of action effects relies on anterior inferior parietal lobe (aIPL), whereas the analysis of body movement relies on superior parietal lobe (SPL). aIPL shows a representational profile that seems ideal for the processing of action effect structures at a high level of generality: Action representations in bilateral aIPL generalize across perceptually variable action exemplars, such as opening a bottle or a box (Wurm and Lingnau, 2015; Hafri et al., 2017; Vannuscorps et al., 2019), as well as structurally similar actions and object events, for example, a girl kicking a chair and a ball bouncing against a chair (Karakose-Akbiyik et al., 2023). Moreover, aIPL is critical for understanding how tools can be used to manipulate objects (Goldenberg and Spatt, 2009; Reynaud et al., 2016). More generally, aIPL belongs to a network important for physical inferences of how objects move and impact each other (Fischer et al., 2016).

Also the recognition of body movements builds on visuospatial and temporal processing, but their representation should be more specific for certain movement trajectories (e.g. pulling the arm toward the body, regardless of the movement's intent to open or close a door). The visual processing of body movements has been shown to rely on posterior superior temporal sulcus (Grossman et al., 2000; Giese and Poggio, 2003; Puce and Perrett, 2003; Peuskens et al., 2005; Peelen et al., 2006). However, recent research found that also SPL, but less so aIPL, encodes observed body movements: SPL is more sensitive in discriminating actions (e.g. a girl jumping over a box) than structurally similar object events (e.g. a ball bouncing over a box) (Karakose-Akbiyik et al., 2023); and point-light-displays (PLDs) of actions, which convey only motion-related action information but not the interactions between the body and other entities, can be decoded with higher accuracy in SPL compared to aIPL (Yargholi et al., 2023). Together, these findings support the hypothesis of distinct neural systems for the processing of observed body movements in SPL and the effect they induce in aIPL.

Using an fMRI-based cross-decoding approach (Fig. 1B), we isolated the neural substrates for the recognition of action effects and body movements in parietal cortex. Specifically, we demonstrate that aIPL encodes abstract representations of action effect structures independently of motion and object identity, whereas SPL is more tuned to body movements irrespective of visible effects on objects. Moreover, cross-decoding between pantomimes and animations revealed that right aIPL represents action effects even in response to implied object interactions. These findings elucidate the neural basis of understanding the physics of actions, which is a key stage in the processing hierarchy of action recognition.

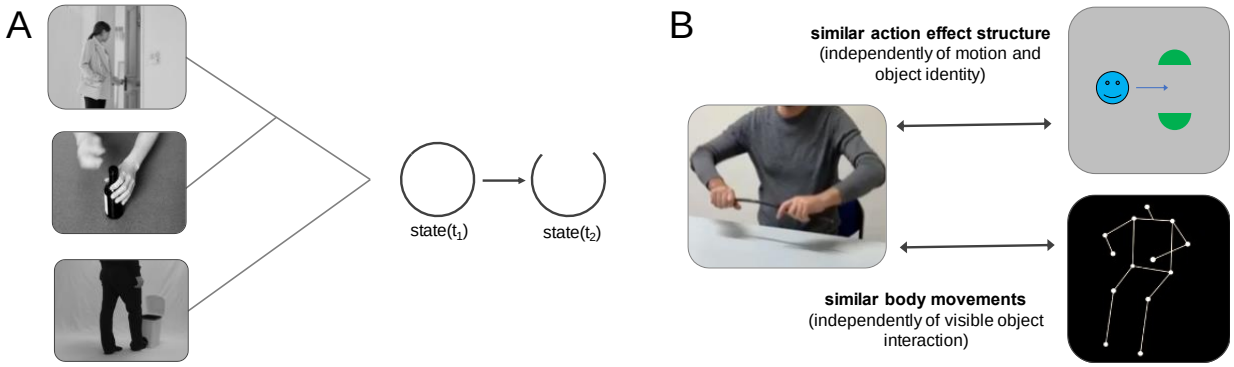

Figure 1. (A) Simplified schematic illustration of the action effect structure of "opening". Action effect structures encode the specific interplay of temporospatial object relations that are characteristic for an action type independently of the concrete object (e.g. a state change from closed to open). (B) Cross-decoding approach to isolate representations of action effect structures and body movements. Action effect structure representations were isolated by training a classifier to discriminate neural activation patterns associated with actions and testing the classifier on its ability to discriminate activation patterns associated with corresponding abstract action animations. Body movement representations were isolated by testing the classifier trained with actions on activation patterns of corresponding PLD stick figures.

Results

To isolate neural representations of action effect structures and body movements from observed actions, we used a cross-decoding approach: In 4 separate fMRI sessions, right-handed participants observed videos of actions (e.g. breaking a stick, squashing a plastic bottle) along with corresponding point-light-display stick figures, pantomimes, and abstract animations of agent-object interactions (Fig. 2) while performing a simple catch-trial-detection task (see Methods for details).

To identify neural representations of action effect structures, we first trained a classifier to discriminate the neural activation patterns associated with the action videos. Then we tested the classifier on its ability to discriminate the neural activation patterns associated with the animations. We thereby isolated the component that is shared between the naturalistic actions and the animations – the perceptually invariant action effect structure – irrespective of other action features, such as motion, object identity, and action-specific semantic information (e.g. the specific meaning of breaking a stick).

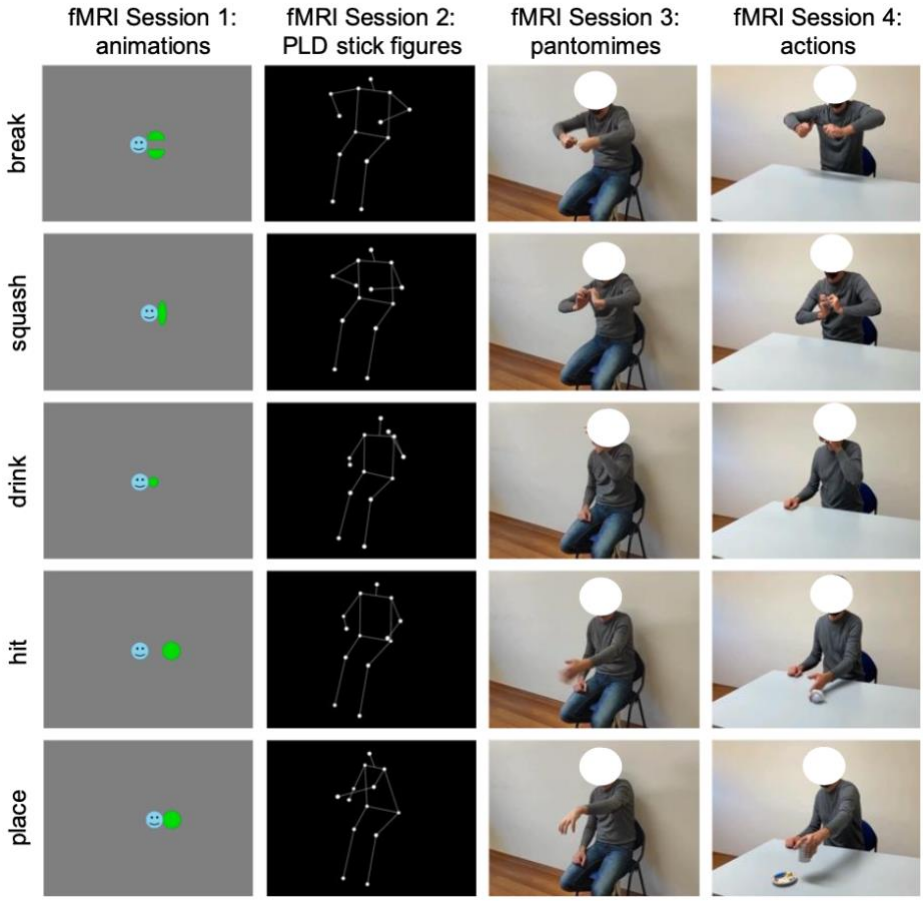

Figure 2. Experimental design. In 4 fMRI sessions, participants observed 2-second-long videos of 5 actions and corresponding animations, PLD stick figures, and pantomimes. For each stimulus type, 8 perceptually variable exemplars were used (e.g. different geometric shapes, persons, viewing angles, and left-right flipped versions of the videos). A fixed order of sessions from abstract animations to naturalistic actions was used to minimize memory and imagery effects.

Likewise, to isolate representations of body movements independently of the effect they have on target objects, we trained a classifier on naturalistic actions and tested it on the point-light-display (PLD) stick figures. We thereby isolated the component that is shared between the naturalistic actions and the PLD stick figures – the coarse body movement patterns – irrespective of action features related to the target object, such as the way they are grasped and manipulated, and the effect induced by the action.

Additionally, we used pantomimes of the actions, which are perceptually richer than the PLD stick figures and provide more fine-grained information about hand posture and movements. Thus, pantomimes allow inferring how an object is grasped and manipulated. Using cross-decoding between pantomimes and animations, we tested whether action effect representations are sensitive to implied hand-object interactions or require a visible object change.

Cross-decoding of action effect structures and body movements. We first tested whether aIPL is more sensitive in discriminating abstract representations of effect structures of actions, whereas SPL is more sensitive to body movements. Action-animation cross-decoding revealed significant decoding accuracies above chance in left aIPL but not left SPL, as well as in right aIPL and, to a lesser extent, in right SPL (Fig. 3A). Action-PLD cross-decoding revealed the opposite pattern of results, that is, significant accuracies in SPL and, to a lesser extent, in aIPL. A repeated measures ANOVA with the factors ROI (aIPL, SPL), TEST (action-animation, action-PLD), and HEMISPHERE (left, right) revealed a significant interaction between ROI and TEST ($F(1,24)=35.03$, $p=4.9E-06$), confirming the hypothesis that aIPL is more sensitive to effect structures of actions, whereas SPL is more sensitive to body movements. Post-hoc t-tests revealed that, for action-animation cross-decoding in both left and right hemispheres, decoding accuracies were higher in aIPL than in SPL (left: $t(23) = 1.81$, $p = 0.042$, right: 4.01, $p = 0.0003$, one-tailed), whereas the opposite effects were found for action-PLD cross-decoding (left: $t(23) = -4.17$ $p = 0.0002$, right: -2.93, $p = 0.0038$, one-tailed). Moreover, we found ANOVA main effects of TEST ($F(1,24)=33.08$, $p=7.4E-06$), indicating stronger decoding for action-animation vs. action-PLD cross-decoding, and of HEMISPHERE ($F(1,24)=12.75$, $p=0.0016$), indicating stronger decoding for right vs. left ROIs. An interaction between TEST and HEMISPHERE indicated that action-animation cross-decoding was significantly stronger in the right vs. left hemisphere ($F(1,24)=9.94$, $p=0.0044$).

These findings were corroborated by the results of a searchlight analysis (Fig. 3B). The whole-brain results further demonstrated the overall stronger decoding for action-PLD throughout the action observation network, which was expected because of the similarity of movement kinematics between the naturalistic actions and the PLDs. Note that we were not interested in the representation of movement kinematics in the action observation network as such, but in testing the specific hypothesis that SPL is disproportionally sensitive to movement kinematics as opposed to aIPL. Interestingly, the action-animation cross-decoding searchlight analysis revealed an additional prominent cluster in right LOTC (and to a lesser extent in left LOTC), suggesting that not only aIPL is critical for the representation of effect structures, but also right LOTC. We therefore include LOTC in the following analyses and discussion.

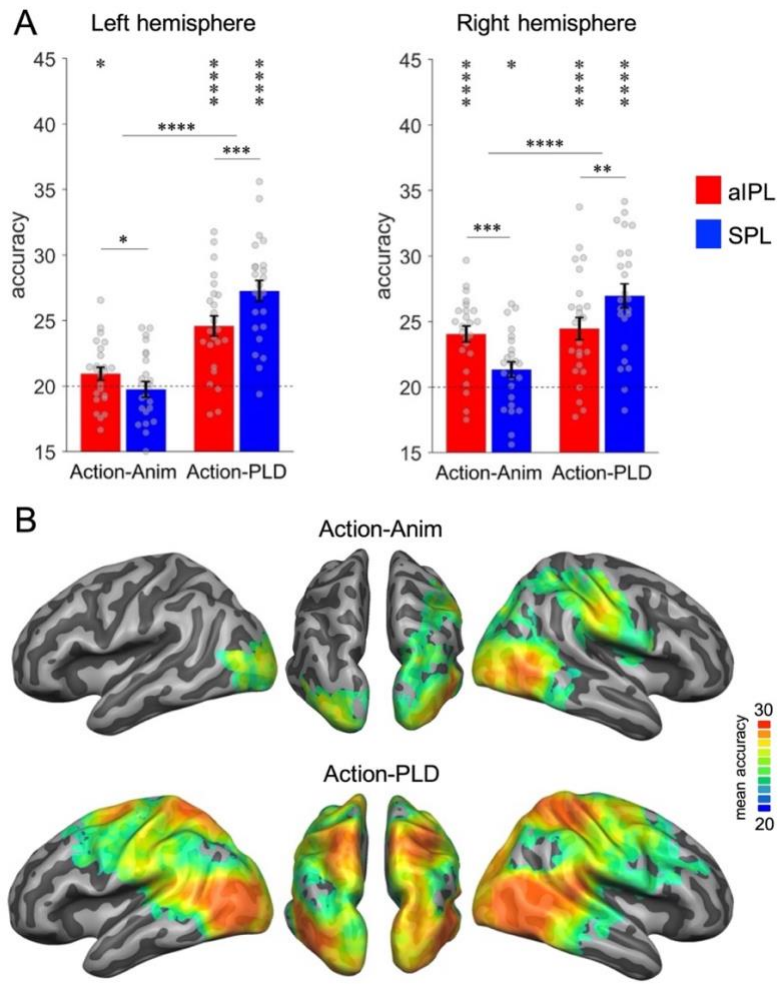

Figure 3. Cross-decoding of action effect structures (action-animation) and body movements (action-PLD). (A) ROI analysis in left and right aIPL and SPL (Brodmann Areas 40 and 7, respectively; see Methods for details). Decoding of action effect structures (action-animation cross-decoding) is stronger in aIPL than in SPL, whereas decoding of body movements (action-PLD cross-decoding) is stronger in SPL than in aIPL. Asterisks indicate FDR-corrected significant decoding accuracies above chance (* $p<0.05$, ** $p<0.01$, *** $p<0.001$, **** $p<0.0001$). Error bars indicate SEM. (B) Mean accuracy whole-brain maps thresholded using Monte Carlo correction for multiple comparisons (voxel threshold $p=0.001$, corrected cluster threshold $p=0.05$). Action-animation cross-decoding is stronger in the right hemisphere and reveals additional representations of action effect structures in LOTC.

A cluster analysis revealed that action effect representations in aIPL and LOTC formed meaningful clusters reflecting the 3 broad categories of object change types (shape/configuration changes, location changes, and ingestion), supporting the interpretation that the cross-decoding between actions and animations isolated the coarse type of action effect (Fig. S1). A cluster analysis in SPL and LOTC for body movements revealed similar representational clusters for bimanual, unimanual, and mouth-directed actions.

Taken together, these findings show that aIPL encodes abstract, that is, perceptually general representations of effect structures independent of motion and object identity (e.g. dividing or compressing object), whereas SPL encodes representations of body movements irrespective of visible interactions with objects, and that the sensitivity to effect structures is generally stronger in the right vs. left hemisphere of the brain.

Representation of implied vs. visible action effects. Humans can recognize many goal-directed actions from mere body movements, as in pantomime. This demonstrates that the brain is capable of inferring the effect that an action has on objects based on the analysis of movement kinematics without the analysis of a visible interaction with an object. Inferring action effects from body movements is easier via pantomimes than with PLD stick figures, because the former provide richer and more fine-grained body information, and in the case of object manipulations, object information (e.g. shape) implied by the pantomimed grasp. Hence, neither pantomimes nor PLDs contain visible information about objects and action effects, but this information is more easily accessible in pantomimes than in PLDs. This difference between pantomimes and PLDs allows testing whether there are brain regions that represent effect structures in the absence of visual information about objects and action effects. We tested whether aIPL is sensitive to implied effect structures by comparing the cross-decoding of actions and pantomimes (strongly implied hand-object interaction) with the cross-decoding of actions and PLDs (less implied hand-object interaction). This was the case in both aIPL and LOTC: action-pantomime cross-decoding revealed higher decoding accuracies than action-PLD cross-decoding (Fig. 4A; (all $t(23) > 3.54$, all $p < 0.0009$; one-tailed). The same pattern should be observed in the comparison of action-pantomime and pantomime-PLD cross-decoding, which was indeed the case (Fig. 3A; (all $t(23) > 2.96$, all $p < 0.0035$, ; one-tailed). These findings suggest that the representation of action effect structures in aIPL does not require a visible interaction with an object. However, the higher decoding across actions and pantomimes might also be explained by the shared information about hand posture and movements, which are not present in the PLDs. A more selective test is therefore the comparison of animation-pantomime and animation-PLD cross-decoding; as the animations do not provide any body-related information, a difference can only be explained by the stronger matching of effect structures between animations and pantomimes. We found higher cross-decoding for animation-pantomime vs. animation-PLD in right aIPL and bilateral LOTC (all $t(23) > 3.09$, all $p < 0.0025$; one-tailed), but not in left aIPL ($t(23) = 0.73$, $p = 0.23$, one-tailed). Together, this suggests that right aIPL and bilateral LOTC are sensitive to implied action effects.

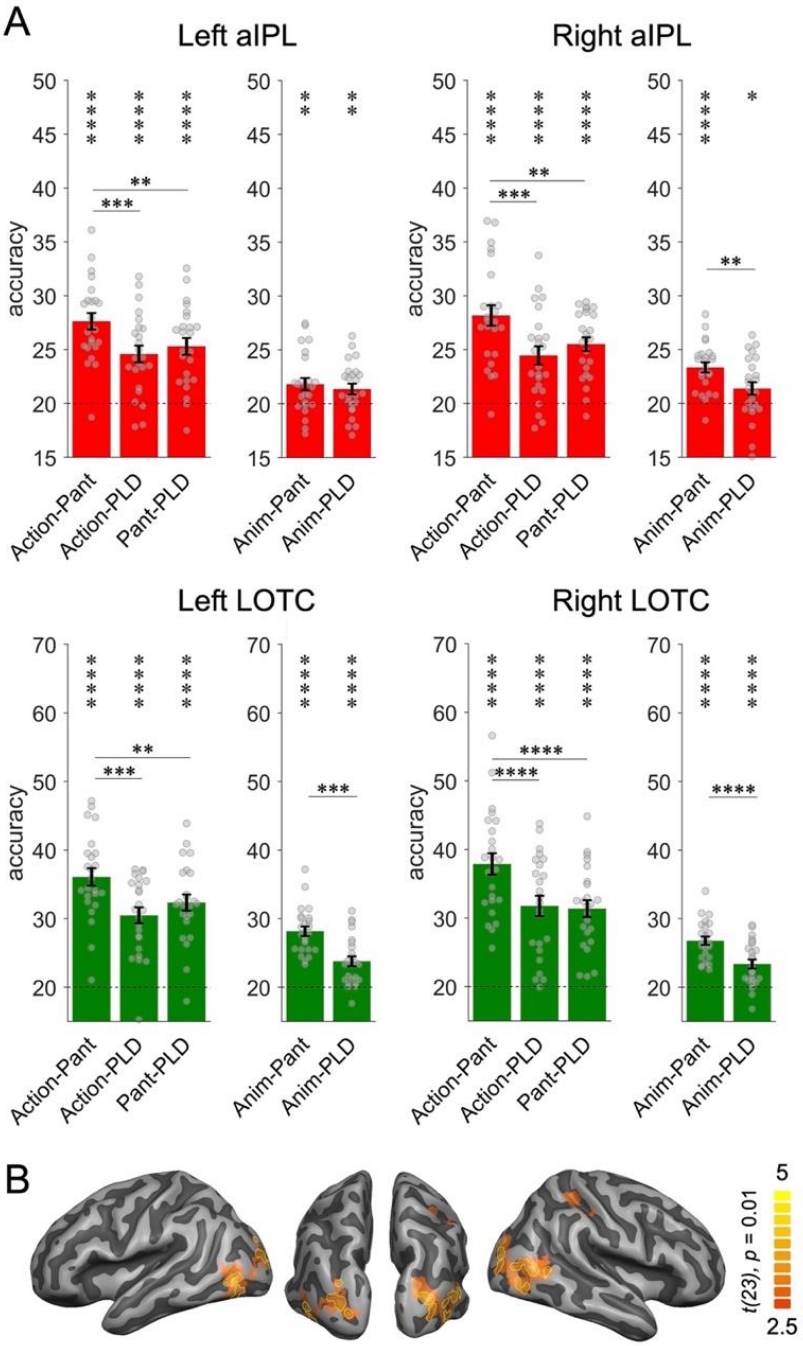

Figure 4. Cross-decoding of implied action effect structures. (A) ROI analysis. Cross-decoding schemes involving pantomimes but not PLDs (*action-pantomime, animation-pantomime*) reveal stronger effects in right aIPL than cross-decoding schemes involving PLDs (*action-PLD, pantomime-PLD, animation-PLD*), suggesting that action effect structure representations in right aIPL respond to implied object manipulations in pantomime irrespective of visuospatial processing of observable object state changes. Same conventions as in Fig. 3. (B) Conjunction of the contrasts *action-pantomime vs. action-PLD, action-pantomime vs. pantomime-PLD*, and *animation-pantomime vs. animation-PLD*. Uncorrected t-map thresholded at *p*=0.01; yellow outlines indicate clusters surviving Monte-Carlo-correction for multiple comparisons (voxel threshold p=0.001, corrected cluster threshold p=0.05).

Discussion

We provide evidence for neural representations of action effect structures in aIPL and LOTC that generalize between perceptually highly distinct stimulus types – naturalistic actions and abstract animations. The representation of effect structures in aIPL is distinct from representations of body movements in SPL. While body movement representations are generally bilateral, action effect structure representations are lateralized to the right aIPL and LOTC. In right aIPL and bilateral LOTC, action effect structure representations do not require a visible interaction with objects but also respond to action effects implied by pantomime.

Recognizing goal-directed actions requires a processing stage that captures the effect an action has on a target entity. Using cross-decoding between actions and animations, we found that aIPL – and surprisingly also LOTC – encode representations that are sensitive to the core action effect structure, that is, the type of change induced by the action. As the animations did not contain biological motion or specific object information matching the information in the action videos, these representations are independent of specific motion characteristics and object identity. This suggests an abstract level of representation of visuospatial and temporal relations between entities and their parts that may support the identification of object change independently of specific objects (e.g. dividing, compressing, ingesting, or moving something). Object-generality is an important feature as it enables the recognition of action effects on novel, unfamiliar objects. This type of representation fits the idea of a more general neural mechanism supporting mechanical reasoning about how entities interact with, and have effects on, each other (Fischer et al., 2016; Karakose-Akbiyik et al., 2023). Action effect structure representations in these regions are organized into meaningful categories, that is, change of shape/configuration, change of location, and ingestion. However, a more comprehensive investigation is needed to understand the organization of a broader range of action effect types (see also Worgotter et al., 2013), which will help to unveil the underlying principles of action structure inference.

In right aIPL and bilateral LOTC, the representation of action effect structures did not depend on a visible interaction with objects but could also be activated by pantomime, i.e., an implied interaction with objects. This suggests that right aIPL and LOTC do not merely represent temporospatial relations of entities in a perceived scene. Rather, the effects in these regions might reflect a more inferential mechanism critical for understanding hypothetical effects of an interaction on a target object.

Interestingly, action effect structures appear lateralized to the right hemisphere. This is in line with the finding that perception of cause-effect relations, e.g., estimating the effects of colliding balls, activates right aIPL (Fugelsang et al., 2005; Straube and Chatterjee, 2010). However, in the context of action recognition, the involvement of aIPL is usually bilateral or sometimes left-lateralized, in particular for actions involving an interaction with objects (Caspers et al., 2010). Also, mechanical reasoning about tools – the ability to infer the effects of tools on target objects based on the physical properties of tools and objects, such as shape, weight, etc. – is usually associated with left rather than right aIPL (Goldenberg and Spatt, 2009; Reynaud et al., 2016; Leshinskaya et al., 2020). Thus, left and right aIPL appear to be disproportionally sensitive to different structural aspects of actions and events: Left aIPL appears to be more sensitive to the type of interaction between entities, e.g. how a body part or an object exerts a force onto a target object, whereas right aIPL

appears to be more sensitive to the effect induced by that interaction. In our study, the animations contained interactions, but they did not show precisely how a force was exerted onto the target object that led to the specific effects: In all animations, the causer made contact with the target object in the same manner. Thus, the interaction could not drive the cross-decoding between actions and animations. Only the effects – the object changes – differed and could therefore be discriminated by the classification. Two questions arise from this interpretation: Would similar effects be observed in right aIPL (and LOTC) if the causer were removed, so that only the object change were shown in the animation? And would effects be observed in the left aIPL for distinguishable interactions (e.g., a triangle hitting a target object with the sharp or the flat side), perhaps even in the absence of the induced effect (dividing or compressing object, respectively)?

Action effect representations were found not only in aIPL but also LOTC. As it appears unlikely that aIPL and LOTC represent identical information, this raises the question of what different functions these regions provide in the context of action effect representation. Right LOTC is associated with the representation of socially relevant information, such as faces, body parts, and their movements (Chao et al., 1999; Pitcher and Ungerleider, 2021). Our findings suggest that right LOTC is not only sensitive to the perception of body-related information but also to body-independent information important for action recognition, such as object change. It remains to be investigated whether there is a dissociation between the action-independent representation of mere object change (e.g., in shape or location) and a higher-level representation of object change as an effect of an action. Left LOTC is sensitive to tools and effectors (Bracci and Peelen, 2013), which might point toward a role in representing putative causes of the observed object changes. Moreover, action representations in left LOTC are perceptually more invariant, as they can be activated by action verbs (Watson et al., 2013), generalize across vision and language (Wurm and Caramazza, 2019), and more generally show signatures of conceptual representation (Lingnau and Downing, 2015; Wurm and Caramazza, 2022). Thus, left LOTC might have generalized across actions and animations at a conceptual, possibly propositional level (e.g., the meaning of dividing, compressing, etc.), rather than at a structural level. Notably, conceptual action representations are typically associated with left anterior LOTC, but not right LOTC and aIPL, which argues against the interpretation that action-animation cross-decoding captured conceptual action representations only, rather than structural representations of the temporo-spatial object change type. From a more general perspective, cross-decoding between different stimulus types and formats might be a promising approach to address the fundamental question of whether the format of certain representations is propositional (Pylyshyn, 2003) or depictive (Kosslyn et al., 2006; Martin, 2016).

In contrast to the abstract representation of action effect structures in aIPL, the representation of body movements is more specific in terms of visuospatial relations between scene elements, that is, body parts. These representations were predominantly found in bilateral SPL, rather than aIPL. This is in line with previous studies demonstrating stronger decoding of PLD actions in SPL than in aIPL (Yargholi et al., 2023) and stronger decoding of human actions as opposed to object events in SPL (Karakose-Akbiyik et al., 2023). Thus, SPL seems to be particularly sensitive to specific visuospatial motion characteristics of human movements. An interesting question for future research is whether movement representation in SPL is particularly tuned to biological motion or equally to similarly complex nonbiological movements.

The distinct representations in aIPL and SPL identified here may not only play a role in the recognition of others' actions but also in the execution of goal-directed actions, which requires visuospatial processing of own body movements and of the changes in the world induced by them (Fischer and Mahon, 2021). This view is compatible with the proposal that the dorsal "where/how" stream is subdivided into sub-streams for the visuomotor coordination of body movements in SPL and the manipulation of objects in aIPL (Rizzolatti and Matelli, 2003; Binkofski and Buxbaum, 2013).

In conclusion, our study dissociated important stages in the visual processing of actions: the representation of body movements and the effects they induce in the world. These stages draw on distinct subregions in parietal cortex – SPL and aIPL – as well as LOTC. These results help clarify the roles of these regions in action understanding and more generally in understanding the physics of dynamic events. The identification of action effect structure representations in aIPL has implications for theories of action understanding (Csibra, 2007; Zentgraf et al., 2011; Kemmerer, 2021; Fischer, 2024), in particular theories that claim key roles for premotor and inferior parietal cortex in motor simulation of observed actions. The recognition of many goal-directed actions critically relies on the identification of state change of the action target, which is usually independent of body movements and thus not well accounted for by motor theories of action understanding. However, this does not rule out additional processes related to action recognition in parietal cortex, for example the generation of putative motor reactions (Orban et al., 2021) and the prediction of unfolding actions and events, which might also rely on premotor regions (Schubotz, 2007). Not all actions induce an observable change in the world. It remains to be tested whether the recognition of, e.g., communication (e.g. speaking, gesturing) and perception actions (e.g. observing, smelling) similarly relies on structural action representations in aIPL and LOTC.

Methods

Participants. Twenty-five right-handed adults (15 females; mean age, 23.7 years; age range, 20-38 years) participated in this experiment. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disease. All procedures were approved by the Ethics Committee for research involving human participants at the University of Trento, Italy.

Stimuli. The stimulus set consisted of videos of 5 object-directed actions (squashing a plastic bottle, breaking a stick, drinking water, hitting a paper ball, and placing a cup on a saucer) that were shown in 4 different formats: naturalistic actions, pantomimes, point light display (PLD) stick figures, and abstract animations (Fig. 1). The actions were selected among a set of possible actions based on two criteria: (1) The actions should be structurally different from each other as much as possible. (2) The action structures (e.g. of dividing) should be depictable as animations, but at the same time the animations should be associated with the corresponding concrete actions as little as possible to minimize activation of conceptual action representations (e.g. of "breaking a stick"). The resulting set of 5 actions belonged to 3 broad categories of changes: shape/configuration changes (break, squash), location changes (hit, place), and ingestion (drink).

For each action and stimulus format, 8 exemplars were generated to increase the perceptual variance of the stimuli. All videos were in RGB color, had a length of 2 s (30 frames per second), and a resolution of 400 x 225 pixels.

Naturalistic actions and corresponding pantomimes were performed by two different persons (female, male) sitting on a chair at a table in a neutral setting. The actions were filmed from 2 different camera viewpoints (approx. 25° and 40°). Finally, each video was mirrored to create left- and right-sided variants of the actions.

For the generation of PLD stick figures, the actions were performed in the same manner as the action videos in a motion capture lab equipped with a Qualisys motion-capture system (Qualisys AB) comprising 5 ProReflex 1000 infrared cameras (100 frames per second). 13 passive kinematic markers (14 mm diameter) were attached to the right and left shoulders, elbows, hands, hips, knees, feet and forehead of a single actor, who performed each action two times. Great care was taken that the actions were performed with the same movements as in the action and pantomime videos. 3D kinematic marker positions were processed using the Qualisys track manager and Biomotion Toolbox V2 (van Boxtel and Lu, 2013). Missing marker positions were calculated using the interpolation algorithm of the Qualisys track manager. To increase the recognizability of the body, we connected the points in the PLDs with white lines to create arms, legs, trunk, and neck. PLD stick figures were shown from two angles (25° and 40°), and the resulting videos were left-right mirrored.

Abstract animations were designed to structurally match the 5 actions in terms of the induced object change. At the same time, they were produced to be as abstract as possible so as to minimize the match at both basic perceptual levels (e.g. shape, motion) and conceptual levels. To clarify the latter, the abstract animation matching the “breaking” action was designed to be structurally similar (causing an object to divide in half) without activating a specific action meaning such as “breaking a stick”. In all animations, the agent object (a circle with a smiley) moved toward a target object (a rectangle or a circle). The contact with the target object at 1 sec after video onset induced different kinds of effects, i.e., the target object broke in half, was compressed, ingested (decreased in size until it disappeared), propelled, or pushed to the side. The animations were created in MATLAB (Mathworks) with Psychtoolbox-3 (Brainard, 1997). The speeds of all agent and target-object movements were constant across the video. To increase stimulus variance, 8 exemplars per action were generated using two target-object shapes (rectangle, circle), two color schemes for the agent-target pairs (green-blue and pink-yellow), and two action directions (left-to-right, right-to-left). To verify that animations were not associated with the specific action meanings of the naturalistic actions we performed a behavioral experiment, in which we asked 14 participants to describe what kind of actions the animations depict. No participant used verb-noun phrases (e.g. "breaking a stick") to describe the animations. Rather, the participants more abstract verbs or nouns to describe them (e.g. dividing, splitting, division; Tab. S1). These results suggest that the animations were not substantially associated with specific action meanings (e.g. "breaking a stick").

Experimental Design. For all four sessions, stimuli were presented in a mixed event-related design. In each trial, videos were followed by a 1 s fixation period. Each of the 5 conditions was presented

4 times in a block, intermixed with 3 catch trials (23 trials per block). Four blocks were presented per run, separated by 8 s fixation periods. Each run started with a 2 s fixation period and ended with a 16 s fixation period. In each run, the order of conditions was pseudorandomized to ensure that each condition followed and preceded each other condition a similar number of times in each run. Each participant was scanned in 4 sessions (animations, PLDs, pantomimes, actions), each consisting of 3 functional scans. The order of sessions was chosen to minimize the possibility that participants would associate specific actions/objects with the conditions in the animation and PLD sessions. In other words, during the first session (animations), participants were unaware that they would see human actions in the following sessions; during the second session (PLDs), they were ignorant of the specific objects and hand postures/movements. Each of the five conditions was shown 48 times (4 trials per block x 4 blocks x 3 runs) in each session. Each exemplar of every video was presented 6 times in the experiment.

Task. We used catch-trial-detection task to ensure that participants paid constant attention during the experiment and were not biased to different types of information in the various sessions. Participants were instructed to attentively watch the videos and to press a button with the right index finger on a response-button box whenever a video contained a glitch, that is, when the video did not play smoothly but jerked for a short moment (300 ms). Glitches were created by selecting a random time window of 8 video frames of the video (excluding the first 10 and last 4 frames) and shuffling the order of the frames in that window. The task was the same for all sessions. Before fMRI, participants were instructed and trained for the first session only (animations). In all four sessions, the catch trials were identified with robust accuracy (animations: 0.73±0.02 SEM, PLDs: 0.65±0.02, pantomimes: 0.69±0.02, actions: 0.68±0.02). Participants were not informed about the purpose and design of the study before the experiment.

Data acquisition. Functional and structural data were collected using a 3 T Siemens Prisma MRI scanner and a 64-channel head coil. Functional images were acquired with a T2*-weighted gradient echo-planar imaging (EPI) sequence. Acquisition parameters were a repetition time (TR) of 1.5 s, an echo time of 28 ms, a flip angle of 70°, field of view of 200 mm matrix size of 66 x 66, voxel resolution 3x3x3 mm. We acquired 45 slices in ascending interleaved odd-even order. Each slice was 3 mm thick. There were 211 volumes acquired in each functional run.

Structural T1-weighted images were acquired using an MPRAGE sequence (Slice number=176, TR=2.53 seconds, inversion time= 1.1 second, flip angle= 7°, 256*256 mm field of view, 1x1x1 mm resolution).

Preprocessing. Data were analyzed using BrainVoyager QX 2.84 (BrainInnovation) in combination with the SPM12 and NeuroElf (BVQXTools) toolboxes and custom software written in Matlab (MathWorks). Anatomical scans of individual subjects were normalized to the standard SPM12 EPI template (Montreal Neurological Institute MNI stereotactic space). Slice time correction was performed on the functional data followed by a three-dimensional (3D) motion correction (trilinear interpolation, with the first volume of the first run of each participant as reference).

Functional data were co-registered with the normalized anatomical scans followed by spatial smoothing with a Gaussian kernel of 8 mm full width at half maximum (FWHM) for univariate analysis and 3 mm FWHM for MVPA.

Multivariate pattern classification. For each participant, session, and run, a general linear model (GLM) was computed using design matrices containing 10 action predictors (2 for each action; based on 8 trials from the first 2 blocks of a run and the second half from the last 2 blocks of a run, to increase the number of beta samples for classification), a catch-trial predictor, 6 predictors for each parameter of motion correction (three-dimensional translation and rotation) and 6 temporal-drift predictors. Each trial was modeled as an epoch lasting from video onset to offset (2s). The resulting reference time courses were used to fit the signal time courses of each voxel. Predictors were convolved with a dual-gamma hemodynamic impulse response function. Since there were 3 runs per session, there were thus 6 beta maps per action condition.

Searchlight classification was done on each subject in volume space using a searchlight sphere of 12 mm and an LDA (linear discriminant analysis) classifier, as implemented in the CosmoMVPA toolbox (Oosterhof et al., 2016). We also tested for robustness of effects across MVPA parameter choices by running the analysis with different ROI sizes (9 mm, 15 mm) and a support vector machine (SVM) classifier, which revealed similar findings, that is, all critical findings were also found with the alternative MVPA parameters.

For the within-session analyses, all 5 action conditions of a given session were entered into a 5-way multiclass classification using leave-one-out cross validation, that is, the classifier was trained with 5 out of 6 beta patterns per action and was tested with the held-out beta pattern. This was done until each beta pattern was tested. The resulting accuracies were averaged across the 6 iterations and assigned to the center voxel of the sphere (see Fig. S2 for results of within-session decoding). For cross-decoding analyses, a classifier was trained to discriminate the voxel activation patterns associated with the 5 action conditions from one session (e.g. actions) and tested on its ability to discriminate the voxel activation patterns associated with the 5 action conditions of another session (e.g. animations). The same was done vice versa (training with animations and testing with actions), and the accuracies were averaged across the two directions (see Fig. S3 for contrasts between cross-decoding directions). In total, there were 6 across-session pairs: action-animation, action-PLD, action-pantomime, animation-pantomime, pantomime-PLD, animation-PLD.

For all decoding schemes, a one-tailed, one-sample t-test was performed on the resulting accuracy maps to determine which voxels had a decoding accuracy that was significantly above the chance level (20%). The resulting t-maps were corrected for multiple comparisons with Monte Carlo Correction as implemented in CosmoMVPA (Oosterhof et al., 2016), using an initial threshold of $p=0.001$ at the voxel level, 10000 Monte Carlo simulations, and a one-tailed corrected cluster threshold of $p = 0.05$ ($z = 1.65$).

Conjunction maps were computed by selecting the minimal z-value (for corrected maps) or t-value (for uncorrected maps) for each voxel of the input maps.

ROI analysis. Regions of interest (ROI) were based on MNI coordinates of the center locations of Brodmann areas (BA) associated with the aIPL/supramarginal gyrus (BA 40; Left: -53, -32, 33; right: 51, -33, 34), SPL (BA 7; left: -18, -61, 55; right: 23, -60, 61), and LOTC (BA 19; left: -45, -75, 11; right: 44, -75, 5), using the MNI2TAL application of the BioImage Suite WebApp (https://bioimagesuiteweb.github.io/webapp/). For each participant, ROI, and decoding scheme, decoding accuracies from the searchlight analysis were extracted from all voxels within a sphere of 12 mm around the ROI center, averaged across voxels, and entered into one-tailed, one-sample t-tests against chance (20%). In addition, paired t-tests and repeated measures analyses of variance (ANOVA) were conducted to test for the differences between ROIs and different decoding schemes. The statistical results of t-tests were FDR-corrected for the number of tests and ROIs (Benjamini and Yekutieli, 2001).

Acknowledgements

We thank Seoyoung Lee for assistance in preparing the video stimuli, Ingmar de Vries for assistance in preparing the PLD stimuli, Ben Timberlake for proof-reading and the Caramazza Lab for helpful feedback on the study design and interpretation of results. This research was supported by the Caritro Foundation, Italy.

Author Contributions

Conceptualization: MFW, Methodology: YE and MFW, Investigation: YE, Software: YE and MFW, Formal Analysis: YE and MFW, Writing - Original Draft: MFW, Writing - Final Draft: MFW, Supervision: MFW.

References

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188.

Binkofski F, Buxbaum LJ (2013) Two action systems in the human brain. Brain Lang 127:222-229.

Bracci S, Peelen MV (2013) Body and object effectors: the organization of object representations in high-level visual cortex reflects body-object interactions. J Neurosci 33:18247-18258.

Brainard DH (1997) The Psychophysics Toolbox. Spatial vision 10:433-436.

Caspers S, Zilles K, Laird AR, Eickhoff SB (2010) ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50:1148-1167.

Chao LL, Haxby JV, Martin A (1999) Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. Nat Neurosci 2:913-919.

Csibra G (2007) Action mirroring and action interpretation: An alternative account. In: P. Haggard, Y. Rosetti, & M. Kawato (Eds.), Sensorimotor Foundations of Higher Cognition. Attention and Performance XXII (pp. 435-459). Oxford: Oxford University Press. Oxford: Oxford University Press:435-459.

Fischer J (2024) Physical reasoning is the missing link between action goals and kinematics A comment on "An active inference model of hierarchical action understanding, learning, and imitation" by Proietti et al. Physics of Life Reviews 48:198-200.
Fischer J, Mahon BZ (2021) What tool representation, intuitive physics, and action have in common: The brain's first-person physics engine. Cognitive neuropsychology 38:455-467.
Fischer J, Mikhael JG, Tenenbaum JB, Kanwisher N (2016) Functional neuroanatomy of intuitive physical inference. Proc Natl Acad Sci U S A 113:E5072-5081.
Fugelsang JA, Roser ME, Corballis PM, Gazzaniga MS, Dunbar KN (2005) Brain mechanisms underlying perceptual causality. Brain Res Cogn Brain Res 24:41-47.
Giese MA, Poggio T (2003) Neural mechanisms for the recognition of biological movements. Nat Rev Neurosci 4:179-192.
Goldenberg G, Spatt J (2009) The neural basis of tool use. Brain 132:1645-1655.
Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci 15:20-25.
Grossman E, Donnelly M, Price R, Pickens D, Morgan V, Neighbor G, Blake R (2000) Brain areas involved in perception of biological motion. J Cogn Neurosci 12:711-720.
Hafri A, Trueswell JC, Epstein RA (2017) Neural Representations of Observed Actions Generalize across Static and Dynamic Visual Input. J Neurosci 37:3056-3071.
Karakose-Akbiyik S, Caramazza A, Wurm MF (2023) A shared neural code for the physics of actions and object events. Nat Commun 14:3316.
Kemmerer D (2021) What modulates the Mirror Neuron System during action observation?: Multiple factors involving the action, the actor, the observer, the relationship between actor and observer, and the context. Progress in Neurobiology:102128.
Kosslyn SM, Thompson WL, Ganis G (2006) The case for mental imagery: Oxford University Press.
Kravitz DJ, Saleem KS, Baker CI, Mishkin M (2011) A new neural framework for visuospatial processing. Nat Rev Neurosci 12:217-230.
Leshinskaya A, Wurm MF, Caramazza A (2020) Concepts of actions and their objects. M Gazzaniga, GR Mangun, D Poepped The Cognitive Neurosciences:757-765.
Lingnau A, Downing PE (2015) The lateral occipitotemporal cortex in action. Trends Cogn Sci 19:268-277.
Martin A (2016) GRAPES-Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. Psychonomic bulletin & review 23:979-990.
Oosterhof NN, Connolly AC, Haxby JV (2016) CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Front Neuroinform 10:27.
Orban GA, Lanzilotto M, Bonini L (2021) From Observed Action Identity to Social Affordances. Trends Cogn Sci.
Peelen MV, Wiggett AJ, Downing PE (2006) Patterns of fMRI activity dissociate overlapping functional brain areas that respond to biological motion. Neuron 49:815-822.
Peuskens H, Vanrie J, Verfaillie K, Orban GA (2005) Specificity of regions processing biological motion. Eur J Neurosci 21:2864-2875.
Pitcher D, Ungerleider LG (2021) Evidence for a Third Visual Pathway Specialized for Social Perception. Trends Cogn Sci 25:100-110.

Puce A, Perrett D (2003) Electrophysiology and brain imaging of biological motion. Philos Trans R Soc Lond B Biol Sci 358:435-445.

Pylyshyn Z (2003) Return of the mental image: are there really pictures in the brain? Trends in Cognitive Sciences 7:113-118.

Reynaud E, Lesourd M, Navarro J, Osiurak F (2016) On the neurocognitive origins of human tool use : A critical review of neuroimaging data. Neurosci Biobehav Rev 64:421-437.

Rizzolatti G, Matelli M (2003) Two different streams form the dorsal visual system: anatomy and functions. Exp Brain Res 153:146-157.

Schubotz RI (2007) Prediction of external events with our motor system: towards a new framework. Trends Cogn Sci 11:211-218.

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

van Boxtel JJ, Lu H (2013) A biological motion toolbox for reading, displaying, and manipulating motion capture data in research settings. J Vis 13.

Vannuscorps G, M FW, Striem-Amit E, Caramazza A (2019) Large-Scale Organization of the Hand Action Observation Network in Individuals Born Without Hands. Cereb Cortex 29:3434-3444.

Watson CE, Cardillo ER, Ianni GR, Chatterjee A (2013) Action concepts in the brain: an activation likelihood estimation meta-analysis. J Cogn Neurosci 25:1191-1205.

Worgotter F, Aksoy EE, Kruger N, Piater J, Ude A, Tamosiunaite M (2013) A Simple Ontology of Manipulation Actions Based on Hand-Object Relations. Ieee T Auton Ment De 5:117-134.

Wurm MF, Lingnau A (2015) Decoding actions at different levels of abstraction. J Neurosci 35:7727-7735.

Wurm MF, Caramazza A (2019) Distinct roles of temporal and frontoparietal cortex in representing actions across vision and language. Nat Commun 10:289.

Wurm MF, Caramazza A (2022) Two 'what' pathways for action and object recognition. Trends Cogn Sci 26:103-116.

Yargholi E, Hossein-Zadeh GA, Vaziri-Pashkam M (2023) Two distinct networks containing position-tolerant representations of actions in the human brain. Cereb Cortex 33:1462-1475.

Zentgraf K, Munzert J, Bischoff M, Newman-Norlund RD (2011) Simulation during observation of human actions--theories, empirical studies, applications. Vision Res 51:827-835.

Supplementary Information

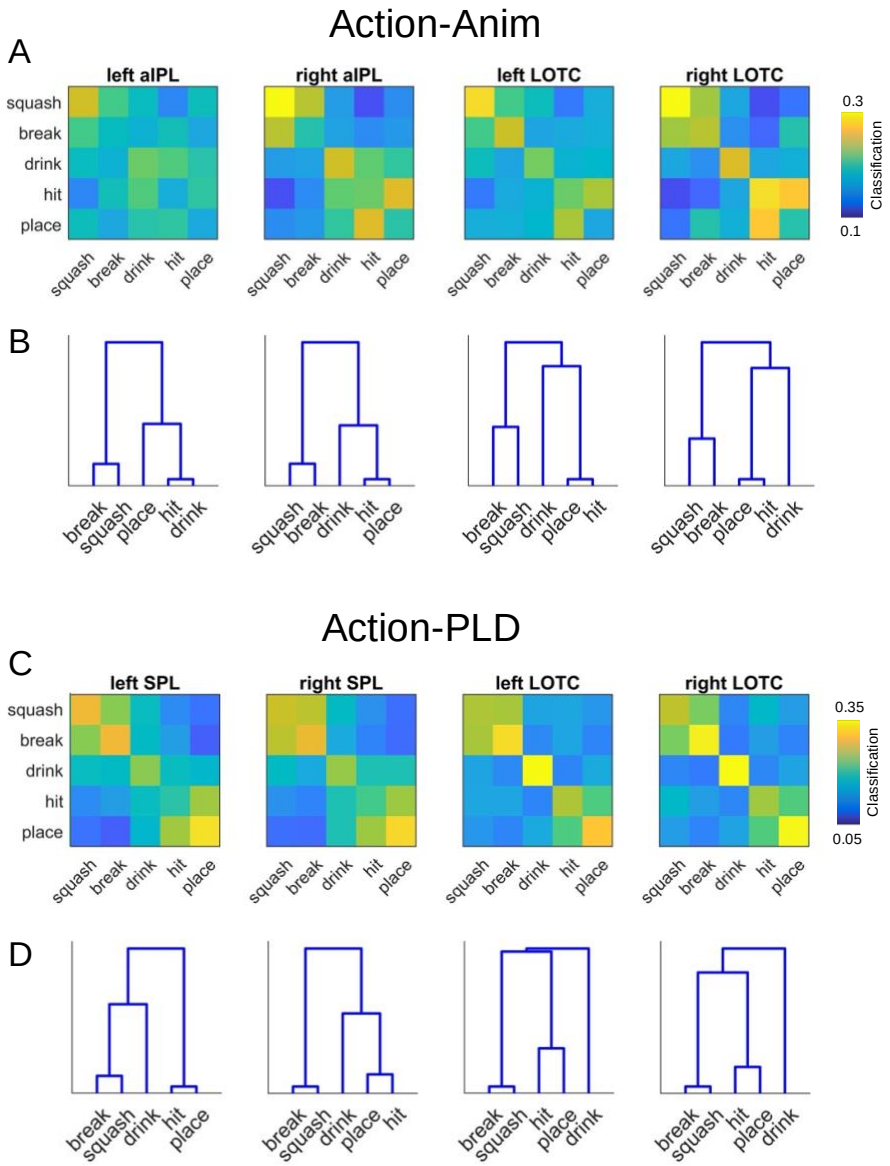

Fig. S1. Representational similarity of action effect structures and body movements. Pairwise classifications of the 5 actions from the action-animation cross-decoding (A) and the action-PLD cross-decoding (C) were extracted from each ROI, averaged across voxels, and entered into a cluster analysis using average distance. The resulting hierarchical cluster trees are displayed as dendrograms (B, D). In aIPL and LOTC, action effect structure representations formed meaningful clusters reflecting the 3 broad categories of change types: object shape/configuration changes (break, squash), location changes (hit, place), and ingestion (drink), supporting the interpretation that the cross-decoding between actions and animations isolated the coarse type of action effect. The cluster analysis in SPL and LOTC for body movements revealed similar representational clusters, which probably reflect categories of body movements, that is, bimanual actions (break, squash), unimanual actions (hit, place), and drinking as a mouth-directed action.

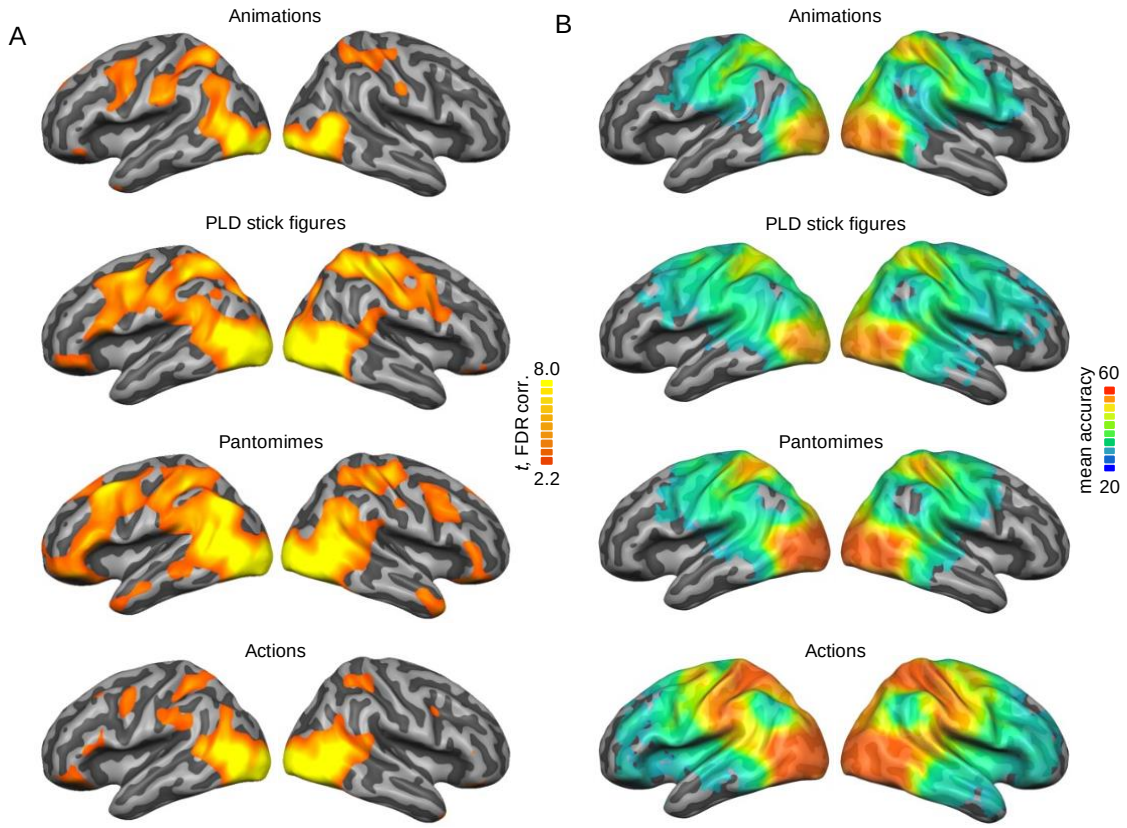

Fig. S2. (A) Univariate activation maps for each session (all 5 actions vs. Baseline; FDR-corrected at $p = 0.05$) and (B) within-session decoding maps (Monte-Carlo-corrected for multiple comparisons; voxel threshold p=0.001, corrected cluster threshold p=0.05).

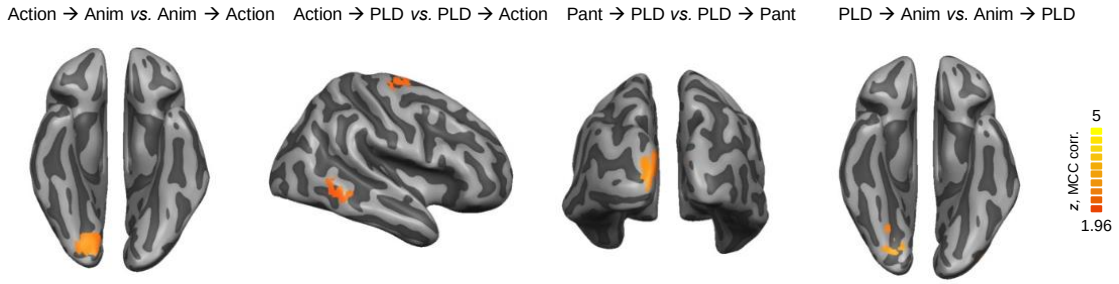

Fig. S3. Direction-specific cross-decoding effects. To test whether there were differences between the two directions in the cross-decoding analyses, we ran, for each of the 6 across-session decoding schemes, two-tailed paired samples t-tests between the decoding maps of one direction (e.g. action → animation) vs. the other direction (animation → action). Direction effects were observed in left early visual cortex for the directions action → animation, PLD → animation, and pantomime → PLD, as well in right middle temporal gyrus and dorsal premotor cortex for action → PLD. These effects do not appear to affect the interpretation of direction-averaged cross-decoding effects in the main text. Monte-Carlo-corrected for multiple comparisons; voxel threshold p=0.001, corrected cluster threshold p=0.05.

Table S1: Results of the behavioral pilot experiment. Verbal descriptions of the abstract animations by N=14 participants.

Subject	break	hit	ingest	move	squash
1	splitting	hitting	swallowing	moving	hitting
2	dividing	explosion	shrinking	passing	colliding
3	break	kinetic energy	suction	touch	crash
4	break	press	eating	press	push
5	divide	crash	swallow	push	turn
6	divide	crash	sucking	pushing	squeeze
7	break	accident	consume	push	hit
8	splitting in two	pushing and splitting	shrinking	slowly pushing	pushing
9	destroying	banging	introjection	hitting	squishing
10	division	division	absorbing	pushing	impacting
11	dividing	force	absorption	action reaction	squeezing
12	divide	collide	absorb	activate	squeeze
13	crash	collision	eating	push	push
14	separation	cooperation	absorption	cause and effect	force