Visual perspective taking without visual perspective taking

Steven Samuel[1,2]

Madeline J. Eacott[1]

Geoff G. Cole[1]

1 Department of Psychology, University of Essex, U.K.

2 Department of Psychology, University of Plymouth, U.K.

Word count: 5442 (excluding abstract, acknowledgements, captions, references and tables, including footnotes)

Key words: perspective-taking, vision, theory of mind.

Address for correspondence: Steven Samuel. Department of Psychology, Portland Square, University of Plymouth, Drake Circus, Plymouth, Devon, PL4 8AA. Email: steven.samuel@plymouth.ac.uk

Abstract

What happens when an observer takes an agent's visual perspective of a scene? We conducted a series of experiments designed to measure what proportion of adults take a *stimulus-centered* rather than *agent-centered* approach to a visual perspective taking task. Adults were presented with images of an agent looking at a number (69). From the perspective of the viewer, the number appeared upside down. We then asked participants what number the agent saw. An agent-centered approach, i.e., one that takes into account the other's visual experience, should produce the correct answer '69'. Even an egocentric error (i.e., the participant's own perspective) would provide the same correct response. We were interested in what proportion of participants would give the incorrect answer '96', which is best explained by a stimulus-centered rather than agent-centered strategy, namely 'flipping' each digit one at a time from left to right. Crucially, such a strategy ignores the alternative visual perspective. We found that, on average, 12-21% of participants made this error. We discuss this finding in the context of the key questions around representation, content, and Theory of Mind in visual perspective taking.

Introduction

Visual perspective taking (VPT) concerns the ability to represent and/or make judgments about the viewpoint of another person, and it is often central to successful communication and interaction with others (Brown-Schmidt, Gunlogson, & Tanenhaus, 2008; Clark & Brennan, 1991; Linde & Labov, 1975). Despite decades of research going back at least as far as Piaget (Piaget & Inhelder, 1956), there is currently no formal model or theory of VPT (Cole & Millett, 2019). A crucial issue for any such model is the question of *what aspects of a viewpoint an observer can reliably represent* (Cole, Millett, Samuel, & Eacott, 2020). Consistent with what might be called an 'intuitive' view of VPT, some scholars have suggested that it is possible to simulate the visual experiences of others in quasi-perceptual, image-like form (Ward, Ganis, & Bach, 2019; Ward, Ganis, McDonough, & Bach, 2020). Others have suggested such representations are theoretically problematic (Cole & Millett, 2019; Cole et al., 2020), and that perspective-takers bring naive and often erroneous concepts of how vision works to bear on VPT problems (Samuel, Hagspiel, Eacott, & Cole, 2021). Some have broken the problem up by proposing two systems, one which spontaneously captures ('registers') simple visual links between agents and objects, and one which is effortful but can generate a richer, more detailed representation of appearance (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013).

An equally important but rarely-asked question is how frequently observers need to represent other visual perspectives in VPT. For instance, it has been demonstrated that some VPT problems can be solved by drawing imaginary lines from agents' eyes to objects, concluding that something is not seen if objects lie in their paths (Michelon & Zacks, 2006). Other VPT tasks might be tackled by relating target objects to the agent in the form of simple spatial propositions such as 'in front of', a process called object-centered spatial coding (Santiesteban, Shah, White, Bird, & Heyes, 2015). The ability to reverse spatial mappings for

someone in front of us can also solve certain VPT problems concerning the left/right axis (Yu & Zacks, 2017). Importantly, none of these strategies would conform with a reasonable definition of a representing another agent's *visual experience* (see Cole et al., 2020).

One possibility is that the strategies described above occur because the tasks are too simple to necessitate a representation of another agent's vision; they concern the relatively basic question of what is *visible*, not how things *appear*, usually defined as Level 1 and Level 2 perspective problems respectively (Flavell, Everett, Croft, & Flavell, 1981; Masangkay et al., 1974). A representation of visual experience should be more likely (and resistant to short-cut heuristics) when the stimulus appears different according to perspective, such as a 6 and a 9 (see Lurz, 2009, for an interesting discussion of this point in relation to comparative cognition). However, even in Level 2 VPT tasks it has been shown that adults sometimes apply strategies that *misrepresent* what an agent sees. For example, when viewing an agent looking at two lines of equal length but where one line is closer to the agent than the other, observers are as likely to judge the closer line to appear visually *shorter* to the agent as longer, despite this response contradicting not only the agent's perspective, but also geometric logic (Samuel et al., 2021). The explanation offered for this effect was that, rather than attempt to represent what the agent *saw*, participants applied an (erroneous) folk theory of how vision works. One such theory could be the belief that, since the two lines were in fact the same length, the more distant one is somehow visually 'stretched' to compensate for this knowledge. Results like these suggest that, even in Level 2 tasks, VPT need not elicit a representation of another agent's visual experience.

Results such as these also suggest that a closer examination of *errors* in VPT could provide useful information about how people approach VPT problems. However, tasks which are designed to measure accuracy or response times typically do not allow clear inferences about pathways to errors to be made. For example, even in the experiment with the identical

lines described above (Samuel et al., 2021), it is possible that participants had generated something that they *thought* was a representation of another agent's vision, even when this was inaccurate. In these circumstances it is difficult to know whether such an error should still be considered a 'representation', because the result fails to reflect the agent's experience. Part of the problem therefore also concerns how one defines a representation in the context of VPT.

We consider that, minimally, such a representation should be *agent-centered.* In essence, this is the consideration of the stimuli in terms of the *perception* of them. Such a representation need not be *true* (i.e., it need not be *accurate*) to be a representation. An agent-centered strategy fits neatly with the argument that VPT is a component of the ability to understand others' mental states (e.g., Apperly & Butterfill, 2009; Ferguson, Apperly, & Cane, 2017), namely our Theory of Mind (Premack & Woodruff, 1978). This contrasts with a *stimulus-centered* approach, in which an operation is performed on the stimuli in the belief that this will *lead* (indirectly) to a correct judgment about a perspective. This strategy is more difficult to identify with Theory of Mind, as the agent's visual experience does not have primacy. Applying this contrast to the experiment where an agent sees two identical lines but one is closer, some errors (judgments that the closer line looked shorter) may have arisen because participants formed a 'bad' representation but one that nevertheless was *conceived of* as a corresponding to the agent's perceptual experience. These errors would thus achieve minimal qualification as representations, by our definition. Others may have applied one or other erroneous rule sourced in a consideration of the stimuli themselves, such as the 'stretching' of the further line already described, or the application of an erroneous geometrical rule (closer things appear *smaller*). These approaches would instead be stimulus-centered. Note that we cannot know whether even *correct* responses were agent- or stimulus-

centered; they may have come about via a 'good' representation or a correct application of the geometric rule 'closer things appear larger'.

Ideally, therefore, it would be possible to know precisely *how* an observer arrived at a response. We therefore designed a paradigm in which one type of error would be best explicable in terms of a stimulus-centered approach. Participants were presented with a picture of a woman looking at a number. The woman's position meant that the number appeared upside-down, compared to the participant's viewpoint. When that number was a 6 from the participant's perspective, the answer to a question about what number the agent saw was therefore '9'. This answer could be arrived at either by representing the agent's visual perspective of the number (agent-centered) or by 'flipping' the number upside-down (stimulus-centered), meaning the response alone could not distinguish between these two strategies. This required a number that would generate a different response depending on strategy. An example of such a number is 69. The number '69' looks the same upside-down, and thus it is also 69 from the agent's perspective. Now the agent-centered and stimulus-centered approached could be distinguished, because only the latter can produce the error '96'. The question was, *what proportion* of participants would take the stimulus-centered route to solving a Level 2 VPT problem and therefore make this error? The outcome would serve as a measure of the frequency and therefore the importance of stimulus-centered rather than agent-centered, representational VPT.

2. All Experiments

2.1 General method

The experiment was performed online using Qualtrics. Participants were told at the start that they should maximize their browser window and switch off their phone/email/music and anything else distracting. They were told that the experiment was investigating people's

ability to accurately recall details of photos, and that they would be asked to view some images and answer questions about aspects of those images. After providing informed consent, participants entered information about their age, gender, and whether they were native English speakers. They were then told that they would be shown a photo (400^2 pixels, see Figure 1) and were instructed that they should pay attention to the photo for as long as it appeared. Participants always saw two images, one per trial. Each photo depicted a woman sitting down looking at a number on the floor. The precise question that participants were asked, when they responded, and whether they had their attention directed towards the agent, varied by experiment. Participants were then debriefed and the experiment ended. Total experiment time was approximately 2-4 minutes. An entirely new sample was recruited for each experiment.

3. Experiment 1

In Experiment 1, participants were informed that they would be shown images of an agent looking at a number. Each image appeared for three seconds before disappearing. When the image disappeared, participants were presented with a text box and were instructed as follows: "Please type the number the woman saw". Up to ten seconds were allowed to respond. In the first trial ('Trial 1') the participant saw a '6' which from the agent's perspective appeared to be a '9'. The correct response was thus '9'. Results from this trial would tell us whether participants were able to take the agent's perspective of the number, but not *how* they did so. It would also induct participants into the knowledge that what they see as a 6 is a 9 when viewed upside down, which could encourage the use of this information in a stimulus-centered response later. This was examined in the second trial ('Trial 2'). In this trial the participant saw '69', the same number that the agent saw. If participants successfully represented the agent's perspective, they should give the answer '69'. Note that giving the

correct answer does not *guarantee* that participants generated a representation, because the same response would be given if participants rotated the digits while maintaining their own frame of reference. A '69' response would also be given even if participants ignored the agent's perspective entirely and simply gave what they themselves saw (an egocentric error). However, a stimulus-centered strategy would be to 'flip' (invert) the digits individually according to the rule: 'a 6 looks like a 9/a 9 looks like a 6'. If participants use this strategy rather than attempting to consider the agent's perception, they will produce the erroneous response '96'. Note that *no-one saw* '96' - this response is best explained in terms of a stimulus-centered strategy of number-flipping. We could therefore be confident that any such errors came from participants who had not taken an agent-centered approach.

We were interested in the proportion of '96' responses rather than statistical analyses of '69' vs. '96' responses because, while the best explanation for a '96' response is clearly a stimulus-centred response, correct '69' responses can be arrived at via distinct strategies, meaning a comparison would not tell us anything interesting about the relative frequencies of different VPT strategies (or indeed an absence of strategy, in the case of fortuitously correct yet egocentric responses). Details of the preregistration of Experiment 1 can be found here: https://osf.io/a3hfn.

3.1 Participants

Examining proportions (Trial 2) comes with no specific strategy for generating a sample size. We therefore chose to conduct a power analysis based on Trial *1*, where participants were shown a single number ('6') and asked what the agent saw. This test was conducted in G*Power (version 3.1.9.5), and was based on a one-sample and one-tailed (V0/V1 of 1.5) chi square test with an alpha of .05 and power of .80. The test resulted in a desired N of 74. We chose a one-tailed test because we expected participants to be more

accurate than not on Trial 1. Adults consistently show above-chance accuracy (approx. 80% and higher) when taking others' perspectives of single digits, including when the angle of perspective is 180 degrees (Surtees, Apperly, & Samson, 2013a, 2013b; Surtees, Butterfill, & Apperly, 2012), and even when there is a second, distractor digit (Samuel, Cole, & Eacott, 2020; Samuel, Legg, Manchester, Lurz, & Clayton, 2019). The reported tests, however, come from two-tailed tests of significance for conservatism.

We recruited participants using Prolific Academic, requiring that they used a laptop or desktop computer (which could not be easily physically re-oriented), were aged 18-35, and spoke English as a first language. We recruited the suggested 74 participants (M_{Age} = 28, range 18-34, 1 non-binary, 27 male, 46 female). Ethical approval for the study was received from the University of Essex Psychology Ethics Committee. All participants were compensated equally for their time.

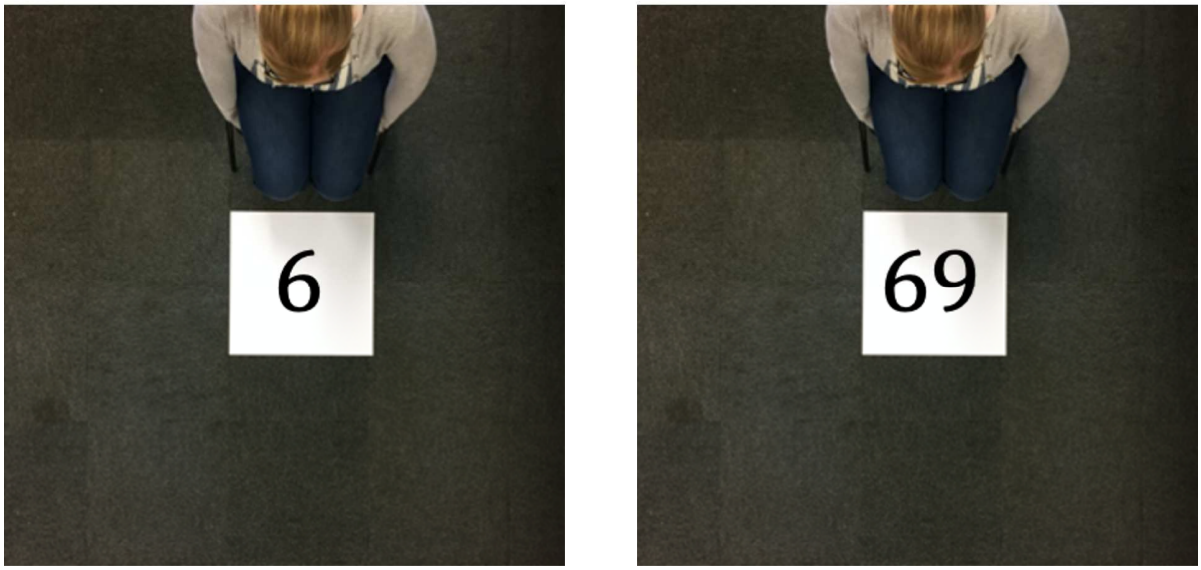

Figure 1. The images used in Trial 1 (left) and Trial 2 (right) in all four experiments.

3.2 Results and Discussion

Trial 1. Of the 74 participants, 42 (56.8%) correctly stated that the woman saw '9', 32 (43.2%) did not, with all but one of these saying she saw the same number they saw ('6'), the exception giving a '2' response. This difference was not statistically significant, Chi Sq (1,74) = 1.351, p = .245.

Trial 2. Of the 74 participants, 65 (87.8%) correctly gave a '69' response, and 9 (12.2%) stated that the woman saw '96', a number that no-one saw. This suggests that these participants 'flipped' each of the two digits in a left-right sequence, a stimulus-centered approach[1]. Of these nine participants, the majority (seven participants or 78%) had provided an accurate response on Trial 1, and had therefore previously responded that what looked like a 6 to themselves looked like a 9 to the agent. Two (22%) had previously provided an egocentric error (‘6’) on Trial 1.

[1] We did not pre-register any statistical analyses of the results of Trial 2, only the reporting of proportions. However, for completeness the results of Trial 2 were as follows, always favouring a minority of '96' responses: Experiment 1: Chi Sq (1,74) = 42.378, p < .001; Experiment 2: Chi Sq (1,74) = 28.595, p < .001; Experiment 3: Chi Sq (1,76) = 25.474, p < .001; Experiment 4: Chi Sq (1,74) = 70.054, p < .001. Similarly, we did not pre-register analyses of the conditional probabilities of providing a 'flipped' response on Trial Two following either a correct or incorrect response on Trial One. Please note that these tests are based on very small sample sizes and should therefore be interpreted with caution. These were: Experiment 1: Chi Sq (1,9) = 2.778, p = .1; Experiment 2: Chi Sq (1,14) = 10.286, p = .014 (please see the relevant results sections for descriptives). In Experiments 3 and 4 all those who gave the 'flipped' response were correct on Trial One.

In sum, 12.2% of participants demonstrated a stimulus-centered approach, 'flipping' the digits on Trial 2 rather than considering any true perspective of the stimulus. This proportion puts those participants who did this in a clear minority. Unexpectedly, on Trial 1 participants were about as likely to indicate what *they* saw (a '6') as they were to provide the correct '9' response. This suggests that even taking someone's perspective of a single digit was difficult. However, we cannot know whether the incorrect responses on Trial 1 are absences of representations or simply *bad* representations.

In a second experiment, we changed the question for Trial 1 so that participants were now asked to give their *own* perspective of the number ('6'). This was done for two reasons. Firstly, it allowed us to assess whether the stimulus-centered strategy found in Trial 2 required the induction to the invertibility of 6/9 on Trial 1, which was suggested by the fact that only two participants who gave incorrect responses on Trial 1 went on to give a stimulus-centered response on Trial 2. In other words, would participants use the stimulus-centered approach as a 'starting strategy', the first time they are asked to take another perspective? Secondly, a correct response on Trial 1would also demonstrate that the difficulty of giving the correct response in Experiment 1was not simply due to forgetting the number in the picture. In addition, in an exploratory test of the data from Experiment 1 we counted the number of '96' responses by gender, with four given by males, four by females, and one by a non-binary individual. Since males comprised a smaller proportion of the sample, these figures corresponded to 15% of males and 9% of women. There was thus a hint that males may have a greater tendency to apply the stimulus-centered strategy. This would be consistent with research that finds females are better empathizers and embodiers of perspectives generally (Baron-Cohen, 2002; Kessler & Wang, 2012). We therefore recruited equivalent numbers of males and females to provide more balanced data on this matter to assess whether it was deserving of more formal attention (i.e., confirmatory testing).

4. Experiment 2

Details of the preregistration of Experiment 2 can be found here: https://osf.io/mf9tc. Data are available in the supplemental materials.

4.1 Method

Only two changes were made from Experiment 1. We recruited an equal number of males (37) and females (37), Final N = 74, M_{Age} = 27. range 18-35, and we changed the question for Trial 1 so that it now said the following: "Please type the number *you* saw" (italics new). Thus, Trial 1 now assessed participants' ability simply to recall the number they themselves saw. For Trial 2, the original question ("Please type the number *the woman* saw") was retained, but with added italics to draw attention to the change from Trial 1.

4.2 Results and Discussion

Trial 1. Of the 74 participants, 65 (87.8%) correctly stated that they had seen a '6', and 9 (12.2%) did not, with all but one of these giving the number the woman saw ('9'), the exception giving a single '3' response. This difference was statistically significant, Chi Sq (1,74) = 42.378, $p < .001$.

Trial 2. Of the 74 participants, 14 (18.9%) stated that the woman saw '96', which was a number that no-one saw. This again suggests that participants 'flipped' each of the two digits in a left-right sequence. Six of these 14 were male, 8 were female. Of these fourteen participants, all but one (93%) had provided an accurate (and this time *egocentric*) response on Trial 1. Of the rest, 59 (79.7%) gave a correct '69' response, and one gave an incorrect '6' response (1.4%).

In Experiment 2 18.9% of participants had 'flipped' the digits on Trial 2 rather than considered any true perspective of them, a slightly larger proportion than previously. Interestingly, this meant the stimulus-centered strategy was applied even when participants were not previously inducted to the invertibility of 6/9 in Trial 1. With a balanced quota of males and females, there was no evidence that males preferred the stimulus-centered approach; instead, two more females than males made this error. Finally, the high accuracy rate on Trial 1 rules out the possibility that participants easily forget the number in the image, and therefore failures to respond correctly on Trial 1 in the previous Experiment are more likely to be failures of perspective taking rather than failures of recall.

An interesting outcome of Trial 1 was that a minority of 8 participants (10.8%) gave the number that the agent saw instead. This was contrary to the explicit instruction to provide the number they themselves saw. This type of error is consistent with evidence from studies investigating 'spontaneous' perspective taking, in which an individual's ability to act egocentrically is compromised when they are aware of an alternative perspective that conflicts with their own (e.g., Samson, Apperly, Braithwaite, Andrews, & Bodley Scott, 2010). However, most such studies take care not to draw explicit attention to the other agent in the scene. In contrast, in our task we state just prior to the presentation of the images that what will follow is a photo in which is woman is looking at a number. Our 10.8% might therefore have been directed to the agent's perspective by means of this textual prime rather than through spontaneous VPT. We tested this possibility in Experiment 3, in which we removed reference to the woman and what she was looking at from the text for Trial 1 (Trial 2 remained the same, as it explicitly concerned the woman’s perspective and thus no response could be classed as 'spontaneous' perspective taking). We hypothesized that the proportion of participants who gave a '9' response on Trial 1 would decline—numerically rather than statistically given the already low numbers—with these references removed.

5. Experiment 3

Details of the preregistration of Experiment 3 can be found here: https://osf.io/8seqt.

5.1 Method

Only one change was made from Experiment 2. We removed reference to the agent in the photo from the introduction to Trial 1. The Intro now simply read: "Now, you will be shown a photo. Please pay attention to the photo for as long as it appears." Trial 2 remained unchanged (reference to the agent remained). Due to an error one extra female participant was recruited, and we therefore recruited one extra male participant for balance (final N = 76, M_{Age} = 25, range 18-35, 38 male, 38 female)

5.2 Results and Discussion

Trial 1. Of the 76 participants, 73 (96.1%) correctly stated that they had seen a '6', and 3 (3.9%) did not, giving the number the woman saw ('9'). This difference was statistically significant, Chi Sq (1,76) = 64.474, $p < .001$.

Trial 2. Of the 76 participants, 16 (21.1%) stated that the woman saw '96'. Six of these were male, 10 were female. All of these had responded correctly (i.e., egocentrically) on Trial 1. Of the rest, 57 (75%) gave a correct '69' response, one gave '9', one '6', one '95' (1.3% each).

In Experiment 3, 21.1% of participants demonstrated a stimulus-centered strategy on Trial 2 rather than considered any true perspective of them, a slightly larger proportion (again) than previously. Additionally, removing references to the agent or her perspective from the text led to a decline in responses made from her perspective on Trial 1, down from 10.8% to a quite negligible 3.9%.

Overall, across three experiments the proportion of participants who applied an (erroneous) stimulus-centered strategy had ranged from a low of 12.2% (Experiment 1) to a high of 21.1% (Experiment 3). In a final experiment, we examined whether this strategy was restricted to circumstances where the scene could no longer be viewed when responding. Previous research has shown that keeping pictures visible when making responses in VPT tasks does *not* increase accuracy relative to responding after a picture disappears; however, the stimuli in question were lines and not alphanumeric stimuli, which have been speculated to be processed more easily than abstract shapes (Samuel et al., 2021). In Experiment 4 we therefore repeated Experiment 1 but allowed participants to make their responses while viewing the images, with no time limits.

6. Experiment 4

Due to an error, Experiment 4 was not pre-registered. However, the method and analyses are consistent with the previous experiments and pre-registrations.

6.1 Method

Only two changes were made from Experiment 1. First, we recruited equal numbers of men and women, as per Experiments 2 and 3. Second, instead of showing the critical image for a fixed period of time before the response prompt appeared, this time the response box was presented beneath the image, and no time limits were set. The Intro now simply read: "Now, you will be shown a photo. In the photo is a woman looking at a number." The prompt beneath the image on the next screen was: "Please type the number the woman sees". Of the 74 participants, 37 were male, 37 female (M_{Age} = 27, range 18-35).

6.2 Results and Discussion

Trial 1. Of the 74 participants, 62 (83.8%) correctly responded that the woman saw a '9', and 12 (16.2%) did not, giving the number they saw ('6'). Unlike in Experiment 1, this difference was statistically significant, Chi Sq (1,74) = 33.784, $p < .001$.

Trial 2. Of the 74 participants, all but one (1.4%) correctly responded that the woman saw '69', with the sole (female) exception giving the number-flipped error '96'. This participant had responded correctly (i.e., that the agent saw a 9) on Trial 1.

In Experiment 4 only one participant demonstrated a clearly stimulus-centered approach on Trial 2. Additionally, participants were now much more likely to give an accurate response to the perspective-taking question in Trial 1 in this Experiment than in Experiment 1. Overall, it was much easier to solve these VPT problems accurately when it was possible both to view the scenes and respond at leisure.

7. General Discussion

In Experiment 1, when asked what single digit another agent saw, adults were about as likely to respond with the number they themselves saw ('6') as the number the agent saw ('9'). While we can reasonably surmise that those who were incorrect made an egocentric error—we cannot know how those 56% who *did* answer correctly came to their response. They may have generated a representation, applied knowledge that a 6 looks like a 9 when upside down, mentally rotated the number 180 degrees while maintaining their own perspective of the scene, etc. In Experiments 1-3, when asked about the agent's perspective of '69', most participants (between 75%-88%) correctly gave the answer '69'. Again, we cannot know how this answer was arrived at, or even if it was a fortuitous egocentric error. Crucially, between 12.2%–21.1% responded with '96', which is a number that no-one saw but can be explained by a stimulus-centered strategy of 'flipping' the numbers '6' and '9' individually. This occurred not only when participants were inducted to the reversibility of

the digit (Exp 1) but also when they were considering another perspective for the first time (Exps 2-3). This suggests that stimulus-centered strategies are also *starting* strategies for VPT tasks. The data therefore show that between 12.2%–21.1% of adults in these experiments derived a response to a VPT task without taking anyone's visual perspective at all (even their own). While this is clearly a minority, since it is unclear whether *correct* responses involved representations this cannot be an *overestimate,* but could be an *underestimate*. Indeed, coupled with the low accuracy (56.8%) on even the first trial in Experiment 1, the data suggest that representations might even occur in only a *minority* of cases on this Level 2 VPT task. Less speculatively, these results militate against the possibility that Level 2 VPT problems are *necessarily* tackled using an agent-centered rather than stimulus-centered approach. By extension, they also suggest that VPT, even Level 2 VPT, need not engage one's Theory of Mind, and that there is no single, dedicated process for VPT questions concerning appearance.

Nevertheless, these conclusions come with a significant caveat because, in Experiment 4, when it was possible to view the image and respond at leisure, the vast majority got the answer correct on both trials. Evidence of 'number flipping' fell to a single participant out of 74. This points to a potential distinction between VPT based on a scene *being remembered* and VPT based on scene *being perceived*, with a stimulus-centered approach more likely for the former than latter. A good explanation for this distinction is not immediately apparent. However, we can rule out two possibilities. Firstly, since the images shown to participants were clearly not live scenes, we can exclude an explanation by which it might be easier to take 'real-time' perspectives. We can also exclude the possibility that participants failed to notice or recall the number. This is because i) participants could key in any number they wished but only two responses in the first three (timed) experiments included any numbers other than 6 and 9; and ii) the vast majority of participants correctly

recalled the precise number they saw when asked in Experiments 2 and 3. If we consider the present data alone, it would thus appear that we are left with two possibilities. First, it might be easier to come up with an accurate response to a VPT problem while the agent and the target stimulus are viewable. Second, it might be easier to come up with an accurate response to a VPT problem when an observer is under no time pressure.

However, a comparison of these results with those from another, similar study with abstract rather than alphanumeric stimuli seems to favour a third possibility. Recall from the Introduction that, in a series of previous experiments, participants were presented with an agent looking at two identical lines (Samuel et al., 2021). Results showed that adults often failed to judge that the closer of two identical lines would appear visually longer to an agent. Importantly, accuracy was no better if responses were made when both the agent and stimuli were viewable while making a response. This contrast is thus very similar to that between the present Experiments 1 and 4, with the exception that there was always a ten-second time limit in the other study, which was ample for the task at hand. However, in Experiment 1, Trial 1 of the present studies, 42% of participants gave an erroneous egocentric response, but in Experiment 4 this figure decreased to 16%. Additionally, evidence of stimulus-centered strategies in Trial 2 was almost non-existent. The instructions participants were given could not explain this difference, as they did not change between experiments. Instead, an explanation by which increasing the salience of the agent serves to facilitate perspective taking with specifically *alphanumeric* stimuli is the better candidate. We typically expect intentional agents to position themselves where they can comfortably read such characters, meaning that agents and characters typically predict each others' orientations. However, we do not have the same expectation for lines, shapes and other stimuli which have no intrinsic 'upright' orientation. We therefore speculate that the viewability of the agent while responding in Experiment 4 increased the salience of her positional cue, increasing accuracy

and by extension decreasing egocentricity on Trial 1, and all but eliminating the use of stimulus-centered strategies on Trial 2.

It is interesting that the level of egocentricity found in Trial 1, Experiment 1, was higher than expected (42%). Although error rates from some VPT tasks with adults have been reported around this level (e.g., Apperly et al., 2010; Samuel et al., 2021; Wardlow, 2013; Wu & Keysar, 2007), it is unusual for studies with numbers as stimuli. We have just speculated that it is easier for participants to take other agents' perspectives of alphanumeric characters than abstract shapes, and thus it may appear that this result runs contrary to this hypothesis. There are however a number of reasons why direct comparisons between accuracy rates in Experiments 1-3 and accuracy rates in many other tasks in the literature is made difficult. Firstly, the agent and stimulus here were almost certainly *less* salient than in most other VPT studies in the literature because they were presented for three seconds alone and were inaccessible while responding. This is relatively unusual for VPT tasks, which usually test participants while the relevant scene is being viewed (Apperly et al., 2010; Michelon & Zacks, 2006; Wardlow, 2013; Wu & Keysar, 2007), with some exceptions (Samuel, Frohnwieser, Lurz, & Clayton, 2020; Samuel et al., 2021). Secondly, participants did not know what they were going to be asked to do until the picture had disappeared. Again, this is different from the majority of explicit VPT tasks, in which participants are given prior instructions to take an agent's perspective, and therefore what to attend to. Thirdly, participants in our experiments saw a total of two trials, and in Experiments 2 and 3 only one of these required perspective taking. Other VPT paradigms often employ multiple trials (Samuel, Cole, et al., 2020; Samuel et al., 2019; Surtees et al., 2013a, 2013b; Surtees et al., 2012). More important for our argument about the importance of salience is therefore the *internal* comparison between Experiments 1 and 4.

In conclusion, we found that adults sometimes (as a minimum, approx. 12-21% of the time) apply a stimulus-centered strategy to a VPT task, one which does not comply with a definition of a representation as being agent-centered. However, this finding was limited to instances where the agent and stimulus were not visible at the time of responding. When they *were* visible, accuracy was very high and evidence for this strategy almost disappeared. Our findings therefore suggest that adults sometimes come up with answers to some VPT problems without representing, either accurately *or* inaccurately, another agent's visual perspective. *Which* problems may depend on task-specific factors such as the salience of the agent and the type of stimulus being viewed.

Acknowledgements: None

Authors' note: All data are available in the supplemental materials. Materials will be published on the OSF. Details of pre-registrations can be found here: https://osf.io/a3hfn (Experiment 1); https://osf.io/mf9tc (Experiment 2); https://osf.io/8seqt (Experiment 3).

Disclosure of interest: The authors report no conflicts of interest.

References

Apperly, I. A., & Butterfill, S. A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological review, 116*(4), 953.

Apperly, I. A., Carroll, D. J., Samson, D., Humphreys, G. W., Qureshi, A., & Moffitt, G. (2010). Why are there limits on theory of mind use? Evidence from adults' ability to follow instructions from an ignorant speaker. *Quarterly Journal of Experimental Psychology, 63*(6), 1201-1217.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in cognitive sciences, 6*(6), 248-254.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition, 107*(3), 1122-1134.

Butterfill, S. A., & Apperly, I. A. (2013). How to construct a minimal theory of mind. *Mind & Language, 28*(5), 606-637.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. *Perspectives on socially shared cognition, 13*(1991), 127-149.

Cole, G. G., & Millett, A. C. (2019). The closing of the theory of mind: A critique of perspective-taking. *Psychonomic bulletin & review*, 1-16.

Cole, G. G., Millett, A. C., Samuel, S., & Eacott, M. J. (2020). Perspective taking: In search of a theory. *Vision, 4*(2), 30.

Ferguson, H. J., Apperly, I., & Cane, J. E. (2017). Eye tracking reveals the cost of switching between self and other perspectives in a visual perspective-taking task. *Quarterly Journal of Experimental Psychology, 70*(8), 1646-1660.

Flavell, J. H., Everett, B. A., Croft, K., & Flavell, E. R. (1981). Young children's knowledge about visual perception: Further evidence for the Level 1–Level 2 distinction. *Developmental psychology, 17*(1), 99-103.

Kessler, K., & Wang, H. (2012). Spatial perspective taking is an embodied process, but not for everyone in the same way: differences predicted by sex and social skills score. *Spatial Cognition & Computation, 12*(2-3), 133-158.

Linde, C., & Labov, W. (1975). Spatial networks as a site for the study of language and thought. *Language*, 924-939.

Lurz, R. (2009). If chimpanzees are mindreaders, could behavioral science tell? Toward a solution of the logical problem. *Philosophical Psychology, 22*(3), 305-328.

Masangkay, Z. S., McCluskey, K. A., McIntyre, C. W., Sims-Knight, J., Vaughn, B. E., & Flavell, J. H. (1974). The early development of inferences about the visual percepts of others. *Child Development*, 357-366.

Michelon, P., & Zacks, J. M. (2006). Two kinds of visual perspective taking. *Perception & Psychophysics, 68*(2), 327-337.

Piaget, J., & Inhelder, B. (1956). *The child's concept of space*: Routledge & Paul.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences, 1*(4), 515-526.

Samson, D., Apperly, I. A., Braithwaite, J. J., Andrews, B. J., & Bodley Scott, S. E. (2010). Seeing it their way: evidence for rapid and involuntary computation of what other people see. *Journal of Experimental Psychology: Human Perception and Performance, 36*(5), 1255-1266.

Samuel, S., Cole, G. G., & Eacott, M. J. (2020). Two independent sources of difficulty in perspective-taking/theory of mind tasks. *Psychonomic bulletin & review*, 1-7.

Samuel, S., Frohnwieser, A., Lurz, R., & Clayton, N. (2020). Reduced egocentric bias when perspective-taking compared to working from rules. *Quarterly Journal of Experimental Psychology*.

Samuel, S., Hagspiel, K., Eacott, M. J., & Cole, G. G. (2021). Visual perspective-taking and image-like representations: We don't see it. *Cognition, 210*.

Samuel, S., Legg, E., Manchester, C., Lurz, R., & Clayton, N. (2019). Where was I? Taking alternative visual perspectives can make us (briefly) misplace our own. *Quarterly journal of experimental psychology (2006)*, 1747021819881097.

Santiesteban, I., Shah, P., White, S., Bird, G., & Heyes, C. (2015). Mentalizing or submentalizing in a communication task? Evidence from autism and a camera control. *Psychonomic bulletin & review, 22*(3), 844-849.

Surtees, A., Apperly, I., & Samson, D. (2013a). Similarities and differences in visual and spatial perspective-taking processes. *Cognition, 129*(2), 426-438.

Surtees, A., Apperly, I., & Samson, D. (2013b). The use of embodied self-rotation for visual and spatial perspective-taking. *Frontiers in human neuroscience, 7*, 698.

Surtees, A., Butterfill, S. A., & Apperly, I. A. (2012). Direct and indirect measures of level-2 perspective-taking in children and adults. *British Journal of Developmental Psychology, 30*(1), 75-86.

Ward, E., Ganis, G., & Bach, P. (2019). Spontaneous vicarious perception of the content of another's visual perspective. *Current Biology, 29*(5), 874-880. e874.

Ward, E., Ganis, G., McDonough, K. L., & Bach, P. (2020). Perspective taking as virtual navigation? Perceptual simulation of what others see reflects their location in space but not their gaze. *Cognition, 199*, 104241.

Wardlow, L. (2013). Individual differences in speakers' perspective taking: The roles of executive control and working memory. *Psychonomic bulletin & review, 20*(4), 766-772.

Wu, S., & Keysar, B. (2007). The effect of culture on perspective taking. *Psychological Science, 18*(7), 600-606.

Yu, A. B., & Zacks, J. M. (2017). Transformations and representations supporting spatial perspective taking. *Spatial Cognition & Computation, 17*(4), 304-337.