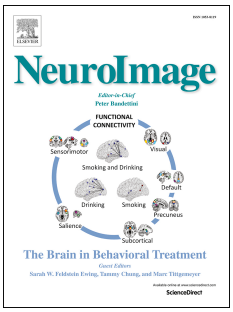
Accepted Manuscript



Decoding intentions of self and others from fMRI activity patterns

Sam J. Gilbert, Hoki Fung

PII: S1053-8119(17)31114-X

DOI: 10.1016/j.neuroimage.2017.12.090

Reference: YNIMG 14605

To appear in: *NeuroImage*

Received Date: 14 September 2017

Revised Date: 28 November 2017

Accepted Date: 28 December 2017

Please cite this article as: Gilbert, S.J., Fung, H., Decoding intentions of self and others from fMRI activity patterns, *NeuroImage* (2018), doi: 10.1016/j.neuroimage.2017.12.090.

Decoding intentions of self and others from fMRI activity patterns

Sam J. Gilbert and Hoki Fung

Institute of Cognitive Neuroscience, University College London, UK

Keywords: intentions, MVPA, fMRI

Acknowledgements:
SJG was supported by a Royal Society University Research Fellowship. We are grateful to Annika Boldt for helpful comments on an earlier version of this article.

Address correspondence to:

Sam Gilbert
Institute of Cognitive Neuroscience
17 Queen Square
London WC1N 3AR
UK
Email: sam.gilbert@ucl.ac.uk
Tel: +44 (0)20 7679 1121
Fax: +44 (0)20 7813 2835

Previous studies using multi-voxel pattern analysis have decoded the content of participants' delayed intentions from patterns of fMRI data. Here we investigate whether this technique can be used to decode not only participants' own intentions, but also their representation of the intentions held by other people. In other words: if Sam is thinking about Hoki, can we decode the content of Hoki's intention by scanning Sam's brain? We additionally distinguished two components of intentions: action-plans versus goals, and included novel control analyses that allowed us to distinguish intending an outcome from simply expecting it to occur or simulating its consequences. Regions of frontal, parietal, and occipital cortex contained patterns from which it was possible to decode intentions of both self and other. Furthermore, crossclasification between self and other was possible, suggesting overlap between the two. Control analyses suggested that these results reflected visuo-spatial processes by which intentions were generated in our paradigm, rather than anything special about intentions per se. There was no evidence for any representation of intentions as mental states distinct from visuospatial processes involved in generating their content and/or simulating their outcomes. These findings suggest that the brain activity patterns decoded in intention-decoding fMRI studies may reflect domain-general processes rather than being intention-specific.

1. Introduction

What are intentions, and how do they relate to patterns of brain activity? The term intention is used in various ways in everyday language and is "notoriously difficult" to define (Pacherie and Haggard, 2010). Nevertheless, recent neuroimaging studies have investigated their underlying brain mechanisms, including several using pattern classification techniques as a method for "decoding intentions". In this study we attempt to extend these findings by asking whether we can decode participants' representations of another person's intentions as well as their own. We also examine some of the different ways in which intentions may be defined, and examine how these may relate to fMRI decoding results. In other words, we aim to explore the following question: when intentions are decoded from fMRI activity, what exactly is it that is being decoded?

1.1 Types of intentions

Intentions are conscious mental states that bear some relation to subsequent action (Bratman, 1987; Mele, 1992; Pacherie, 2008; Pacherie and Haggard, 2010; Searle, 1983). They may be subdivided according to various factors such as their temporal proximity to action. For example, philosophers have distinguished prior intentions versus intentions-in-action (Searle, 1983); future-directed versus present-directed intentions (Bratman, 1987); prospective versus immediate intentions (Brand, 1984); and distal versus proximal intentions (Mele, 1992). Pacherie (2008) proposes an "intentional cascade" between 1) distal intentions, associated with initial deliberation and planning; 2) proximal intentions, which develop an action plan within a specific context; and 3) motor intentions, which are involved in motor guidance and control during the execution of overt movements.

In the psychology literature, memory for intentions has been investigated in the field of "prospective memory" (PM) research (Brandimonte et al., 1996; Kliegel et al., 2008). Researchers in this field have also made various theoretical distinctions. For example some authors distinguish "vigilance", where an intention is consciously rehearsed over a brief period, versus prospective memory "proper", where the intention must be brought to mind at the appropriate time (Graf and Uttl, 2001). Others distinguish the different ways in which intentions can be cued (McDaniel and Einstein, 2007), such as event-based PM (intending to do something when a particular cue occurs), time-based PM (intending to do something at a particular time) and activity-based PM (intending to do something after completing a particular activity). A further distinction within event-based PM (McDaniel and Einstein, 2000) is between "focal" tasks (where the individual will already be attending to the cue for intended action as part of an ongoing activity) versus "nonfocal" tasks (where the ongoing task does not direct attention towards the cue).

Clearly, the term "intention" has a rich set of meanings and is used in a variety of different ways. This complicates any attempt to investigate putative neural correlates. It could be the case that phenomena related to the concept of intention are so diverse at the level of brain function as to make this concept unsuitable as a target for neuroscientific investigation (Uithol et al., 2014). According to this view, searching for the neural correlates of intention would be like searching for the neural correlates of "thinking" or "believing": the concepts have such "wildly disjunctive" meanings (Fodor, 1974) as to resist interpretation in terms of a limited set of neural correlates. Alternatively, it may be the case that a core set of brain systems play a key role across a broad range of situations described in terms of intention. Deciding between these views will require careful consideration of how neuroimaging paradigms relate to specific cognitive processes.

Insofar as intentions can be linked to particular brain processes, a further question is the extent to which these processes are intention-specific or domain-general. On one hand, intentions might be seen as special type of mental state, with distinct neural correlates. For example, some artificial intelligence-inspired models of the cognitive system such as SOAR (Laird, 2012; Newell, 1990) and ACT-R (Anderson, 1996; Anderson et al., 2004) posit a system called the "goal stack" that plays a unique cognitive role. According to these models, the goal stack is structurally distinct from other systems (i.e. it has unique computational properties). Such models might predict a unique pattern of brain activity associated with intentions. On the other hand, other models (Altmann and Trafton, 2002) analyse the concept of intention wholly in terms of domain-general processes such as memory, priming, and so on. Models of this type would not predict intention-specific patterns of brain activity that could not be observed in the context of domain-general cognitive processes.

1.2 Neuroimaging and intention decoding

Initial PET studies investigating neural correlates of intentions focused on comparing brain activity between situations where participants held pending intentions while performing an additional "ongoing" task versus situations where they simply performed the ongoing task by itself. These studies revealed brain regions showing increased (Burgess et al., 2001; Okuda et al., 1998) and decreased (Burgess et al., 2003) activity associated with remembering delayed intentions. Methodologically, several of these studies used an approach of "multiple task averaging", where two or more tasks were investigated in order to detect patterns of intention-related signal change that were relatively invariant to the precise cognitive operations required by specific tasks (Burgess

et al., 2003, 2001; Gilbert et al., 2009; Simons et al., 2006). Despite wide variety in the intentions and tasks used in these studies, results from these paradigms revealed surprisingly consistent patterns of signal change, particularly in regions of rostral prefrontal cortex (reviewed by Burgess et al., 2011; Cona et al., 2015). Subsequent studies characterised these effects in further detail, for example extending results from event-based to time-based paradigms (Oksanen et al., 2014; Okuda et al., 2007), investigating subtypes of event-based tasks (Gilbert et al., 2009; Simons et al., 2006), distinguishing activity related to encoding, storage, and retrieval of intentions (Gilbert et al., 2012) and examining whether the roles of regions showing activation increases versus decreases could be distinguished (Landsiedel and Gilbert, 2015).

More recently, studies have begun to use neuroimaging to investigate the content of intentions. That is, rather than seeking to identify differences in brain activity between having any intention versus no intention, these studies have attempted to identify differences in brain activity between two specific intentions (Gilbert, 2011; Haynes et al., 2007; Momennejad and Haynes, 2012, 2013; Soon et al., 2013, 2008; Wisniewski et al., 2016). This represents an inversion of the multiple task averaging logic of earlier studies. Whereas multiple task averaging was used in an attempt to identify brain activity that is relatively invariant from one intention to another, recent studies have complemented this by seeking to detect intention-specific patterns of brain activity. These studies have typically used multivariate decoding approaches (Haynes and Rees, 2006; Norman et al., 2006), whereby pattern classifiers are trained on examples of activity associated with two categories of trial (e.g. two different intentions), and their ability to correctly classify novel data is tested.

For example, Haynes et al. (2007) decoded participants' prospective intentions to add or subtract a pair of to-be-presented numbers, by analysing patterns of activity in medial prefrontal cortex. Subsequent studies showed that it was possible to decode participants' free choices to make one of two motor responses (Soon et al., 2008) or perform one of two abstract tasks (Soon et al., 2013), up to 10 seconds before they executed those responses. Gilbert (2011) decoded participants' intentions to make a special response if they saw a particular stimulus in the future, with distinct brain regions containing patterns that predicted A) the type of stimulus participants were expecting, and B) the specific response they were preparing to make when they saw it. Momennejad and Haynes (2012, 2013) decoded participants' intentions to execute high/low or odd/even discriminations once they encountered subsequently-presented digits, with significant effects observed in various medial and lateral prefrontal regions.

1.3 Components of intentions

The studies reviewed above indicate that different intentions can be associated with specific patterns of brain activity, allowing them to be decoded using pattern classification techniques. But what exactly do these different patterns of brain activity reflect? There are various dimensions along which intentions can vary, potentially in an orthogonal manner, making it unclear what exactly the classifiers are detecting. For example, consider a hypothetical pattern classifier that can decode whether a person intends to warm or cool a room, using a thermostat. Below we consider some of the dimensions that the classifier might be sensitive to.

A. Goals or outcomes. An obvious way in which two intentions can differ from one another is in the goals or outcomes they aim towards. Our classifier might be distinguishing representations of a desired goal of warmth or coolness.

B. Action-plans. A second dimension on which intentions can differ is in the action or task one intends to perform in order to achieve a goal. Suppose that the thermostat is operated digitally with a left or right button to cool or warm the room, respectively. In this case, a person who intends to cool the room may also intend to press the left button. If the button mappings were reversed, an intention to cool the room may also be an intention to press the right button. Therefore, intended goals and intended action-plans can vary independently of one another. This distinction between goals and action plans has been recognised in the distinction made between "goal intentions", which specify an intended outcome, and "implementation intentions", which specify a means of bringing that outcome about (Gilbert et al., 2009; Gollwitzer, 1999). Similarly, discussions of the brain's so-called mirror system have noted that "there must be a clear distinction between goals and the motor routines that are implemented in a given circumstance to achieve those goals" (Hickok, 2008, p. 1241).

The distinction between goals and action-plans may be interpreted in terms of a multi-level hierarchy (e.g. Cooper and Shallice, 2000), whereby a higher-level goal (e.g. to fill a glass with water) is decomposed in terms of progressively more specific subgoals and subplans (e.g. to turn the tap, by executing a particular type of reaching movement with the arm, adopting a particular kind of grip with the hand, etc.), eventually leading to specific patterns of muscle activity. As noted by Tomasello et al. (2005), a goal such as to fill a glass with water can itself be seen as an action-plan to satisfy a higher-level goal (e.g. to quench thirst). Thus, "what is a goal when viewed from beneath is a means when

viewed from above" (p. 677). Accordingly, the distinction between goals and action-plans does not necessarily identify two dichotomous kinds of representation. Rather, describing intentions in terms of distinct goals and action-plans can highlight multiple hierarchically-organised representations that play a role in the control of action.

C. Reasons. A third dimension on which intentions can differ is the reason for forming that intention to begin with. Intentions can be seen as "the primary link between reasons and actions" (Mele, 1992, p.3). Thus, a currently hot room may motivate an individual to form an intention to cool it down, but a currently cool room may motivate the opposite intention. Our hypothetical classifier might be sensitive to current room temperature or the reasoning involved in comparing the current temperature to the desired temperature in order to infer whether the appropriate action would be to warm or cool the room.

D. Expectations. Another possibility is that our classifier is decoding an expectation that the room will warm up or cool down, rather than the intention to bring this about. Expectations and intentions are intimately related. Indeed, in formal models of "active inference" the distinction between intending an outcome and expecting it to occur can collapse (Friston et al., 2011; Kilner et al., 2007). Nevertheless, intention and expectation have distinct meanings in ordinary language. For example, someone might intend to execute a tricky basketball shot without expecting that they are likely to be successful (Mele, 1992). Conversely, it is possible to expect an outcome without intending it to occur. For example, someone who failed to prepare for a test might expect to fail it without intending to.

The expectations associated with intentions could relate to outcomes (in which case they may overlap with goals), or the expected actions by which those goals might be achieved

(in which case they would overlap with action-plans). They might also relate to other features associated with goals and action-plans such as the expected level of effort associated with an intended action. For example, in the Haynes et al. (2007) study addition may have been less effortful than subtraction, and therefore the classifier may have decoded an expected level of cognitive control that would subsequently be required, rather than any specific representation of arithmetical operations. Another possibility would be that expectations represent the rewards that can be obtained if an intention is fulfilled (see Wisniewski et al., 2015 for discussion).

E. Commitment. One crucial difference between intending an outcome and merely expecting it to occur is that in the former case, but not the latter, the agent has a disposition to act in order to bring it about. Thus, along with their representational aspects (goals and action-plans), intentions also have a motivational aspect (Brand, 1984). Without this motivational aspect, one may have a "predictive awareness" (Mele, 2009) that something is about to occur (e.g. a sneeze) without intention. In the example of our hypothetical pattern classifier, it is unclear whether it is sensitive to a prediction that a particular outcome will occur, a commitment to that outcome, or both. One way to address this issue is to investigate whether a pattern classifier trained to distinguish two particular intentions also makes correct predictions when tested on states of the world associated with those intentions (e.g. stimuli associated with expected perceptual outcomes) in the absence of any intention to bring those states about. Insofar as such a cross-classification is possible, this implies that the classifier is sensitive to the predictive aspect of intentions.

1.4 Aims of the present study

The foregoing discussion demonstrates that if a pattern classifier can distinguish between a person intending X versus Y, it is not always certain exactly what it is distinguishing. Perhaps the classifier is sensitive to brain representations of the goals attached to the two intentions, or their associated action-plans. It may be sensitive to an expectation of the intentions' perceptual consequences, the required level of cognitive control, or the state of the world that rationally justifies forming one or the other intention to begin with. Perhaps it is sensitive to two or more of these factors. One aim of the present study is to apply an intention decoding methodology to a paradigm that allows us to disentangle some of these possibilities, by orthogonally manipulating distinct intention components (e.g. goals versus action-plans) and training separate pattern classifiers to decode these components.

A second aim of this study is to investigate not only whether we can decode the intentions of the scanned participant, but also their representation of another person's intention. A large body of research has identified brain regions involved in thinking about the mental states of other people, particularly those associated with the brain's "mentalizing system" (Frith and Frith, 2012; Kennedy and Adolphs, 2012). For example, many studies show increased activation of regions such as medial prefrontal cortex and temporo-parietal junction when participants think about the goals and intentions of other people (Van Overwalle and Baetens, 2009). These studies might be considered analogous to univariate investigations of first-person intentions which compare performance of an ongoing task by itself with performance of the same task while also remembering a delayed intention (e.g. Burgess et al., 2003, 2001). In other words, they compare thinking about someone's intention versus not thinking about their intention. Here, we seek to extend this by searching for patterns of brain activity that distinguish thinking that an agent has intention X versus thinking that they have intention Y. To our

knowledge, this type of 'meta-decoding' has not previously been attempted, i.e. decoding the scanned participant's mental decoding of another agent's intention.

As well as training separate classifiers to decode the scanned-participant's own intention and their representation of another agent's intention, we also investigated whether we could "cross-classify" between the two. That is, we investigated whether a classifier trained to decode the scanned participant's intention could also perform successfully when tested on the other agent's intention, and vice versa. This would imply some overlap between representation of our own mental states and those of other agents, as suggested by "simulation" theories of mindreading. These theories posit that we understand the mental states of other agents – at least in part – by simulating those mental states within our own cognitive systems (Apperly, 2008; Carruthers and Smith, 1996; Gordon, 1986; Heal, 1986; Ramnani and Miall, 2004).

Finally, as discussed above, as well as investigating cross-classification between 1st-person and 3rd-person intentions, we also investigated cross-classification between intentions and states of the world associated with those intentions. Insofar as classifiers trained to decode intentions could also decode these non-intended outcomes, this would suggest that the classifiers are sensitive to information that is not specific to intentions but could also apply to expectations for what is about to occur, simulation of their perceptual consequences, and so on. To our knowledge, this is the first study to compare brain activity associated with intending an outcome versus merely expecting it.

We scanned participants using fMRI as they performed a task alongside another agent, orthogonally manipulating the scanned-participant's own intention and the intention of the other agent. In one condition, participants believed that the other agent was another

person they had met prior to the scanning session, in another they believed that the other agent was a computer. This allowed us to investigate whether putative brain representations of another agent's intention are specific to thinking about people, or also apply to thinking about other types of agent. Some previous studies indicate patterns of brain activation, particularly involving medial prefrontal cortex, that apply selectively when participants believe they are interacting with a person rather than a computer (Gallagher et al., 2002; Gilbert et al., 2007). However other studies, particularly those investigating the mirror system, show no difference between interacting with a person versus a robot (Cross et al., 2012; Gazzola et al., 2007).

2. Methods

2.1 Participants

24 right-handed participants attended an initial behavioural training session followed by a MRI scanning session 1-4 days later (mean age: 24.4 years; range 19-36; 10 males). A further three participants took part in the behavioural training setting but were not invited back for MRI scanning due to poor performance of the behavioural task. All participants provided written informed consent before taking part and the study was approved by the UCL Research Ethics Committee (1584/002).

2.2 Behavioural task

Participants took part in a collaborative task with a partner. They were told that they were playing a game together. For half of the trials the participants were told that their partner was another person outside the scanner; in the other half they were told that it

was a computer program. There were two roles in this task: the experimenter or computer program always took the role of player A, who acted first on each trial. The experimental participant took the role of player B, who acted second on each trial. A schematic representation of this task is shown in Figure 1 and the sequence of events on each trial is shown in Figure 2.

Each trial began with the players viewing a picture of two vertical pipes, side by side, with a ball above one of them. Initially, two sections of the pipes were missing. Subsequently, the ball began to fall downwards and the players' task was to fill in the missing sections of the pipes to guide it towards goal locations (left or right). During this trajectory there were three "switch points" where the ball could either switch from one pipe to the other (in which case the pipes could be seen crossing over) or stay in the same pipe (in which case the pipes continued straight down). Player A's task was to configure the first switch point (i.e. switch versus stay) and player B's task was to configure the third switch point. The second switch point was already shown in a switch or stay configuration from the beginning of the trial and was not under control of either player.

Before the pipes were filled in, the first switch point was filled with either the name 'Hoki' (when player A was the experimenter) or 'Computer' (when player A was the computer program). The third switch point was filled with the participant's first name. The colour of these names indicated goal locations for the two players (left versus right). These locations determined each player's goal immediately after the switch point that they controlled. Therefore, player A's goal referred the ball's location immediately after the first switch point and player B's goal referred to the location immediately after the third switch point. The players' names could be shown in four possible colours: red,

green, blue, or orange. Two of these colours indicated left and two indicated right, so that there were two pairs of colours which indicated left versus right. One pair of colours was used on odd-numbered trials and the other was used on even-numbered trials. By cueing each position with two different colours, we were able to decode fMRI patterns corresponding to each position unconfounded with the presentation of particular colours (see below).

[Figure 1 about here]

Each trial began with an initial period of 'thinking time', where the players could form intentions for future behaviour (i.e. their decisions to set their respective switch points to switch or stay configurations). It is only this period of the trial that will be examined in our neuroimaging analyses. Suppose that a trial begins as follows: 1) ball above the left pipe; 2) player A's name in a colour indicating the right pipe as a goal, 3) the second switch point set to a switch configuration, 4) player B's name in a colour indicating the left pipe as a goal. The full crossing of these four factors defined the 16 possible trial types in this task (see Table 1 for a list). From the starting point described above, it is possible to reason as follows: given that the ball's starting point is on the left but player A's goal is on the right, player A should form an intention to set the first switch point to a *switch* configuration. Following this, the ball will pass through the second switch point, which is also set to a *switch* configuration, so it will transfer the ball back to the left pipe. Therefore, given that player B's goal is for the ball to be on the left, this player should form an intention to set the third switch point to a *stay* configuration. In this way, player B forms the appropriate *stay* intention, and in doing so s/he considered six pieces of information: 1) player A's goal (right); 2) player A's action-plan (switch); 3) player B's

goal (left); 4) player B's action-plan (stay); 5) the ball's initial position (left); 6) configuration of the second switch point (switch).

[Figure 2 about here]

[Table 1 about here]

Importantly, in our experimental design these six variables are perfectly orthogonal to each other across the 16 trial types. This allowed us to investigate patterns of brain activity that were sensitive to each of these six variables in turn, without being confounded with any of the other variables. Although the two action-plans could be deduced by thinking about the four other pieces of information (as described in the example above), the correlation between each action-plan and the other five variables, across the 16 trial types, was zero (Table 2). Formally, predicting the action-plans from the other factors is a linearly inseparable problem (Duda et al., 2000). This means that a linear classifier such as the one used in the present study would by definition be unable perform better than chance in predicting action-plans from the other variables.

[Table 2 about here]

Following the thinking time at the beginning of each trial two images were shown to the left and right of the first switch point, one showing pipes in a switch configuration and the other showing pipes in a stay configuration. This was the cue for player A to press the left or right button to configure the first switch point. As soon as this occurred, the ball began falling continuously down the pipes and player B was presented with switch and stay options either side of the third switch point. Player B then configured the third switch point in anticipation of the arrival of the ball. Finally when the ball reached the

bottom of the pipes, feedback was provided in the form of a flashing dot shown under player B's goal location. If player B's switch point was configured correctly, this was shown in green; otherwise a black outline was shown. Whenever switch / stay options were presented on the screen they were randomly assigned to left versus right. This meant that during the initial thinking period the player could only form an abstract intention to switch or stay rather than a specific motor response to press the left or right button.

2.3 Behavioural procedure

Participants took part in an initial behavioural training session to learn the task. First they practiced the task without player A, with the first two switch points already set to the *stay* configuration from the beginning of each trial. During this part of the practice session they saw the words 'LEFT' and 'RIGHT' presented in their associated colours on the left and right of the screen so that they could learn these colour-position mappings. Following this, they performed the task without these reminders being presented on the screen until they reached a criterion of at least 75% correct. Next, they practiced the task with the experimenter (Hoki) also taking part as player A, controlling the first switch point. The second switch point was still always set to *stay*. During this part of the task the experimenter sat beside the participant and pressed keys on the keyboard to configure the first switch point, so that the participant would see that player A was a real person taking part in the task. Next, they practised the task with the second switch point set to a *switch* configuration, followed by practice trials with the second switch point randomly set to *switch* or *stay* configurations. After this, a new element of the task was introduced. Participants performed 16 practice trials, and on two of these trials Hoki configured the first switch point incorrectly (i.e. catch trials). As a result player B had to reverse their

intended configuration of the third switch point to guide the ball to their goal location. The purpose of this was to provide a measure of whether participants were generating an expectation for player A's behaviour in each trial's initial thinking period. If so, an incorrect choice from player A would be surprising, as a result of which player B would be required to rapidly change their action-plan, which could be detected behaviourally by increased response time and/or error rate. Alternatively, if player B simply reacted to the position of the ball after player A configured the first switch point, there would be no reason to expect them to behave differently between catch and noncatch trials. At the end of this practice session, participants were presented with the question 'Did Hoki sometimes surprise you with her choice?', which they answered by pressing a button corresponding to the words 'No' and 'Yes' presented on the left and right of the screen. This was included as a means of encouraging participants to form expectations for player A's behaviour throughout the task. Finally, in the last practice session, participants performed 16 trials where player A was Hoki, followed by the question 'Did Hoki sometimes surprise you with her choice?', then 16 trials were player A was Computer, followed by the question 'Did the Computer sometimes surprise you with its choice?'. This matched the task that would subsequently be performed in the fMRI scanning session. In order to make clear the distinction between 'Hoki' and 'Computer' trials, whereas Hoki was engaged in the task and pressed keys on the computer while she was player A, when the Computer was player A the switch point was configured without requiring any key press and Hoki looked away from the laptop and read a magazine to make it clear that she was no longer involved in the task.

During the scanning session, there were six runs, each consisting of 32 trials as described above, with the exception that unbeknownst to the participant, the computer was now always in control of player A so that Hoki / Compuer conditions were perfectly

matched. The order of Hoki / Computer for player A in the first run was counterbalanced between participants, and subsequently reversed from each run to the next. Each set of 16 trials consisted of the full set of possible trial types, presented in random order. The two colour-pairs used to cue goal locations alternated from each trial to the next. On two randomly-selected runs there were no catch trials; two runs each contained two catch trials out of the 16 where player A was Hoki, and two runs each contained two catch trials out of the 16 where player A was the Computer. Following the MRI scan, participants were asked to fill out a questionnaire giving their impressions of the task (see below).

The timings on each trial were as follows. The initial thinking time (the focus of the neuroimaging analyses) followed an approximately exponential distribution (min: 3s, max: 18.5s, mean: 8s). By jittering the duration in this manner we were able to decorrelate the haemodynamic response to this period of the trial from subsequent parts (Visscher et al., 2003). Following this, player A's response options were presented on the screen. Player A responded 0.5 – 1.1 seconds later (mean: 0.7). These timings were matched between the Hoki and Computer conditions, and were similar to the timings produced by Hoki in the practice session. At this point the ball began falling down the screen and player B's response options were presented. As soon as player B responded, the third switch point was filled in accordingly. If player B had not responded by the time the ball reached the position of the third switch point (which took 2 seconds), the switch point was set to the incorrect configuration. Once the ball reached the third switch point, i.e. the final moment at which it was possible for player B to make a choice, it then sped up and took 0.4s to reach the bottom of the screen, followed by a 1s feedback period, then a 0.5 pause until the next trial. Therefore mean total trial duration was 12.6s. Each run consisted of two miniblocks of 16 trials, with each miniblock

followed by a 6 second period where participants were asked if they were surprised by player A's behaviour. The two miniblocks were identical except that the agency for player A switched between Hoki and Computer.

2.4 fMRI procedure

A 1.5T Siemens TIM Avanto scanner was used to acquire both T1-weighted structural images and T2*-weighted echoplanar images (64 x 64; 3.2 x 3.2mm pixels; echo time: 40 ms) with blood oxygen level-dependent (BOLD) contrast. Each volume comprised 40 axial slices (3.2mm thick, oriented approximately to the anterior commissure-posterior commissure plane). Functional scans were acquired in six sessions, each comprising 428 volumes (~7 min). Volumes were acquired continuously using a multi-band sequence (acceleration factor: 4), with an effective repetition time of 1s per volume. The first nine volumes in each session were discarded to allow for T1 equilibration effects. Between the third and fourth functional scan, a 6 min T1-weighted structural scan was performed.

2.5 Data analysis

Multi-voxel pattern analyses were conducted as follows. Separate analyses were conducted to decode each of the following pieces of information: player A's goal (left/right), player A's action-plan (switch/stay), player B's goal (left/right), player B's action-plan (switch/stay). Seeing as player B was the scanned participant, this means that we were decoding either player B's representation of player A's goal/action-plan, or player B's representation of their own goal/action-plan. We did this by partitioning the 16 possible trial types into two sets of 8, depending on the piece of information we were decoding, and training a classifier to distinguish the two categories. For example, suppose

that we were decoding player B's representation of player A's goal. In this case, we would train a classifier to distinguish trial types 1,2,5,6,9,10,13,14 versus 3,4,7,8,11,12,15,16 (see Table 1). Note that these categories differ only in player A's goal and are perfectly balanced in terms of the initial ball position, player B's goal, the second switch configuration, player A's action plan, and player B's action plan. Therefore player A's goal could be examined, unconfounded with the other pieces of information. Next, to decode player B's representation of their own goal, we would classify trial types 1,3,5,7,9,11,13,15 versus 2,4,6,8,10,12,14,16. Again, the two categories classified here are perfectly balanced in terms of each of the other pieces of information. This procedure was therefore repeated to decode each possible type of information, separately for the human-partner and computer-partner conditions.

fMRI data were analysed using SPM12 software running on MATLAB R2016b for Mac, along with custom-written MATLAB code. First-level models were set up to analyse fMRI data after realignment with SPM12 but prior to normalisation or smoothing. For each MVPA analysis, the thinking period at the beginning of each trial was modelled with eight separate boxcar regressors: four representing one of the categories being decoded (e.g. player A's goal: left), the other four representing the other category (e.g. player A's goal: right). The four regressors for each decoding category resulted from the crossing of two possible agents (Hoki / Computer) and two possible pairs of colours used to represent left vs right goals. Additional regressors coded for 1) Hoki's non-catch response; 2) Hoki's catch response; 3) Computer's non-catch response; 4) Computer's catch response; 5) The participant's own response; 6) Feedback following correct responses; 7) Feedback following incorrect responses; 8) Post-miniblock question about whether there had been any surprising responses. These were modelled with delta functions, apart from the feedback regressors (modelled with a duration of 1s) and the

post-miniblock regressor (modelled with a duration equivalent to the participant's response time). Additional regressors coded for the six movement parameters derived from realignment and the mean over scans. This comprised the full model for each session.

Parameter estimates from first-level models were used for MVPA analyses. These analyses used a searchlight approach (Kriegeskorte et al., 2006): we investigated decoding accuracy from a sphere of voxels centered on each voxel in the brain in turn (radius: three voxels). Each sphere yielded a vector of voxelwise parameter estimates representing one of the two conditions being distinguished (e.g. player A's goal left vs player A's goal right), in one of the six sessions. These vectors were individually normalised to mean 0, SD 1 before being entered into MVPA analysis, and the resulting decoding accuracy was assigned to the central voxel after subtracting 50%, yielding a whole-brain decoding map where zero indicated chance performance. Separate analyses were conducted for the human-partner and computer-partner conditions. For each condition, linear support vector machines were trained to distinguish patterns corresponding to the two categories (LIBSVM implementation; http://www.csie.ntu.edu.tw/~cjlin/libsvm; regularization parameter: 1). A leave-one-out crossvalidation approach was used so that classifiers were trained on five sessions and tested on the sixth, rotating training/testing sets over the six sessions. Furthermore, classifiers were always trained on data from one colour-goal mapping and tested on the other (averaging results after flipping the two mappings). This ensured that classification could not be based on the colours used to cue the two goals.

To evaluate decoding accuracy at the group level, decoding maps were normalized into 3mm cubic voxels using Montreal Neurological Institute reference space and fourth-

degree B-spline interpolation, based on normalisation parameters derived from segmenting the coregistered structural scan[1]. The normalised maps were then smoothed using a 4mm Gaussian kernel (as in Gilbert, 2011) and entered into a one-sample *t* test at the second level. This allows evaluation of regions showing consistently above-chance decoding accuracy across participants. Results were considered significant if they passed a familywise error corrected extent threshold of $p < .05$, based on a cluster forming threshold of $p < .001$. This follows the statistical approach and threshold used in other recent MVPA studies (e.g. Loose et al., 2017), avoiding the inflated false positive rates seen at more liberal statistical thresholds by Eklund et al. (2016).

3. Results

3.1 Questionnaire measures

No participant spontaneously guessed that the human-partner and computer-partner conditions were in fact identical. All participants said that they were at least somewhat surprised to discover that the experimenter was no longer involved in the task once the scanning session began. However, participants reported that they approached the task similarly in the two conditions, with only one participant out of 24 reporting any difference in behaviour between the two (this participant said that they tried harder to perform well in the human-partner condition). For full results from the questionnaire, see Supplementary Materials.

3.2 Behavioural results

[1] For one participant no structural scan was available and the normalisation parameters were based on the mean functional scan.

Behavioural results are summarised in Figure 3. Participants generally performed accurately; however, accuracy was reduced and reaction time increased on catch trials, where player A configured the first switch point unexpectedly. This indicates that participants formed prior expectations during the initial thinking period on each trial. If they had simply reacted to player A's configuration of the first switch point there would be no reason to expect any difference between the two types of trial. Accuracy data was entered into a repeated measures ANOVA with factors Catch (catch vs non-catch) and Partner (human vs computer). This showed a significant effect of Catch ($F(1,23) = 21$, $p < .001$, $\eta^2_p = .48$) but no significant effect of Partner or Catch x Partner interaction ($F(1,23) < 2.6$, $p > .12$, $\eta^2_p < .11$). A similar analysis of the reaction time data showed a significant effect of Catch ($F(1, 23) = 86$, $p < .001$, $\eta^2_p = .79$), but no significant effect of Partner or Catch x Partner interaction ($F(1,23) < 2.7$, $p > .11$, $\eta^2_p < .11$). Accuracy for participants' report at the end of each block whether there had been any catch trials was 78%, which was significantly above chance ($t(23) = 7.2$, $p < .001$). Note that answering this question correctly requires the catch/non-catch status of all 16 trials to be accurately evaluated, therefore this represents a much higher level individual-trial accuracy than 78%.

3.3 Decoding intentions of self and other

Eight separate decoding analyses were conducted as a result of crossing the following three factors: A. own intention or other intention (below we refer to this as 1st-person and 3rd-person respectively); B. goal (left-right) or action-plan (switch-stay); C. human-partner or computer-partner. Therefore, the eight analyses attempted to decode the

following: 1) 1st-person goal, human-partner; 2) 1st-person goal, computer-partner; 3) 1st-person action-plan, human-partner; 4) 1st-person action-plan, computer-partner; 5) 3rd-person goal, human-partner; 6) 3rd-person goal, computer-partner; 7) 3rd-person action-plan, human-partner; 8) 3rd-person action-plan, computer-partner. Rather than report each of these analyses here (with the attendant multiple comparisons problem this would cause), we first averaged them and entered the mean decoding map into a one-sample t-test (for results from each of the individual decoding analyses, see Supplementary Materials). This provides a single test potentially sensitive to any type of information, from which it is possible to generate regions of interest which are unbiased with respect to the three factors described above. We then consider the evidence for significant decoding of each specific type of information below.

The mean decoding map identified four regions at a whole-brain corrected threshold showing significant effects, all in the left hemisphere (Table 3, Figure 4): dorsal posterior frontal cortex (BA 6/8), superior parietal cortex (BA 7), posterior cingulate (BA 23/30) and occipital cortex (BA 18). Mean decoding accuracies were extracted across all voxels in each of these regions of interest (ROIs) and entered into a repeated measures ANOVA with factors Region, Perspective (1st person / 3rd person), Component (goal / action-plan) and Partner (human / computer). Note that these ROIs are biased towards above-chance decoding accuracy seeing as this was the statistical criterion by which they were selected (based on a familywise error correction for multiple comparisons). Therefore in the analyses below we only report significance tests for effects which are orthogonal to this selection criterion (Kriegeskorte et al., 2009). There was a main effect of Component ($F(1,23) = 7.1$, $p = .014$, $\eta^2_p = .235$), qualified by a Component x Region interaction ($F(3,69) = 7.3$, $p = .0003$, $\eta^2_p = .240$). There were no other significant effects ($p > .079$, $\eta^2_p < .094$).

[Table 3 about here]

[Figure 4 about here]

These results are illustrated in Figure 5. Overall, accuracies for decoding action-plans (i.e. switch versus stay) were higher than accuracies for goals (left versus right). However, these results were modulated by region. Numerically, decoding accuracy for action-plans increased in the following sequence: 1) occipital cortex, 2) posterior cingulate, 3) superior parietal cortex and 4) frontal cortex. Decoding accuracy for goals increased in exactly the reverse sequence. Thus, when considering each intention component separately, there was a main effect of Region for both action-plans ($F(3,69) = 4.54$, $p = .006$, $\eta^2_p = .165$) and goals ($F(3,69) = 4.41$, $p = .007$, $\eta^2_p = .161$). This indicates that the decoding map shown in Figure 4 reflects significant contributions from both goal decoding and action-plan decoding. If mean decoding accuracy had been driven by above-chance levels for just one intention component, there would be no reason for the other component to differ significantly between regions.

[Figure 5 about here]

3.4 Cross-classification between 1st and 3rd person perspectives

Next we investigated whether it was possible to crossclassify between the intentions of the scanned-participant and their partner. To do this, we trained classifiers on categories based on one classification (e.g. own intention switch versus stay) and tested them on categories based on another (e.g. partner's intention switch versus stay). These analyses were performed using a searchlight approach across the whole brain, with classifiers

trained on one colour-goal mapping and tested on the other, as in the decoding analyses above. For all analyses, we averaged over classification direction (i.e. train on partner A, test on partner B / train on partner B, test on partner A). Note that classifications for the two partners were strictly orthogonal, by design. This means that a classifier that was 100% accurate for decoding one partner's intention would necessarily be at chance level for the other partner. Nevertheless, it is still possible for a single classifier to decode both partners' intentions with above chance accuracy. For example, consider a classifier that predicts switch if *either* partner has an intention to switch, otherwise stay. This classifier would correctly classify 75% of trials.

[Figure 6 about here]

We averaged results from the cross-classification analysis in the four regions of interest defined by the analysis above. Note that unlike the decoding accuracies analysed in section 3.3, the decoding accuracies in the present analysis are unbiased at our regions of interest because the trial categorisations for training and testing data were orthogonal to one another. As a result, the analyses under investigation were independent of those used to define the regions of interest (Kriegeskorte et al., 2009). Therefore, we include tests for above-chance decoding accuracies in the present section, which we refrained to do in section 3.3 because it would have been biased. Decoding accuracies were analysed in a repeated-measures ANOVA with factors Region, Component (goal / action-plan) and Partner (human / computer), after subtracting 50 so that zero represented chance performance. The intercept for this ANOVA was greater than zero ($F(1,23) = 4.4$, $p = .047$, $\eta^2_p = .161$), indicating that significant crossclassification occurred. There were also main effects of Region ($F(3,69) = 5.77$, $p = .001$, $\eta^2_p = .200$) and Component ($F(1,23) = 8.95$, $p = .007$, $\eta^2_p = .280$). Results are shown in Figure 6, which shows significant cross-

classification of action-plans in frontal and superior parietal cortex ($p < .0012$). Both of these effects survived a Bonferroni correction for multiple comparisons (corrected alpha = .006); none of the other cross-classifications was significant, even without Bonferroni correction ($p > .28$).

3.5 Cross-classification between mental-state decoding and physical-state decoding

One of the aims of the present study was to investigate cross-classification between intention components and states of the world corresponding to those intentions. Such cross-classification might be expected if, for example, holding a particular intention was associated with simulating its perceptual consequences. This would be consistent with prior evidence of overlapping fMRI activation patterns for perception and imagery (Cichy et al., 2012; Stokes et al., 2009).

In order to investigate cross-classification of this type, we first collapsed over the human-partner and computer-partner conditions (which failed to show any significant differences). We also collapsed across the two colour-goal mappings. It was no longer necessary to keep these separate because the results of the following analyses could not be biased by classification of colour cues. We then performed separate cross-classification analyses for the two intention components and report results from the four ROIs as above. For goal decoding, we cross-classified between left versus right goals, and trials that began with an image of the ball above the left versus right pipe. For plan decoding, we cross-classified between switch versus stay intentions, and trials that began with the second switch-point shown in the switch versus stay configuration. Therefore these analyses cross-classified between invisible goals or plans inferred by the participants, and visible states of the world that corresponded to these goals or plans. We

conducted these analyses separately for 1st-person and 3rd-person intentions. As with our earlier analyses, these cross-classifications were unbiased in our experimental design, e.g. intentions with left versus right goals were equally likely to occur on trials with the ball starting above the left versus right pipe (see Tables 1 and 2).

[Figure 7 about here]

Using the same ROIs as sections 3.3 and 3.4 above, decoding accuracies were analysed in a Region x Perspective (1st-person / 3rd-person) x Component (goal / plan) Repeated Measures ANOVA. The intercept for this ANOVA was greater than zero, indicating that significant cross-classification was possible ($F(1,23) = 45.6$, $p < .001$, $\eta^2_p = .67$). There were main effects of Region ($F(3,69) = 9.58$, $p < .001$, $\eta^2_p = .29$) and Component ($F(1,23) = 15.1$, $p < .001$, $\eta^2_p = .40$), along with interactions between Region x Component ($F(3,69) = 12.1$, $p < .001$, $\eta^2_p = .34$), Component x Perspective ($F(1,23) = 7.69$, $p = .011$, $\eta^2_p = .25$) and Region x Component x Perspective ($F(3,69) = 16.4$, $p < .001$, $\eta^2_p = .42$). These results are illustrated in Figure 7. Significant cross-classification was possible for both 1st-person and 3rd-person action-plans in frontal and superior parietal cortex. In occipital cortex, cross-classification was possible for 1st-person action-plans and 3rd-person goals. Each of these effects survived a Bonferroni correction over the 16 tests conducted ($t(23) > 4.2$, $p < .00031$; corrected alpha = .003). None of the other cross-classification effects was significant, even at an uncorrected threshold ($t(23) < 1.7$, $p > .11$).

3.6 Intention-specific activity patterns?

In a final set of analyses, we tested whether there were any regions for which the initial intention decoding analyses yielded better decoding accuracies than the subsequent cross-classification analyses. This might be expected if there were intention-specific patterns of brain activity that did not generalise between 1st-person and 3rd-person intentions, or between intention representations and representations of physical states of the world. We also searched for regions in which 3rd-person intention decoding was significantly different between the human-partner and computer-partner conditions. None of these planned analyses produced any significant results. Thus, there was no evidence for intention-specific patterns of brain activity. We report a further set of exploratory analyses in Supplementary Materials. While the majority of these analyses produced nonsignificant results, a small number of significant results were found. However, we do not consider that these results provide strong evidence for intention-specific activity patterns unless they can be replicated, given the high risk of false positives when a large number of exploratory analyses are conducted.

We also investigated the strength of evidence for a *null* effect of Perspective (1st person vs 3rd person) and Partner (human vs computer) on intention decoding in the regions of interest shown in Table 3. To do this we repeated the analysis described in section 3.3 above, using a Bayesian Repeated Measures ANOVA in the statistics package JASP 0.8.3.1 (JASP Team, 2017). Our null model included the factors Region, Component, and Region x Component, to account for the significant Region x Component interaction described above. We then calculated Bayes Factors BF_{01} for models additionally including factors of Perspective and Partner, and all interactions with other factors. This provides a measure of the evidence for a null effect of including these additional factors. In all cases, models including the Perspective factor and its interaction with other factors had a BF_{01} greater than 7.8, and those including Partner and its

interactions had a BF_{01} greater than 10.9. Bayes factors greater than 10 are conventionally interpreted as providing "strong" evidence, while those in 3-10 range are described as "substantial" (Jeffreys, 1961). Thus, the evidence for a null effect of Perspective and Partner in our regions of interest was in the substantial-to-strong range.

4. Discussion

In this study we aimed to extend previous "intention decoding" fMRI experiments with a paradigm incorporating three novel features: A) attempting to decode both the scanned-participant's own intention and their representation of another agent's intention; B) distinguishing goal versus action-plan components of intentions; C) attempting cross-classification between intention components and states of the world corresponding to those components. Results showed that it was possible to decode both 1st person and 3rd person intentions, with distinct brain regions showing selectivity for goal versus action-plan components of those intentions. However, follow-up analyses suggested that the decoding analyses reflected low-level processes involved in generating the content of intentions and/or expecting their outcomes, rather than any intention-specific pattern of brain activity.

The network of brain regions from which intentions could be decoded included regions of dorsal prefrontal, superior parietal, posterior cingulate, and occipital cortices. Similar regions of dorsal prefrontal and superior parietal cortex frequently co-activate in functional neuroimaging studies (Yeo et al., 2011) and have been proposed to play a key role in awareness of conscious intentions and motor plans (Cona et al., 2015; Desmurget et al., 2009; Desmurget and Sirigu, 2009; Haggard, 2005). These regions - termed the 'dorsal frontoparietal network' by some authors (Ptak et al., 2017) - have also been

implicated in a range of visuospatial processes relevant to the present paradigm, such as generation of saccades (Grosbras et al., 2005), shifts in visual attention (Corbetta et al., 1998; de Haan et al., 2008), simulation of spatial transformations and mental rotation (Zacks, 2008), motor imagery, and emulation of visuomotor processes (Hétu et al., 2013; Zabicki et al., 2016).

While the present results are certainly consistent with a prominent role of dorsal frontoparietal regions in representing intentions, there are at least at least three possible roles that these regions could have played that can be described without invoking the concept of intention. First, it is possible that intentions to switch versus stay were associated with different patterns of eye movements and that our decoding analyses were simply detecting saccade-related activity patterns in this network. In the absence of eye-tracking data from this study, we cannot exclude this possibility. A second potential role of this network is in performing the spatial computations required to compare the starting position of the ball with its goal position, and thereby derive an appropriate intention to switch versus stay. This would be consistent with its role in other paradigms involving computation of spatial transformations, such as mental rotation tasks (Zacks, 2008). Third, this network may have played a role in emulating intended action-plans and/or simulating the perceptual consequences of plans to switch versus stay. This would be consistent with "motor emulation" theories of the role of this network (Ptak et al., 2017). These three possibilities are not mutually exclusive, and would all be compatible with the significant cross-classification between intended action-plans to switch versus stay and the perceptually-presented configuration of the second switch point in our experimental design. Each of these possible functions could play an important role in supporting visuomotor intentions. However, none of them can be interpreted as an intention-specific processes.

The posterior cingulate region identified in our decoding analyses has also been proposed to play a key role in processing intentions. For example, Cona et al.'s (2015) meta-analysis showed consistent activation of posterior cingulate associated with both encoding and retrieving intentions. Gilbert et al. (2012) found that voxelwise patterns of activity in this region showed greater similarity between encoding and retrieval when intentions were successfully fulfilled rather than missed (see also Qiao et al., 2017 for evidence that similar medial parietal regions are involved in representing intentional task set). This region also plays an important role in visuospatial tasks such as those involving eye movements (Berman et al., 1999), spatial attention (Hopfinger et al., 2000), and translation between egocentric and allocentric spatial reference frames (Vogt et al., 1992).

Turning now to occipital cortex, this region showed the greatest decoding of goals (left versus right), but no cross-classification between 1st-person and 3rd-person representations. However, occipital cortex did show significant cross-classification between visual presentation of the ball on the left versus right of the screen and the 3rd-person goal of partner A to move the ball to the left versus right pipe. It also showed cross-classification between visual presentation of the second switch point in a switch versus stay configuration and player B's action-plan to switch versus stay. Both of these effects could be explained if this region held a visual template of the scanned participant's expectation of the outcome of player A's intended behaviour, along with a visual template of how the scanned-participant (player B) should then respond. The first of these pieces of information would be required as an intermediate step in the inference as to whether the correct plan is to switch or stay. The second would then provide a template that would allow selection of the correct response by matching the visual

presentation of the two response options against a representation of the correct one to choose.

Unlike the majority of previous intention decoding studies (reviewed by Haynes, 2014; Momennejad and Haynes, 2013), the present study did not show any evidence of intention decoding from medial prefrontal cortex, a region thought to play a key role in intentional action (Brass et al., 2013). There are at least two relevant differences between the present paradigm and previous ones that may explain this discrepancy (see also Zhang et al., 2013 for a relevant discussion of rule encoding in medial prefrontal cortex). First, the intentions contrasted in previous studies tended to represent abstract task sets (e.g. perform an addition versus a subtraction task) rather than visuospatial options used here (left vs right and switch vs stay). Second, in the present study there was no filler task during the intention planning period: participants simply focused on thinking about their intention until the response cues appeared. Previous work suggests that medial prefrontal decoding of delayed intentions is particularly associated with periods filled with a distracting task (Momennejad and Haynes, 2013). A possible explanation of this – at least with respect to the rostral aspects of this brain region - comes from the 'gateway hypothesis' of Burgess and colleagues (Burgess et al., 2007; Gilbert et al., 2005). This hypothesis suggests remembering intentions whilst also dealing with a distracting ongoing task may particularly recruit rostral prefrontal cortex due to a role of this brain region in attentional selection between stimulus-oriented thought (prompted by the ongoing task) and maintenance of stimulus-independent representations of intended behaviour. Accordingly, rostral prefrontal cortex would not be expected to play a role in task such as the present one, where participants are free to focus on intentions without simultaneously having to deal with a distracting ongoing task.

Our design included two features that would have allowed us to distinguish brain activity specifically related to processing of intentions versus non-specific visuo-spatial processes. First, we were able to compare patterns of brain activity that distinguished the scanned participant's own intention to switch versus stay from their representation of another agent's intention. Seeing as the visuo-spatial computations required to generate appropriate 1st-person and 3rd-person intentions were matched, this would have revealed patterns that could not be attributed simply to lower-level visuo-spatial processes. Second, when examining participant's representation of 3rd-person intentions we could compare their representation of a human partner's intention versus a computer partner. A brain region from which it was possible to decode participants' representations of a human partner's intention, but not a computer partner in a matched task, would be a candidate region for supporting representations which were specific to the content of other humans' mental states.

Neither of these comparisons revealed any significant differences in decoding accuracy. On the basis of this, we argue that the present results provide no evidence for unique brain representations of intentions as a special type of mental state, distinct from processes such as generation and elaboration of visuomotor plans, simulating expected perceptual outcomes, and so on. Of course, this is not to deny that alternative experimental paradigms might be more successful, or that alternative analysis techniques with the present dataset might have detected intention-specific patterns. To facilitate re-analyses of our data by other researchers, we have provided the full fMRI dataset for download at the following location: [the dataset will be uploaded upon acceptance of this manuscript]. We cannot exclude the possibility that our failure to find intention-specific patterns of activity merely reflects low statistical power. We tested a total of 24 participants, and 96 trials contributed to each of the decoding analyses performed (or 192

trials for analyses collapsing over human-partner and computer-partner conditions). This is comparable to previous fMRI intention decoding studies, e.g. Momennejad and Haynes (2012, 2013), who tested 20 and 23 participants respectively, and used 60 and 72 trials respectively for their main decoding analyses. Nevertheless, the problem of low power in neuroscience studies should not be underestimated (Button et al., 2013) and it is clearly possible that studies with greater power might detect intention-specific patterns of brain activity.

Our results could have implications on a practical as well as a theoretical level. Recent speculation has suggested that intention decoding neuroimaging methodologies could play a role in legal contexts, for example decoding criminal intent from a brain scan (see Haynes, 2014 for discussion). We are sceptical of this possibility. This would require decoding of 1st-person commitment to an intended act, rather than a representation of someone else's intention, or thinking about a particular action-plan without any commitment to actually bring it about (e.g. an expectation that an outcome will occur without motivation to bring it about). Our results suggest that it could be difficult to distinguish which of these is being decoded when we "decode intentions". We also suggest that previous intention decoding neuroimaging studies have failed to demonstrate patterns of brain activity that are unequivocally intention-specific, rather than potentially reflecting other factors such as expectations for future events regardless of any intention to bring them about. We propose that a clearer understanding of what intention decoding analyses are actually decoding can come from paradigms that distinguish various intention components from each other and compare decoding of intentions with decoding of expectations. This can help us to make progress with the following three questions:

1) How selectively are intention components such as goals and action-plans coded by distributed brain networks?
2) To what extent are patterns of brain activity involved in representing our own intentions also involved in representing the intentions of others? The two types of brain representation must in some sense be distinct, seeing as we are can represent other agents' intentions without those intentions controlling our own behaviour. How might this distinction between 1st-person and 3rd-person intentions be explained?
3) Can we distinguish patterns of brain activity involved in commitment to an intention as opposed to expectation or simulation of its consequences?

Answering these questions is likely to benefit from further consideration of domain-general processes that play a role in generating and representing intentions, rather than necessarily conceiving of intentions as a special type of mental state with distinct neural correlates.

References

Altmann, E.M., Trafton, J.G., 2002. Memory for Goals: An Acctivation-Based Model, Cognitive Science. doi:10.1016/S0364-0213(01)00058-1

Anderson, J.R., 1996. ACT: A simple theory of complex cognition. Am. Psychol. 51, 355–365. doi:10.1037//0003-066X.51.4.355

Anderson, J.R., Bothell, D., Byrne, M.D., Douglass, S., Lebiere, C., Qin, Y., 2004. An integrated theory of the mind. Psychol. Rev. 111, 1036–60. doi:10.1037/0033-295X.111.4.1036

Apperly, I.A., 2008. Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study "theory of mind." Cognition 107, 266–283. doi:10.1016/j.cognition.2007.07.019

Berman, R.A., Colby, C.L., Genovese, C.R., Voyvodic, J.T., Luna, B., Thulborn, K.R., Sweeney, J.A., 1999. Cortical networks subserving pursuit and saccadic eye movements in humans: An FMRI study. Hum. Brain Mapp. 8, 209–225. doi:10.1002/(SICI)1097-0193(1999)8:4<209::AID-HBM5>3.0.CO;2-0

Brand, M., 1984. Intending and acting. MIT Press, Cambridge, MA.

Brandimonte, M.A., Einstein, G.O., McDaniel, M.A., 1996. Prospective Memory: Theory and Applications. Psychology Press.

Brass, M., Lynn, M.T., Demanet, J., Rigoni, D., 2013. Imaging volition: what the brain can tell us about the will. Exp. brain Res. doi:10.1007/s00221-013-3472-x

Bratman, M.E., 1987. Intention, plans, and practical reason. Cambridge University Press, Cambridge, MA.

Burgess, P.W., Dumontheil, I., Gilbert, S.J., 2007. The gateway hypothesis of rostral prefrontal cortex (area 10) function. Trends Cogn. Sci. 11, 290–8. doi:10.1016/j.tics.2007.05.004

Burgess, P.W., Gonen-Yaacovi, G., Volle, E., 2011. Functional neuroimaging studies of

prospective memory: what have we learnt so far? Neuropsychologia 49, 2246–57. doi:10.1016/j.neuropsychologia.2011.02.014

Burgess, P.W., Quayle, A., Frith, C.D., 2001. Brain regions involved in prospective memory as determined by positron emission tomography. Neuropsychologia 39, 545–55.

Burgess, P.W., Scott, S.K., Frith, C.D., 2003. The role of the rostral frontal cortex (area 10) in prospective memory: a lateral versus medial dissociation. Neuropsychologia 41, 906–918. doi:10.1016/S0028-3932(02)00327-5

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. 14, 365–76. doi:10.1038/nrn3475

Carruthers, P., Smith, P.K. (Eds.), 1996. Theories of theories of mind. Cambridge University Press, Cambridge.

Cichy, R.M., Heinzle, J., Haynes, J.-D., 2012. Imagery and perception share cortical representations of content and location. Cereb. Cortex 22, 372–80. doi:10.1093/cercor/bhr106

Cona, G., Scarpazza, C., Sartori, G., Moscovitch, M., Bisiacchi, P.S., 2015. Neural bases of prospective memory: A meta-analysis and the "Attention to Delayed Intention" (AtoDI) model. Neurosci. Biobehav. Rev. 52, 21–37. doi:10.1016/j.neubiorev.2015.02.007

Cooper, R., Shallice, T., 2000. Contention Scheduling and the Control of routine activities. Cogn. Neuropsychol. 17, 297–338. doi:10.1080/026432900380427

Corbetta, M., Akbudak, E., Conturo, T.E., Snyder, A.Z., Ollinger, J.M., Drury, H.A., Linenweber, M.R., Petersen, S.E., Raichle, M.E., Van Essen, D.C., Shulman, G.L., 1998. A common network of functional areas for attention and eye movements. Neuron 21, 761–773. doi:10.1016/S0896-6273(00)80593-0

Cross, E.S., Liepelt, R., Antonia, A.F., Parkinson, J., Ramsey, R., Stadler, W., Prinz, W., 2012. Robotic movement preferentially engages the action observation network. Hum. Brain Mapp. 33, 2238–2254. doi:10.1002/hbm.21361

de Haan, B., Morgan, P.S., Rorden, C., 2008. Covert orienting of attention and overt eye movements activate identical brain regions. Brain Res. 1204, 102–111. doi:10.1016/j.brainres.2008.01.105

Desmurget, M., Reilly, K.T., Richard, N., Szathmari, A., Mottolese, C., Sirigu, A., 2009. Movement intention after parietal cortex stimulation in humans. Science 324, 811–3. doi:10.1126/science.1169896

Desmurget, M., Sirigu, A., 2009. A parietal-premotor network for movement intention and motor awareness. Trends Cogn. Sci. 13, 411–419. doi:10.1016/j.tics.2009.08.001

Duda, R.O., Hart, P.E., Stork, D.G., 2000. Pattern Classification, New York: John Wiley, Section. doi:10.1038/npp.2011.9

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure : Why fMRI inferences for spatial extent have inflated false-positive rates 1–6. doi:10.1073/pnas.1602413113

Fodor, J., 1974. Special sciences and the disunity of science as a working hypothesis. Synthese 28, 71–115.

Friston, K., Mattout, J., Kilner, J., 2011. Action understanding and active inference. Biol. Cybern. doi:10.1007/s00422-011-0424-z

Frith, C.D., Frith, U., 2012. Mechanisms of social cognition. Annu. Rev. Psychol. 63, 287–313.

Gallagher, H.L., Jack, A.I., Roepstorff, A., Frith, C.D., 2002. Imaging the Intentional Stance in a Competitive Game. Neuroimage 16, 814–821. doi:10.1006/nimg.2002.1117

Gazzola, V., Rizzolatti, G., Wicker, B., Keysers, C., 2007. The anthropomorphic brain:

The mirror neuron system responds to human and robotic actions. Neuroimage 35, 1674–1684. doi:10.1016/j.neuroimage.2007.02.003

Gilbert, S.J., 2011. Decoding the content of delayed intentions. J. Neurosci. 31, 2888–94. doi:10.1523/JNEUROSCI.5336-10.2011

Gilbert, S.J., Armbruster, D., Panagiotidi, M., 2012. Similarity between brain activity at encoding and retrieval predicts successful realization of delayed intentions. J. Cogn. … 93–105.

Gilbert, S.J., Frith, C.D., Burgess, P.W., 2005. Involvement of rostral prefrontal cortex in selection between stimulus-oriented and stimulus-independent thought. Eur. J. Neurosci. 21, 1423–31. doi:10.1111/j.1460-9568.2005.03981.x

Gilbert, S.J., Gollwitzer, P.M., Cohen, A.-L., Burgess, P.W., Oettingen, G., 2009. Separable brain systems supporting cued versus self-initiated realization of delayed intentions. J. Exp. Psychol. Learn. Mem. Cogn. 35, 905–15. doi:10.1037/a0015535

Gilbert, S.J., Williamson, I.D.M., Dumontheil, I., Simons, J.S., Frith, C.D., Burgess, P.W., 2007. Distinct regions of medial rostral prefrontal cortex supporting social and nonsocial functions. doi:10.1093/scan/nsm014

Gollwitzer, P.M., 1999. Implementation intentions: Strong effects of simple plans. Am. Psychol. doi:10.1037/0003-066X.54.7.493

Gordon, R.M., 1986. Folk Psychology as Simulation. Mind Lang. 1, 158–171. doi:10.1111/j.1468-0017.1986.tb00324.x

Graf, P., Uttl, B., 2001. Prospective memory: a new focus for research. Conscious. Cogn. 10, 437–450. doi:10.1006/ccog.2001.0504

Grosbras, M.H., Laird, A.R., Paus, T., 2005. Cortical regions involved in eye movements, shifts of attention, and gaze perception, in: Human Brain Mapping. pp. 140–154. doi:10.1002/hbm.20145

Haggard, P., 2005. Conscious intention and motor cognition. Trends Cogn. Sci.

doi:10.1016/j.tics.2005.04.012

Haynes, J.-D., 2014. The Neural Code for Intentions in the Human Brain, in: Singh, I., Sinnott-Armstrong, W.P., Savulescu, J. (Eds.), Bioprediction, Biomarkers, and Bad Behavior: Scientific, Legal, and Ethical Challenges. Oxford University Press, Oxford, pp. 173–187.

Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–34. doi:10.1038/nrn1931

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S.J., Frith, C., Passingham, R.E., 2007. Reading hidden intentions in the human brain. Curr. Biol. 17, 323–8. doi:10.1016/j.cub.2006.11.072

Heal, J., 1986. Replication and functionalism, in: Butterfield, J. (Ed.), Language, Mind and Logic. Cambridge University Press, Cambridge.

Hétu, S., Grégoire, M., Saimpont, A., Coll, M.P., Eugène, F., Michon, P.E., Jackson, P.L., 2013. The neural network of motor imagery: An ALE meta-analysis. Neurosci. Biobehav. Rev. doi:10.1016/j.neubiorev.2013.03.017

Hickok, G., 2008. Eight Problems for the Mirror Neuron Theory of Action Understanding in Monkeys and Humans 1229–1243.

Hopfinger, J.B., Buonocore, M.H., Mangun, G.R., 2000. The neural mechanisms of top-down attentional control. Nat. Neurosci. 3, 284–91. doi:10.1038/72999

Jeffreys, H., 1961. Theory of Probability, Theory of Probability.

Kennedy, D.P., Adolphs, R., 2012. The social brain in psychiatric and neurological disorders. Trends Cogn. Sci. 16, 559–572.

Kilner, J.M., Friston, K.J., Frith, C.D., 2007. Predictive coding: An account of the mirror neuron system. Cogn. Process. 8, 159–166. doi:10.1007/s10339-007-0170-2

Kliegel, M., McDaniel, M.A., Einstein, G.O., 2008. Prospective memory: Cognitive, neuroscience, developmental, and applied perspectives., Prospective memory:

Cognitive, neuroscience, developmental, and applied perspectives. Erlbaum, Mahwah. doi:10.4324/9780203809945

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. U. S. A. 103, 3863–3868. doi:10.1073/pnas.0600244103

Kriegeskorte, N., Simmons, W.K., Bellgowan, P.S.F., Baker, C.I., 2009. Circular analysis in systems neuroscience: the dangers of double dipping. Nat. Neurosci. 12, 535–40. doi:10.1038/nn.2303

Laird, J.E., 2012. The Soar Cognitive Architecture. MIT Press, Cambridge, MA.

Landsiedel, J., Gilbert, S.J., 2015. Creating external reminders for delayed intentions: Dissociable influence on "task-positive" and "task-negative" brain networks. Neuroimage 104, 231–240. doi:10.1016/j.neuroimage.2014.10.021

Loose, L.S., Wisniewski, D., Rusconi, M., Goschke, T., Haynes, J.-D., 2017. Switch independent task representations in frontal and parietal cortex. J. Neurosci. 37, 3656–16. doi:10.1523/JNEUROSCI.3656-16.2017

McDaniel, M.A., Einstein, G.O., 2007. Prospective Memory: An Overview and Synthesis of an Emerging Field. Sage Publications Ltd, Los Angeles.

McDaniel, M.A., Einstein, G.O., 2000. Strategic and automatic processes in prospective memory retrieval: a multiprocess framework. Appl. Cogn. Psychol. 14, S127–S144. doi:10.1002/acp.775

Mele, A.R., 2009. Effective Intentions: The Power of Conscious Will, Effective Intentions: The Power of Conscious Will. doi:10.1093/acprof:oso/9780195384260.001.0001

Mele, A.R., 1992. Springs of action. Oxford University Press, Oxford.

Momennejad, I., Haynes, J.-D., 2013. Encoding of Prospective Tasks in the Human Prefrontal Cortex under Varying Task Loads. J. Neurosci. 33, 17342–17349.

doi:10.1523/JNEUROSCI.0492-13.2013

Momennejad, I., Haynes, J.-D., 2012. Human anterior prefrontal cortex encodes the "what" and "when" of future intentions. Neuroimage 61, 139–48. doi:10.1016/j.neuroimage.2012.02.079

Newell, A., 1990. Unified Theories of Cognition. Harvard University Press, Cambridge, MA.

Norman, K. a, Polyn, S.M., Detre, G.J., Haxby, J. V, 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–30. doi:10.1016/j.tics.2006.07.005

Oksanen, K.M., Waldum, E.R., McDaniel, M.A., Braver, T.S., 2014. Neural mechanisms of time-based prospective memory: Evidence for transient monitoring. PLoS One 9. doi:10.1371/journal.pone.0092123

Okuda, J., Fujii, T., Ohtake, H., Tsukiura, T., Yamadori, A., Frith, C.D., Burgess, P.W., 2007. Differential involvement of regions of rostral prefrontal cortex (Brodmann area 10) in time- and event-based prospective memory. Int. J. Psychophysiol. 64, 233–46. doi:10.1016/j.ijpsycho.2006.09.009

Okuda, J., Fujii, T., Yamadori, a, Kawashima, R., Tsukiura, T., Fukatsu, R., Suzuki, K., Ito, M., Fukuda, H., 1998. Participation of the prefrontal cortices in prospective memory: evidence from a PET study in humans. Neurosci. Lett. 253, 127–30.

Pacherie, E., 2008. The phenomenology of action: A conceptual framework. Cognition 107, 179–217. doi:10.1016/j.cognition.2007.09.003

Pacherie, E., Haggard, P., 2010. What are intentions?, in: Nadel, L., Sinnott-Armstrong, W. (Eds.), Conscious Will and Responsibility: A Tribute to Benjamin Libet. Oxford University Press, Oxford, pp. 70–84.

Ptak, R., Schnider, A., Fellrath, J., 2017. The Dorsal Frontoparietal Network: A Core System for Emulated Action. Trends Cogn. Sci. 21, 589–599.

doi:10.1016/j.tics.2017.05.002

Qiao, L., Zhang, L., Chen, A., Egner, T., 2017. Dynamic Trial-by-Trial Re-Coding of Task-Set Representations in Frontoparietal Cortex Mediates Behavioral Flexibility. J. Neurosci. 37, 935–17. doi:10.1523/JNEUROSCI.0935-17.2017

Ramnani, N., Miall, R.C., 2004. A system in the human brain for predicting the actions of others. Nat. Neurosci. 7, 85–90. doi:10.1038/nn1168

Searle, J., 1983. Intentionality. Cambridge University Press, Cambridge.

Simons, J.S., Schölvinck, M.L., Gilbert, S.J., Frith, C.D., Burgess, P.W., 2006. Differential components of prospective memory? Evidence from fMRI. Neuropsychologia 44, 1388–97. doi:10.1016/j.neuropsychologia.2006.01.005

Soon, C.S., Brass, M., Heinze, H.-J., Haynes, J.-D., 2008. Unconscious determinants of free decisions in the human brain. Nat. Neurosci. 11, 543–5. doi:10.1038/nn.2112

Soon, C.S., He, A.H., Bode, S., Haynes, J.-D., 2013. Predicting free choices for abstract intentions. Proc. Natl. Acad. Sci. U. S. A. 110, 6217–22. doi:10.1073/pnas.1212218110

Stokes, M., Thompson, R., Cusack, R., Duncan, J., 2009. Top-down activation of shape-specific population codes in visual cortex during mental imagery. J. Neurosci. 29, 1565–72. doi:10.1523/JNEUROSCI.4657-08.2009

Thomas Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zollei, L., Polimeni, J.R., Fischl, B., Liu, H., Buckner, R.L., 2011. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. 106, 1125–1165. doi:10.1152/jn.00338.2011

Tomasello, M., Carpenter, M., Call, J., Behne, T., Moll, H., 2005. Understanding and sharing intentions: the origins of cultural cognition. Behav. Brain Sci. 28, 675-91-735. doi:10.1017/S0140525X05000129

Uithol, S., Burnston, D.C., Haselager, P., 2014. Why we may not find intentions in the brain. Neuropsychologia 56C, 129–139. doi:10.1016/j.neuropsychologia.2014.01.010

Van Overwalle, F., Baetens, K., 2009. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. Neuroimage 48, 564–84. doi:10.1016/j.neuroimage.2009.06.009

Visscher, K.M., Miezin, F.M., Kelly, J.E., Buckner, R.L., Donaldson, D.I., McAvoy, M.P., Bhalodia, V.M., Petersen, S.E., 2003. Mixed blocked/event-related designs separate transient and sustained activity in fMRI. Neuroimage 19, 1694–1708. doi:10.1016/S1053-8119(03)00178-2

Vogt, B.A., Finch, D.M., Olson, C.R., 1992. Functional heterogeneity in cingulate cortex: The anterior executive and posterior evaluative regions. Cereb. Cortex 2, 435–443. doi:10.1093/cercor/2.6.435-a

Wisniewski, D., Goschke, T., Haynes, J., 2016. Similar coding of freely chosen and externally cued intentions in a fronto-parietal network. Neuroimage. doi:10.1016/j.neuroimage.2016.04.044

Wisniewski, D., Reverberi, C., Momennejad, I., Kahnt, T., Haynes, J.-D., 2015. The Role of the Parietal Cortex in the Representation of Task-Reward Associations. J. Neurosci. 35, 12355–12365. doi:10.1523/JNEUROSCI.4882-14.2015

Zabicki, A., De Haas, B., Zentgraf, K., Stark, R., Munzert, J., Krüger, B., 2016. Imagined and Executed Actions in the Human Motor System: Testing Neural Similarity Between Execution and Imagery of Actions with a Multivariate Approach. Cereb. Cortex 1–14. doi:10.1093/cercor/bhw257

Zacks, J.M., 2008. Neuroimaging studies of mental rotation: a meta-analysis and review. J. Cogn. Neurosci. 20, 1–19. doi:10.1162/jocn.2008.20013

Zhang, J., Kriegeskorte, N., Carlin, J.D., Rowe, J.B., 2013. Choosing the Rules: Distinct

and Overlapping Frontoparietal Representations of Task Rules for Perceptual Decisions. J. Neurosci. 33, 11852–11862. doi:10.1523/JNEUROSCI.5193-12.2013

Trial type	Initial ball position	Player A's goal	Player B's goal	Second switch configuration	Player A's action-plan	Player B's action-plan
1	0	0	0	0	1	0
2	0	0	1	0	1	1
3	0	1	0	0	0	1
4	0	1	1	0	0	0
5	0	0	0	1	1	1
6	0	0	1	1	1	0
7	0	1	0	1	0	0
8	0	1	1	1	0	1
9	1	0	0	0	0	0
10	1	0	1	0	0	1
11	1	1	0	0	1	1
12	1	1	1	0	1	0
13	1	0	0	1	0	1
14	1	0	1	1	0	0
15	1	1	0	1	1	0
16	1	1	1	1	1	1

Table 1: possible trial types. 0 = left (columns 2-4) or switch (columns 5-7). 1 = right (columns 2-4) or stay (columns 5-7)

	Initial ball position	Player A's goal	Player B's goal	Second switch configuration	Player A's action-plan	Player B's action-plan
Initial ball position	1	0	0	0	0	0
Player A's goal	0	1	0	0	0	0
Player B's goal	0	0	1	0	0	0
Second switch configuration	0	0	0	1	0	0
Player A's action-plan	0	0	0	0	1	0
Player B's action-plan	0	0	0	0	0	1

Table 2: correlation matrix between the factors shown in Table 1, across the 16 trial types

Region	**BA**	**Peak co-ordinate**	**Z_{max}**	**N voxels**	**Decoding accuracy**
Dorsal posterior frontal cortex	6/8	-36, -4, 47	3.81	61	51.4%
Superior parietal cortex	7	-21, -52, 56	5.07	290	51.4%
Posterior cingulate	23/30	-9, -58, 17	4.38	78	51.2%
Occipital cortex	18	-21, -88, 8	4.38	60	51.3%

Table 3: Regions of significant intention decoding. BA = approximate Brodmann Area. Decoding accuracies represent the mean of all searchlights with central voxels within each region. Furthermore, they represent the mean of eight separate decoding analyses, therefore a region with a decoding accuracy of 58% in one analysis and 50% in the remaining seven would have a mean decoding accuracy of 51%.

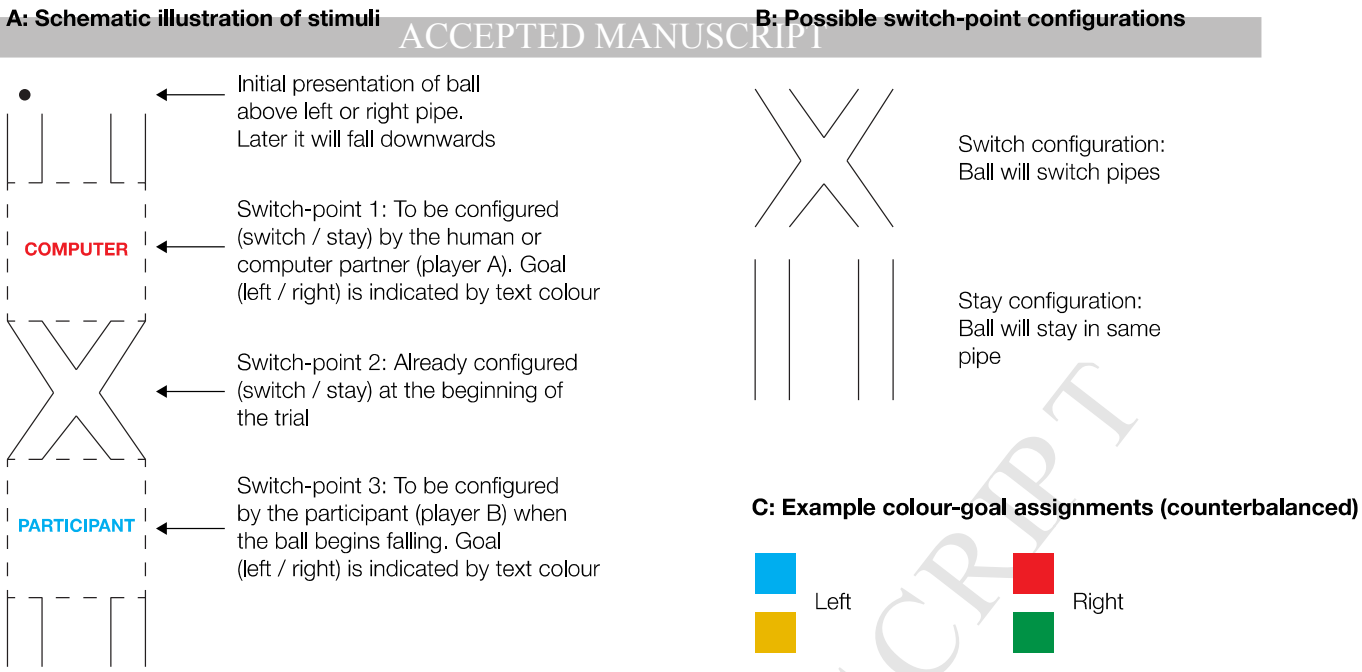


Figure 1. Schematic illustration of stimuli for behavioural task

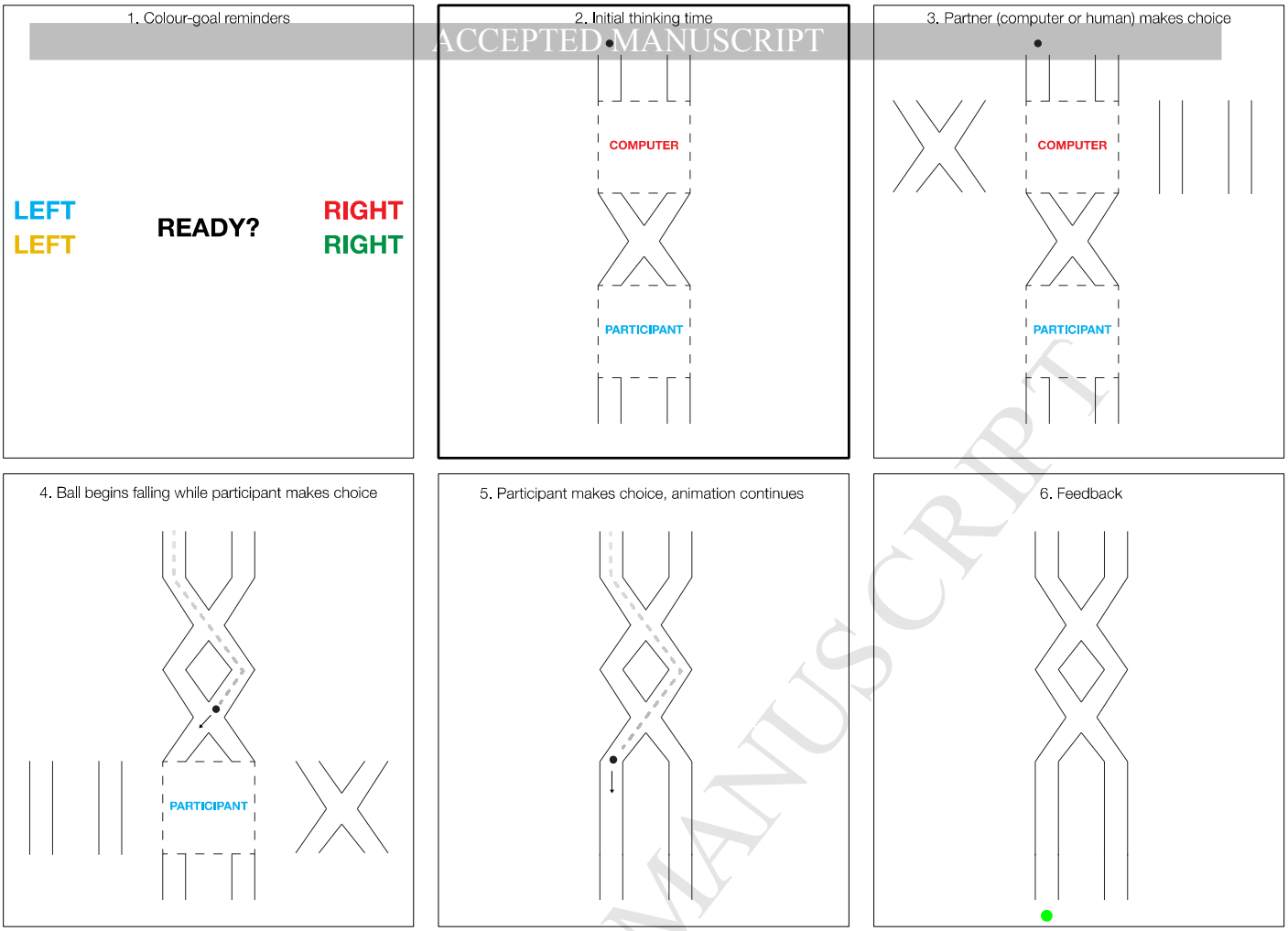


Figure 2. Sequence of events on one trial of the behavioural task. Colour-goal reminders were shown at the beginning of each block of 32 trials, then the subsequent five stages occurred on every trial. Only the initial thinking time, shown here with thicker border, was used for fMRI analyses.

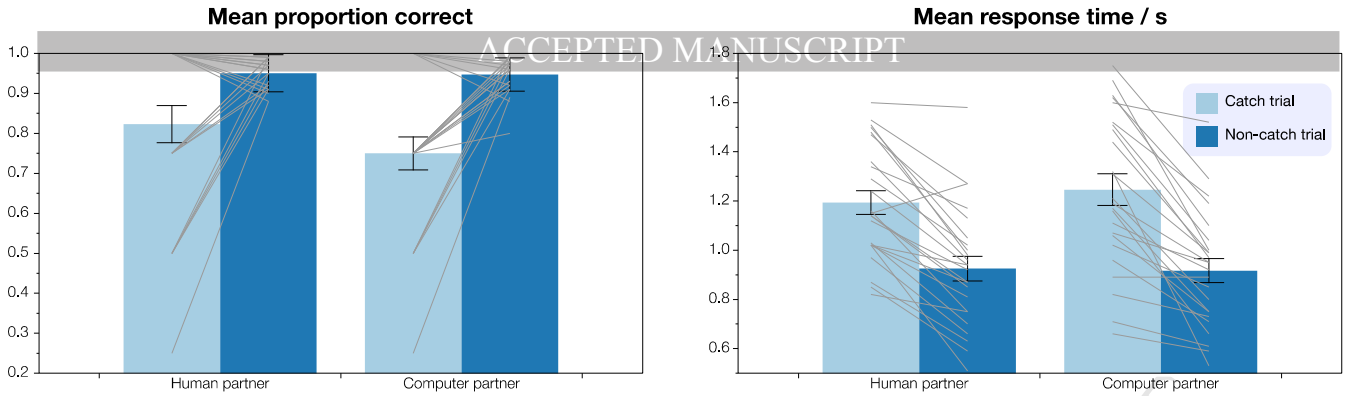


Figure 3. Behavioural results. Error-bars indicate 95% confidence intervals for the within-subject contrast between catch/non-catch trials (Loftus & Masson, 1994), such that nonoverlapping bars indicate a significant difference. Grey lines indicate data from individual participants. This shows a highly consistent pattern in the response time data. The pattern is less consistent for the accuracy data, likely because there are only four catch trials in each condition, hence the only possible values for catch-trial accuracy are 0, 0.25, 0.5, 0.75, and 1, whereas the non-catch trial accuracies can take a larger range of values.

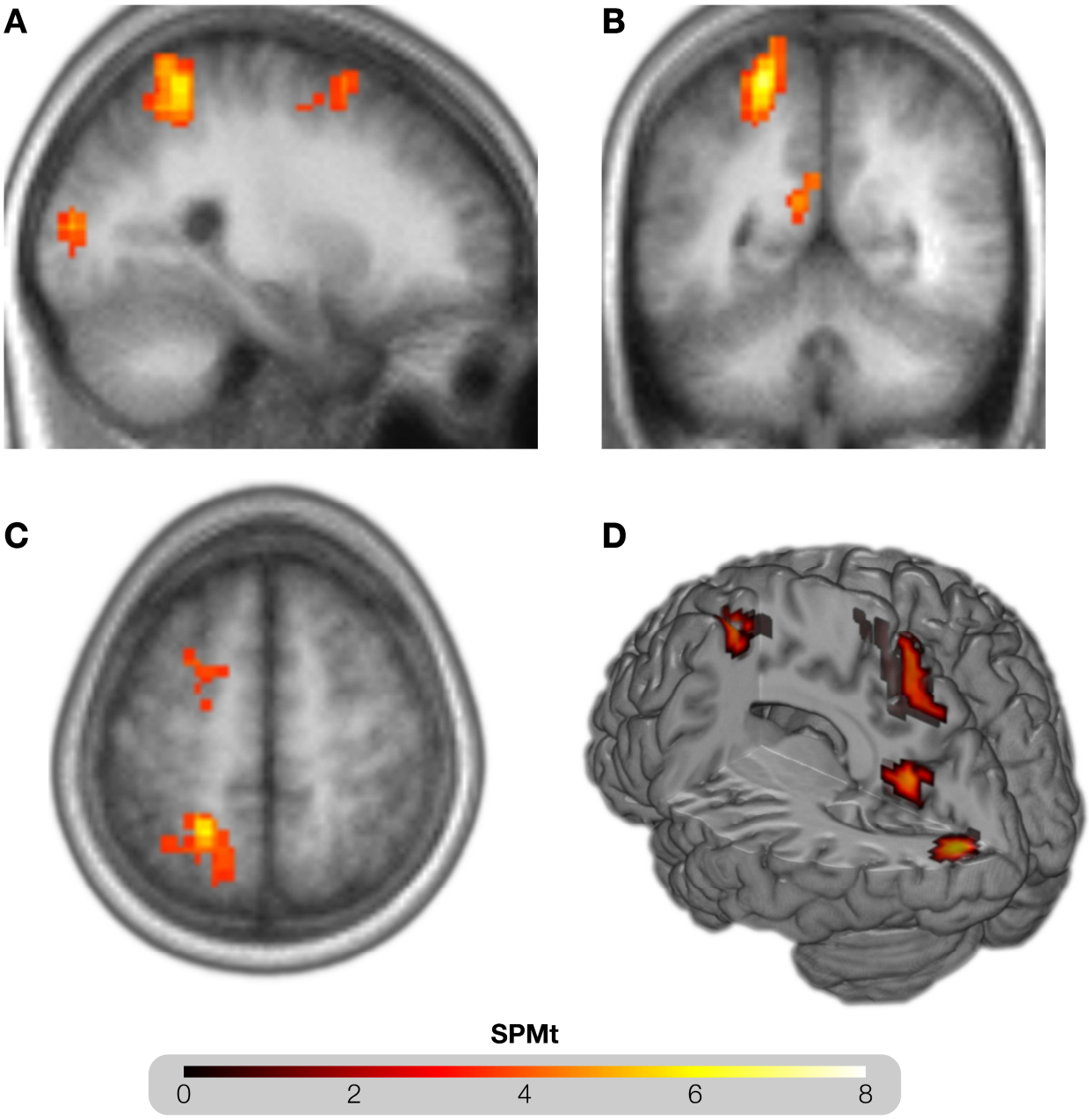


Figure 4. Regions of consistently above-chance intention decoding. Panels A-C display results on sagittal (x = -10), coronal (y = -50), and axial (z = 52) slices respectively of the mean normalised structural scan. Decoding was possible from four regions: occipital cortex (panel A), posterior cingulate (panel B), superior parietal cortex (panels A-C) and dorsal prefrontal cortex (panels A and C). Panel D shows a 3D rendering of results derived from MRIcroGL software.

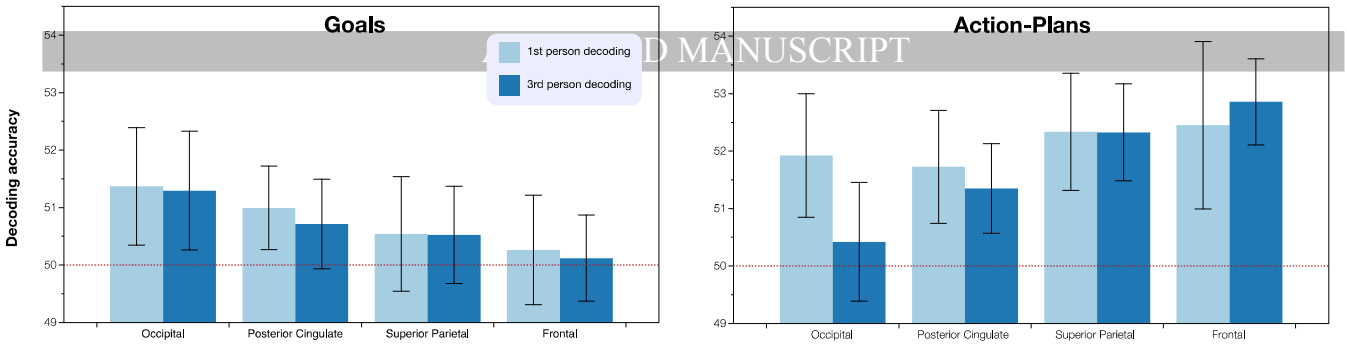


Figure 5. Accuracy of intention decoding analyses in four regions shown in Figure 4. Horizontal red line indicates chance accuracy. 1st-person decoding refers to classification of the scanned participant's intention. 3rd-person decoding refers to classification of their representation of the partner's intention. Error bars indicate 95% confidence intervals for the comparison of each bar with chance performance. Note that the regions of interest were selected on the basis of showing above-chance decoding accuracy in these analyses, therefore we refrain from significance-testing of individual bars seeing as results could be inflated by selection bias.

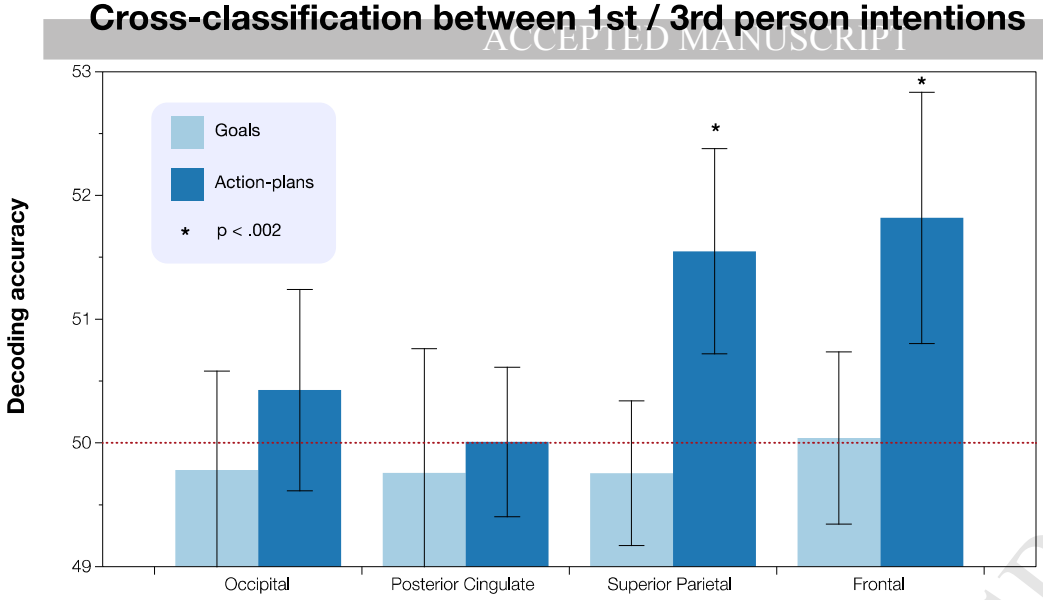


Figure 6. Cross-classification between 1st-person and 3rd-person intentions. Note that these analyses are unbiased with respect to the analyses used to generate regions of interest. Error-bars indicate 95% confidence intervals for the comparison of each bar against chance performance.

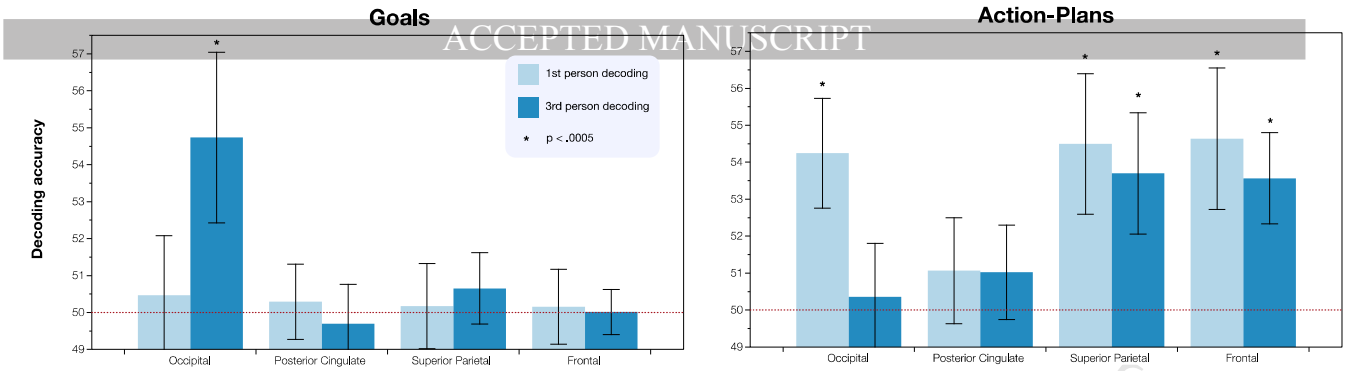


Figure 7. Cross-classification between visually presented stimulus-characteristics (initial presentation of ball on left vs right / initial configuration of second switch point as switch vs stay) and inferred intention components (goal: left vs right / action plan: switch vs stay). 1st-person decoding refers to decoding of the scanned participant's intention; 3rd-person decoding refers to decoding of their representation of their partner's intention. Error bars indicate 95% confidence intervals for the comparison of each bar against chance performance.