LOW-LEVEL ACTION SCHEMAS SUPPORT GAZE GUIDANCE BEHAVIOR FOR ACTION PLANNING AND EXECUTION IN NOVEL TASKS

A PREPRINT

Ashima Keshava
University of Osnabrück
Germany
akeshava@uos.de

Farbod Nosrat Nezami
University of Osnabrück
Germany

Henri Neumann
SALT AND PEPPER
Software GmbH & Co.KG.
Germany

Krzysztof Izdebski
SALT AND PEPPER
Software GmbH & Co.KG.
Germany

Thomas Schüler
SALT AND PEPPER
Software GmbH & Co.KG.
Germany

Peter König
University of Osnabrück
University Medical Center,
Hamburg-Eppendorf
Germany

February 9, 2022

ABSTRACT

Eye movements in the natural environment have primarily been studied for over-learned activities such as tea-making, sandwich-making, and hand-washing, which have a fixed sequence of actions associated with them. These studies indicate an interplay of low-level cognitive schemas that facilitate task completion. However, it is unclear if this strategy is also in play when the task is novel and a sequence of actions must be planned. Here, to study attention mechanisms, we recorded gaze and body movements in a naturalistic sorting task. In a virtual environment, subjects moved objects on a life-size shelf to achieve a given order. To study the action planning and execution related gaze guidance behavior we also controlled the complexity of the tasks, i.e. the objects were to be sorted based on a single or two different object features. We show that subjects are close to optimal while performing the easy tasks and are more sub-optimal while performing the more complex tasks. Fixations aligned with action onset showed gaze is tightly coupled with the action sequence and less so with task complexity. Task complexity elicited differences at specific time points where oculomotor behavior could be categorized into searching, guiding, checking fixations. Task complexity also affected the scan-paths on the task-relevant regions of interest during action planning and execution where subjects exhibit a greater search and action monitoring behavior in more complex tasks. Furthermore, task complexity also affected the timing of first fixations on the task-relevant regions of interest for action planning but not for action execution. Finally, the scan paths and the temporal sequence of first fixations suggest a just-in-time strategy of gaze allocation. Based on the spatio-temporal features of gaze control, we can conclude that eye movements during novel tasks correspond closely to low-level cognitive schemas that support the production and execution of actions.

Keywords action planning, gaze-guided planning, gaze control, action schemas, virtual reality

Author Summary

Eye movements in the natural environment have primarily been studied for over-learned and habitual everyday activities (tea-making, sandwich-making, hand-washing) which have a fixed sequence of actions associated with them. These studies typically categorize eye movements related to low level action plans of locating the object of interest, directing the body or hand to the object, and monitoring the action execution. In this study, we were interested in generalizing these task-oriented eye movements for novel tasks that are not associated with an inherent action sequence. To that end, we asked subjects to sort objects based on object features on a life-size shelf in a virtual environment as we recorded their eye and body movements. We then analyzed gaze features relative to the action onset i.e. when the hand made contact with an object. By only studying the sequence of fixations based on past, current and next actions, we aimed to elucidate the general characteristics of gaze behavior relative to action planning and execution and independent of object identity. With our analysis, we could categorize fixations into searching, guiding, monitoring fixations in a data-driven manner. Based on the spatio-temporal features of gaze control, we can conclude that novel tasks predominantly elicit eye movements corresponding to low-level action plans that support the production and execution of actions.

1 Introduction

In the pragmatic turn in cognitive science, there is a significant push towards incorporating the study of cognitive processes while interacting with the external world (Parada and Rossi, 2020). Moreover, Engel et al. (2013) proposed that cognition encompasses the body, and in turn, bodily action can be used to infer cognition. To this effect, understanding the control of eye movements in natural, real-life situations requires a mobile setup that allows for a subject to be recorded in tandem with voluntary actions in a controlled yet unconstrained environment. tudying eye movements in mobile subjects might give us a richer picture of cognitive processing in more naturalistic settings.

Humans actively use vision during everyday activities to gather and refine information about the environment. Since the seminal works of Yarbus (2013) and Buswell (1935) there has been consistent evidence that eye movements depend on the viewing task the observer is performing. Kollmorgen et al. (2010) demonstrated stimulus-dependent features, spatial viewing biases, and task-dependent features all influence the exploration of a visual scene. This is further supported by studies that emphasize the relevance of semantics in the guidance of eye movements (Henderson and Hayes, 2017; Einhäuser et al., 2008). Thus, we have growing evidence that task demands affect eye movement behavior.

Seminal studies have investigated eye movement behavior in the natural environment with fully mobile participants. In the pioneering studies of Land et al. (1999) and Hayhoe et al. (2003), subjects performed everyday activities of tea-making and sandwich-making, respectively. These studies required a sequence of actions that involved manipulating objects one at a time to achieve the goal. Both studies showed that nearly all the fixations were task-related. Further studies investigated eye movements under a plethora of natural conditions while walking (Matthis et al., 2018), driving (Mars and Navarro, 2012; Sullivan et al., 2012; Navarro et al., 2020), hand-washing (Pelz and Canosa, 2001), hitting a

ball (Land and McLeod, 2000), and free exploration (Schumann et al., 2008). These experiments in naturalistic settings have revealed several distinct functions of the eye movements during habitual everyday tasks.

Studies investigating habitual tasks uncovered a systematic timing of visual fixations and object manipulation. Specifically, fixations are made to target objects about 600ms before manipulation. Importantly, Ballard et al. (1995) proposed a "just-in-time" strategy that universally applies to this relative timing of fixations and actions. In other words, fixations that provide information for a particular action immediately precede that action and are crucial for the fast and economical execution of the task.

Moreover, while performing habitual tasks, Land and Hayhoe (2001) fixations have broadly categorized into four functional groups. “Locating” fixations retrieve visual information. “Directing” fixations acquire the position information of an object, accompany a manipulation action and facilitate reaching movements. “Guiding” fixations alternate between objects being manipulated, e.g., fixations on the knife, bread, and butter while buttering the bread. “Checking” fixations monitor where the task constraints in the scene have been met. These findings have also been corroborated by Pelz and Canosa (2001) and Mennie et al. (2007). Pelz and Canosa (2001) showed a similar just-in-time strategy of gaze allocation while performing a hand-washing task. They also reported a small number of fixations of about 5% that did not serve the immediate sub-task but rather provided the information needed for future action. The authors hypothesize these “look-ahead” fixations provide a mechanism to stabilize the visual input stream that results from a sequence of actions, facilitate task-switching, and reduce the conscious effort required to complete the actions in a given sequence. Hence, look-ahead fixations can be explained as a perceptual strategy to ease the cognitive load of attending to complex tasks in the real world. In sum, the wide-ranging functions of eye movements are well documented in natural routine tasks.

Based on the above observations Land and Hayhoe (2001) proposed a framework that outlines the flow of visual and motor control during task execution Figure 1A. The process summarizes the various operations that *must* occur during an ’object-related action’ i.e., individual actions performed on separate objects to achieve the desired goal. Each schema "specifies the object to be dealt with, the action to be performed on it, and the monitoring required to establish that the action has been satisfactorily completed." (Land, 2006). Further, the gaze control samples the information about the location and identity of the object and directs the hands to it. Subsequently, the motor control system of the arms and hands implements the desired actions. Here, vision provides the information of where to direct the body, which state to monitor, and determine when the action must be terminated. Taken together, a ’script’ of instructions is sequentially implemented where the eye movements earmark the task-relevant locations in the environment that demand attentional resources for that action.

The line of argument in the above studies investigating natural tasks (e.g., tea-making, sandwich-making, hand-washing) share a common theme, i.e., the tasks have an organized and well-known structure. They involve specific object-related actions such as picking up the knife, picking up the teapot, etc. and have a predefined ’script’ for the execution of the tasks. The studies, therefore, study eye movements that are under strict control of a task sequence. Moreover, these tasks

are over-learned as they are part of a habitual action repertoire for an adult human being. As discussed by Land (2006) the low-level schemas (locating, directing, guiding, monitoring) defined above are likely not executed under deliberate conscious control. This distinction corresponds to James (2007) distinction between "ideo-motor" and "willed" acts. As James described, ideo-motor actions correspond to movements where we are "aware of nothing between the conception and execution" of the said action. In contrast, the willed actions require "an additional conscious element in the shape of a fiat, mandate, or expressed consent." Hence, it is unclear whether these low-level schemas of gaze control operate similarly for deliberate or planned actions where an internal task script is not already known.

Norman and Shallice (1986) proposed a theoretical framework for the components of attentional mechanisms that govern deliberate/planned actions. In comparison to the low-level schema proposed by Land (2006), which can account for routine, well-learned tasks, the Norman and Shallice (1986) model suggests another supervisory module that selects a schema to implement. In well-learned tasks, a schema is triggered automatically without conscious control. However, when a task is fairly complex and requires planning, multiple low-level schemas might compete for resources at the same time and require contention scheduling. For example, contentions can arise on whether to monitor the current action with respect to previous actions or future planned actions to fulfill the task-relevant goals. Such a scheduling mechanism is then required to provide conflict resolution for potentially relevant schemas either by inhibition or activation. Taken together, the model predicts that a failure of the supervisory control can lead to an instability of attention and heightened distraction.

To further pursue this line of research, it is desirable to perform such experiments under tightly controlled laboratory conditions. In recent years, virtual reality (VR) and mobile sensing have offered great opportunity to create controlled, natural environments. Here, subjects' eye and body movements can be measured reliably along with their interactions with the environment (Keshava et al., 2020, 2021; Clay et al., 2019; Mann et al., 2019). Experiments in virtual environments have grown popular in recent years and have shown promise towards studying cognition in naturalistic and controlled environments.

In the present study, we investigate the mechanisms of allocation of attention while performing a novel task in a naturalistic environment. We created two types of tasks that varied in complexity and required performing a sequence of actions to accomplish the cued goal. We asked subjects to sort objects on a life-size shelf based on the object features. The complexity of the task depended on sorting based on one object feature or both. We designed the tasks to be novel in a way that subjects had to plan their action sequences on the fly and in the absence of a pre-defined action script. We concurrently measured the eye and body movements while subjects performed the tasks. In the tasks, an abundance of look-ahead saccades would be evidence of elaborate cognitive planning. That is, high-level planning processes with matching eye movements would also support optimal decision-making. Further, the concurrence of cognitive processing and actions would emphasize a strict sequence of fixations with specific purposes, e.g., locating, directing, guiding, checking where cognitive schemas and actions would evolve in parallel. In this way we test the generalization of the theoretical frameworks of attentional mechanisms in novel tasks.

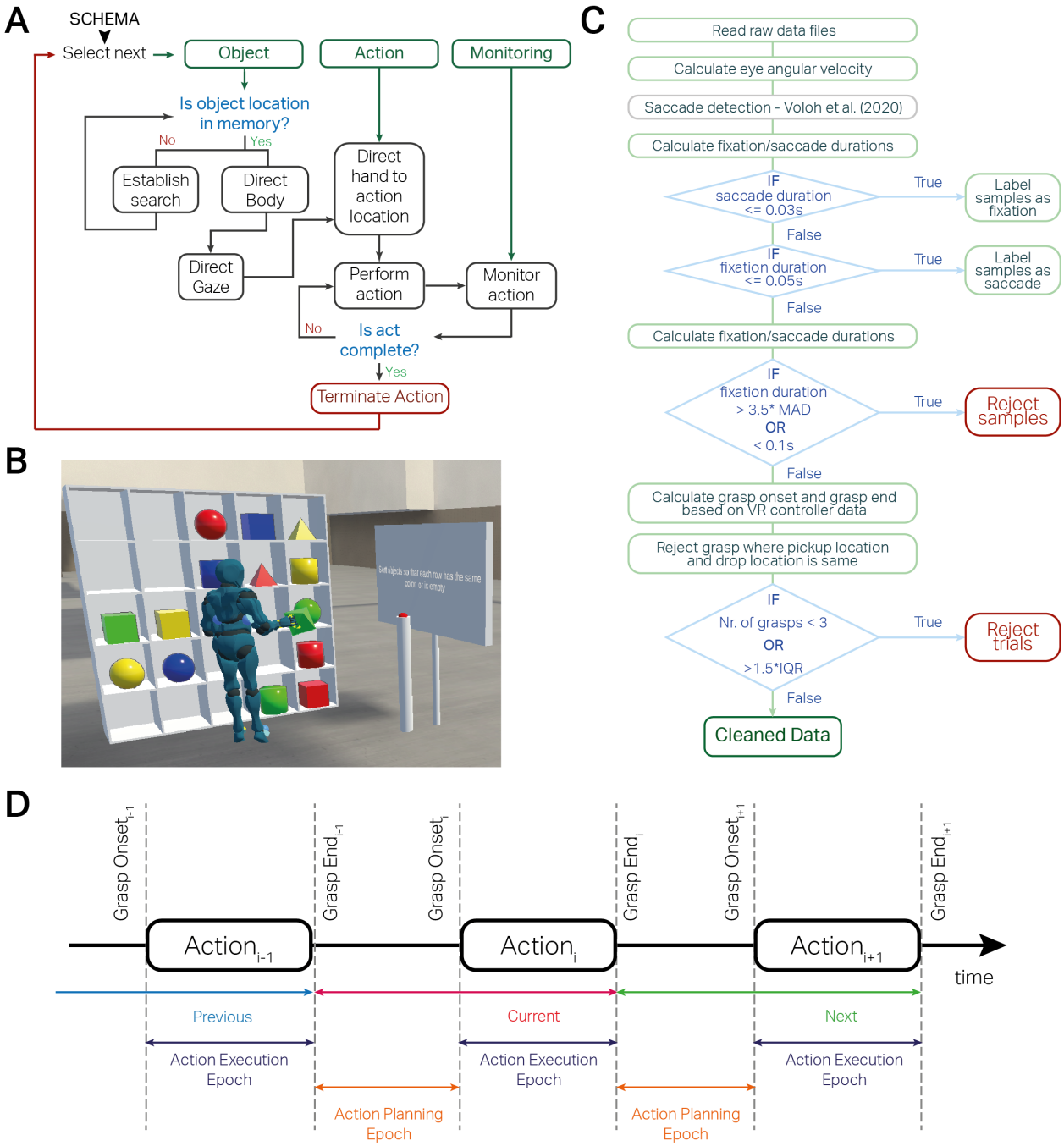

Figure 1: **A.** Schematic of motor and gaze control during performance of natural tasks, adapted from Land and Hayhoe (2001). **B.** Experimental Task. In a virtual environment participants sorted 16 different objects based on 2 features color or shape while we measured their eye and body movements. The objects were randomly presented on a 5x5 shelf at the beginning of each trial and were cued to sort objects by shape and/or color. Trials where objects were sorted based on just one object feature (color or shape) were categorized as EASY trials. Conversely, in the trials where sorting was based on both features (color and shape) were categorized as HARD trials. All participants performed 24 trials in total (16 easy trials and 8 hard trials) with no time limit. **C.** Data preprocessing steps for fixation/saccade detection and data rejection. **D.** Action execution and planning epochs. In order to study the function of eye movements we divided each trial into action planning and action execution epochs. The action execution epochs start from grasp onset till grasp end for each object displacement, whereas the action planning epochs start from grasp end of previous object displacement and grasp onset of current object displacement.

2 Results

Our experiment measured the eye and body movements as participants performed a sorting task in a virtual environment Figure 1B. The participants sorted objects based on the color and/or shape where we modulated the task complexity into EASY and HARD trials. We further divided the trials into planning and execution epochs where participants planned the selection of the target objects to grasp and then executed the action of displacing them to target shelves, respectively. In this section, we report the behavioral and oculomotor differences of the subjects for the two task types (EASY, HARD), and the planning and execution epochs.

2.1 Task-based Behavioral Differences

In the present study, the primary object-related action was to iteratively plan and execute the pickup and placement of objects until they were sorted according to the task instructions. To account for the behavioral differences between the task complexities, we used the measure of trial duration and the number of object displacements required to complete the tasks. Figure 2A shows the differences in EASY and HARD trials based on the time taken to finish the sorting task. A two-sample independent t-test showed that the trial duration for the two trial types was substantially and significantly different ($t = -10.13, p < 0.001$) where EASY trials that required the objects to be sorted based on a single feature were shorter (Mean = $54.12s$, SD = $\pm 13.33s$) as compared to HARD trials (Mean = $111.57 \pm 36.92s$) where subjects had to sort taking into account both features (color and shape) of the objects. Notably, the increase in task complexity increased the duration of the trials by a factor of 2. This indicates that the differences in task instructions of EASY and HARD tasks had the desired behavioral effects.

To determine if the participants sorted the objects during the trials optimally, i.e., if there was a broader cognitive plan, we compared their number of object displacements with an optimal depth-first search algorithm. Figure 2B shows the comparison of the object displacements made by the subjects and the optimal number of displacements as elicited by the model for both EASY and HARD trials. Subjects made lower number of object displacements in the EASY trials (Mean = 10.2 ± 1.99) compared to HARD trials (Mean = 15.52 ± 2.59). An independent t-test revealed that the participants made a significantly higher number of object displacements in the HARD trials ($t = 11.21, p < 0.001$). In the EASY trials the optimal model required a lower number of object displacements (Mean = 9.42 ± 1.48) displacements, whereas, in the HARD trials, the model required a slightly higher number of displacements (Mean = 11.24 ± 2.77). We compared the human and model performance for the two trial types using independent t-tests. In the EASY trials, there was a significant difference between the model and human object displacements ($t = 3.64, p < 0.001$). In the HARD trials, there was also a significant and larger difference between the model and human performance ($t = 10.61, p < 0.001$). This indicates that the participants did not plan their actions optimally and might have used other heuristics to complete the tasks.

To check if there were any salient spatial heuristics employed by the participants, we calculated the propensity to pick up and drop off objects to preferred locations on the shelf. In Figure 2C shows that subjects preferred to pick up objects

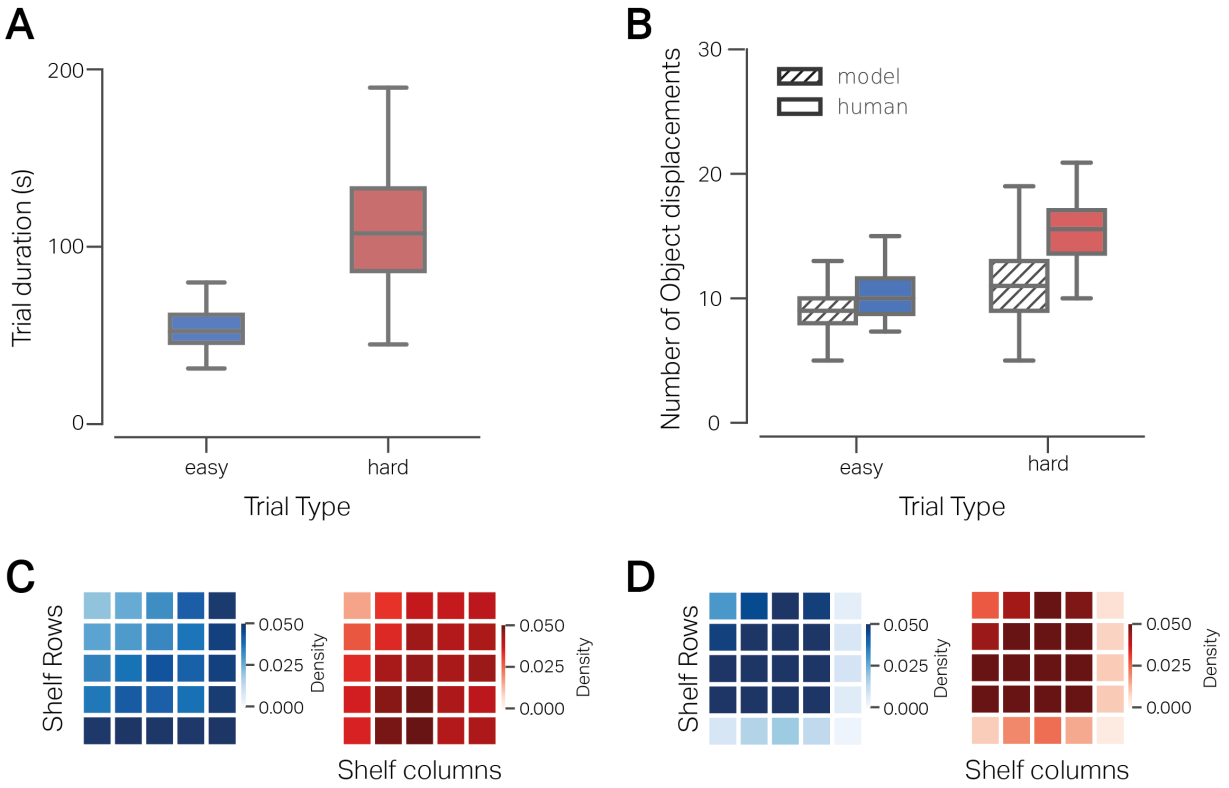

Figure 2: A. Trial duration of the EASY and HARD trials. The box plots show the interquartile range (IQR) of the duration of the trials for the two different trial types for all trials and participants. The whiskers represent 1.5 times the IQR. **B.** Distribution of the number of object displacements for EASY (blue) and HARD (red) trials. The colored box plots show the interquartile range of the number of object displacements made by subjects per trial and per participant. The whiskers represent 1.5 times the IQR. The dashed box plots show the optimal number of displacements required to sort the objects for a model computed with a depth-first search algorithm for 5000 random trial initialization for each trial type. **C.** Propensity of picking up objects from a preferred location on the 5x5 shelf locations with blue heatmaps showing the probability density over EASY trials and red heatmaps showing density over HARD trials. The probability density shows that subjects have a propensity to pick up objects from the rightmost column and bottom column for EASY trials (left) and conversely, in the HARD trials (right) subjects pick up objects from central locations. **D.** Propensity of dropping-off objects to a preferred location on the 5x5 shelf locations with blue the heatmap showing probability density over EASY trials and red heatmap showing density over HARD trials. The probability density shows that subjects display a systematic propensity to place the objects everywhere other than the bottom row or rightmost column.

from the right-most column and bottom-most row of the shelf. Figure 2D shows that subjects had a propensity to drop the objects leaving out the right column and bottom row of the shelf for both EASY and HARD trials. Given the sorting tasks, where subjects were presented with random initial configurations of the objects on the shelf locations, we did not expect any systematic spatial biases at play. Further, the expectation was that the subjects would move the objects randomly and not display a preference for object pickup and drop-off locations. These results show that subjects systematically, displace the objects leftward and upward employing an arbitrary and a non-task-specific heuristic to complete the tasks. As the objects are instantiated on the shelf randomly, an optimal strategy would not show this behavior. We can conclude from the above that subjects offset their cognitive load of optimally completing the task by employing simple spatial heuristics. In other words, in lieu of optimally performing the task and finishing it in a shorter time, subjects preferred to offload cognitive effort on the environment by adopting a more sub-optimal strategy.

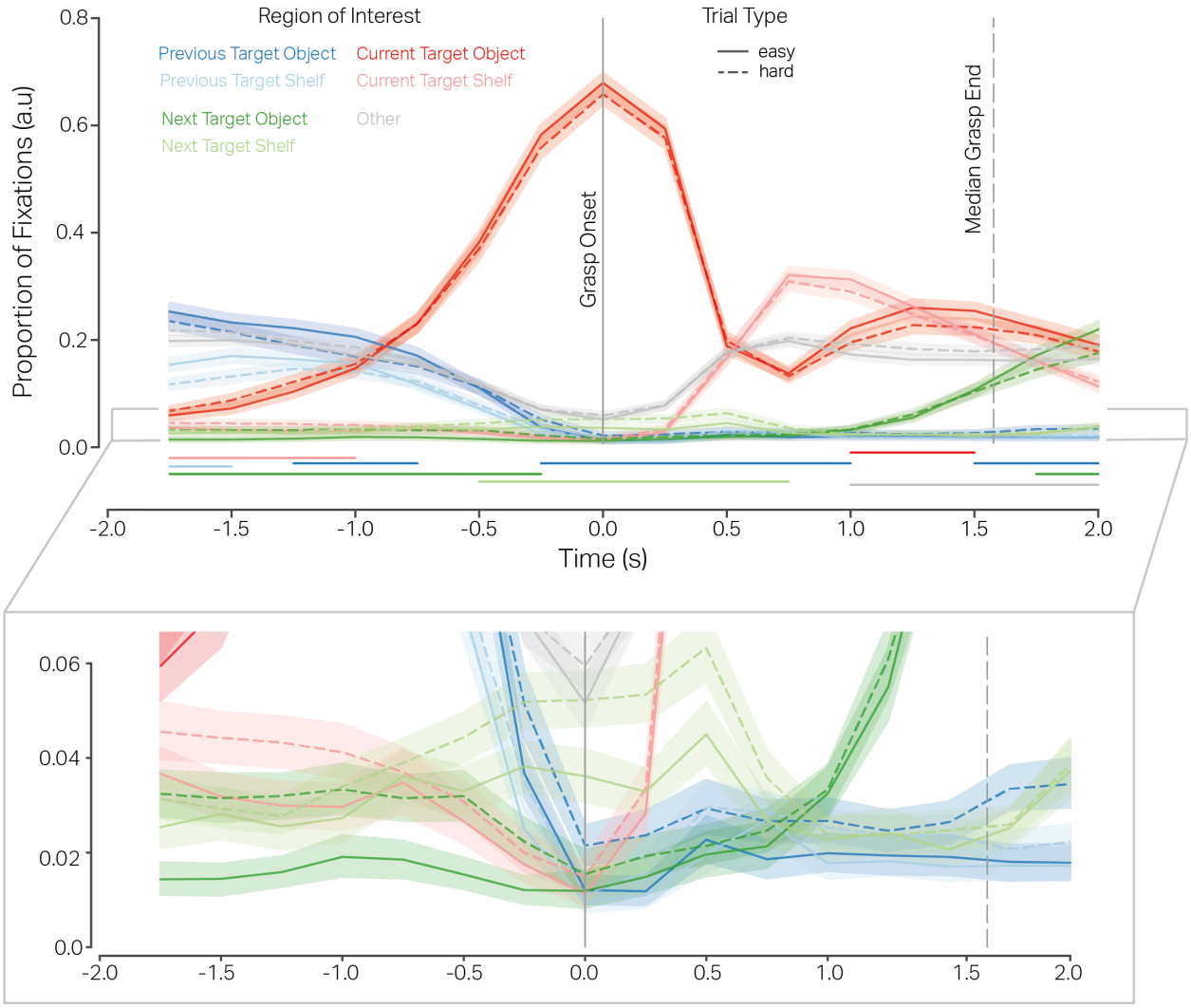

Figure 3: Proportion of fixations on the seven regions of interest, time-locked to grasp onset (time=0s) and 2 seconds before and after in the EASY (solid trace) and HARD (dashed trace) tasks. In each time bin of 0.25s, the proportion of fixations on all 7 ROIs for a trial-type add to 1. The dashed vertical line (time=1.57s) denotes the median end of the action execution. The shaded regions show the 95% confidence interval of the mean proportion of gaze at each time-bin across all subjects. The horizontal traces at the bottom correspond to the significant time periods where the proportion of fixations on a ROI was significantly different for the EASY and HARD trials.

2.2 Action Locked Gaze Control

In the next step, we investigated the average fixation profiles over the seven ROIs time-locked to the grasp onset, i.e., the time when the hand makes contact with the current target object. For the analysis, we chose the period from 2s before action onset to 2s after. We were also interested in the differences in the average fixation behavior as a consequence of task complexity. Figure 3 shows the time course of the proportion of fixations on the seven ROIs as described in section 4.5.2 for the EASY and HARD tasks.

In the action planning phase, the average fixation profile shows important aspects of oculomotor control 2s before the grasp onset. At the beginning of the planning phase, there is an equal proportion of fixations on the previous target object (dark blue trace) and 'other' regions (grey trace) on the shelf. Subsequently, as the fixations on the previous target

object and shelf (light. blue trace) decrease, there is a linear increase in the proportion of fixations on the current target object (dark red trace) and is maximum at the point of grasp onset (time= 0s). Throughout the action planning period, the proportion of fixations on the next target object (dark green trace) and shelf (light green trace) remain comparatively low. The large proportion of fixations in the action planning phase devoted to the current target object, shortly before the hand makes contact with the object is indicative of just-in-time fixations used to direct and guide the hand towards the target object.

In the action execution phase, up to 2s after the grasp onset, the proportion of fixations on the different ROIs reveal further action-relevant oculomotor behaviors. Firstly, the proportion of fixations decrease with time from the current target object and shift towards the current target shelf (light red trace) where the proportion of fixations increases and is maximum approximately 1 second after grasp onset and well before the median end of the action execution at 1.57s. Secondly, before the median end of the action execution, there is an increase in the proportion of fixations on the next target object. The presence of fixations on the next target object even before the end of the current action execution shows that the target for the next action is queued in shortly before ending the current action. Moreover, the proportion of fixations on the 'other' ROI although lower than on the current target shelf, indicates a search process for task-relevant objects/shelves. In sum, in the action execution phase, the results show two primary functions of gaze, first to assist in the completion of the current action and second to locate or search the objects for the next action.

Furthermore, the proportion of fixations for the EASY and HARD tasks are similar for the ROIs across time. Even though there are large behavioral differences in the EASY and HARD tasks, both in terms of time taken to complete the tasks and the number of actions required, the fixation profiles time-locked to the grasp onset are comparable. Most importantly, we see that the fixations are made in a just-in-time manner where the gaze shifts occur systematically in sync with the previous, current, and next action sequence one after the another. Hence, we can conclude that the oculomotor behavior described above is driven just-in-time by overt actions and less so by the cognitive load of the different sorting tasks.

To further quantify the differences in the proportion of fixations elicited due to the task complexity, we used a cluster permutation analysis (section 4.5.3) of the time course over the ROIs. The analysis revealed several time periods where the proportion of fixations was different for the 7 ROIs for the two task types.

The proportion of fixations on the previous target object differed between EASY and HARD trials for three time periods. The first time period spanned from -1.25s to -0.75s ($p < 0.001$) with a lower proportion of fixations in the HARD tasks compared to the EASY tasks indicating allocation of gaze to other ROIs relevant in the action sequence. The second significant time period was from -0.25s to 1s ($p < 0.001$) where the proportion of fixations was higher in the HARD tasks. This time period spans the time of action initiation till the execution of the object displacement strongly indicating that these fixations are related to the monitoring of the execution of the current action with respect to the previous task-relevant object and could be classified as 'checking' fixations. The third significant time period was from 1.5s to 2s ($p < 0.001$) with a higher proportion of fixations in the HARD trials than EASY trials. These differences

might be constitutive of further 'checking' fixations in the case of HARD trials while executing the object displacement. The differences in the proportion of fixations on the previous target object, especially in the time period concurrent with the action execution is strongly indicative of checking fixations that monitor the task progression.

The proportion of fixations on the previous target shelf was lower in the HARD trials compared to the EASY trials from -1.75s to -1.5s ($p < 0.001$) before grasp onset. As this difference, although significant, occurs in a smaller time period it should be interpreted cautiously and could suggest the allocation of gaze to ROIs relevant to the current task sequence happens earlier in the HARD tasks as more planning is required.

The proportion of fixations on the current target object differed between tasks for a single time period. There were differences that occurred from 1s to 1.5s ($p < 0.001$) after grasp onset with a lower proportion of fixations in the HARD tasks compared to the EASY trials. These differences indicate that towards the end of action execution in HARD tasks fixations are allocated towards other task-relevant ROIs in the action sequence. More importantly, there are no differences in the proportion of fixations on the current target object before the grasp onset, suggesting that task complexity does not play a role in 'directing' fixations that direct the hand to the action-relevant object.

The proportion of fixations on the current target shelf also differed between the task types for a single time period. There was a higher proportion of fixations on the current target shelf in the HARD trials compared to EASY trials from -1.75s to -1s ($p < 0.001$) before grasp onset. From these differences, we can infer that some proportion of fixations are utilized to plan the current task, well before the action has been initiated. These fixations could be classified as 'locating' fixations or look-ahead fixations which are predominantly present due to the complexity of the HARD tasks and due to the need for elaborate planning required in these tasks.

For the next target object, the proportion of fixations were different in two time periods. There were a higher proportion of fixations on the next target object from -1.75s to -0.25s ($p < 0.001$) and a lower proportion from 1.75 s to 2s ($p < 0.001$) in the HARD tasks compared to the EASY tasks. The differences in the first time period suggest that these fixations also serve as 'locating' fixations to queue the next action-relevant object in the sequence indicating some planning of the next actions in the HARD tasks. The differences in the latter time period show a lower proportion of fixations for the HARD tasks indicating that prior locating fixations made it unnecessary to allocate attention to that object of interest. Hence, in the EASY trials, the lower proportion of these locating fixations are further evidence of an ad-hoc or just-in-time gaze allocation for the next action-relevant item.

There was also a higher proportion of fixations on the next target shelf from -0.5s to 0.75s ($p < 0.001$) in the HARD tasks as compared to EASY tasks. These fixations are made in concert with the onset of the action execution and indicate that these fixations play the role of both 'checking' fixations to monitor the task progression as well as 'locating' fixations to queue the locations in the scene that are pertinent for the next action in the sequence.

Finally, the proportion of fixations on the other objects and shelves in the scene were higher in the HARD tasks from 1s to 2s ($p < 0.001$) after grasp onset. These differences indicate a search behavior towards the end of the current task

execution. Given the task complexity of the HARD trials, this search behavior might function to queue further objects or shelves of interest in the subsequent actions in the sequence.

To summarize, the results above support two broad findings. First, despite large behavioral differences between the two task types, the eye movements do not show corresponding large deviations. This supports the idea that these eye movements are tightly coupled to the actions and less so with the cognitive load added by task complexity. Secondly, the proportion of fixations on the ROIs for the two task types differ at specific time periods and relate to relevant cognitive schemas such as searching for and queuing objects into the working memory for subsequent actions and monitoring the progression of the ongoing action. Consequently, even though eye movements reveal the different cognitive schemas during action planning and execution, these schemas do not correspond to broader planning strategies for task completion.

2.3 Sequential Gaze Control in Action Planning and Execution

Next, we focused our analysis from fixations on single regions-of-interest to sequences of attended regions. That is, we study the scanning behavior in individual trials while subjects plan and execute an action in order to further elucidate the concurrent cognitive processes. Specifically, we computed transition matrices to capture fixations to and from each of the seven ROIs as described in section 4.5.3. Figure 4A shows the exemplar transition matrix for one planning epoch in an EASY trial. The total transitions represent the total number of fixations to and from one ROI to another. The net transitions, on the other hand, denote the asymmetry in the allocation of fixation from one ROI to another. For example, an equal number of fixations to and from pairs of ROIs would result in zero net transitions between those ROIs. Similarly, greater transitions from one ROI to another than vice versa would result in greater net transitions from that ROI to the other. Furthermore, we analyzed the relative number of net transitions within the planning and execution epochs of a trial. With maximum relative net transitions equal to 1, we expect an equal number of net and total transitions between ROIs, indicating all saccades between ROIs are made parsimoniously and show purposeful gaze guidance to task-relevant ROIs. Conversely, the minimum relative net transition of 0 would indicate an equal number of saccades from and to ROI pairs, either due to searching or due to monitoring the task progression. Hence, when subjects use saccades purposefully such as for guiding their hand or body towards the current target object we would expect high relative net transitions and with a pattern of fixations related to multiple task-relevant schemas we would expect lower relative net transitions.

As evidenced by the fixation transition behavior, the relative net transitions were different for the trial types and the action planning and execution epochs Figure 4B. In the EASY trials, for the PLANNING epochs the mean relative net transition was $0.66 SD = \pm 0.10$ and 0.63 ± 0.14 for the execution epochs. In the HARD trials, for the PLANNING epochs, the mean relative net transition was 0.61 ± 0.11, whereas it was 0.62 ± 0.14 in the execution epochs. There was also a positive correlation of the mean relative net transitions in a trial and the number of object displacements made in the trial Figure 4C. Specifically, in the EASY trials, for PLANNING epochs the relative net transition correlated significantly with the number of object displacements $(\rho = 0.19, p < 0.001)$ as well as for the EXECUTION epoch

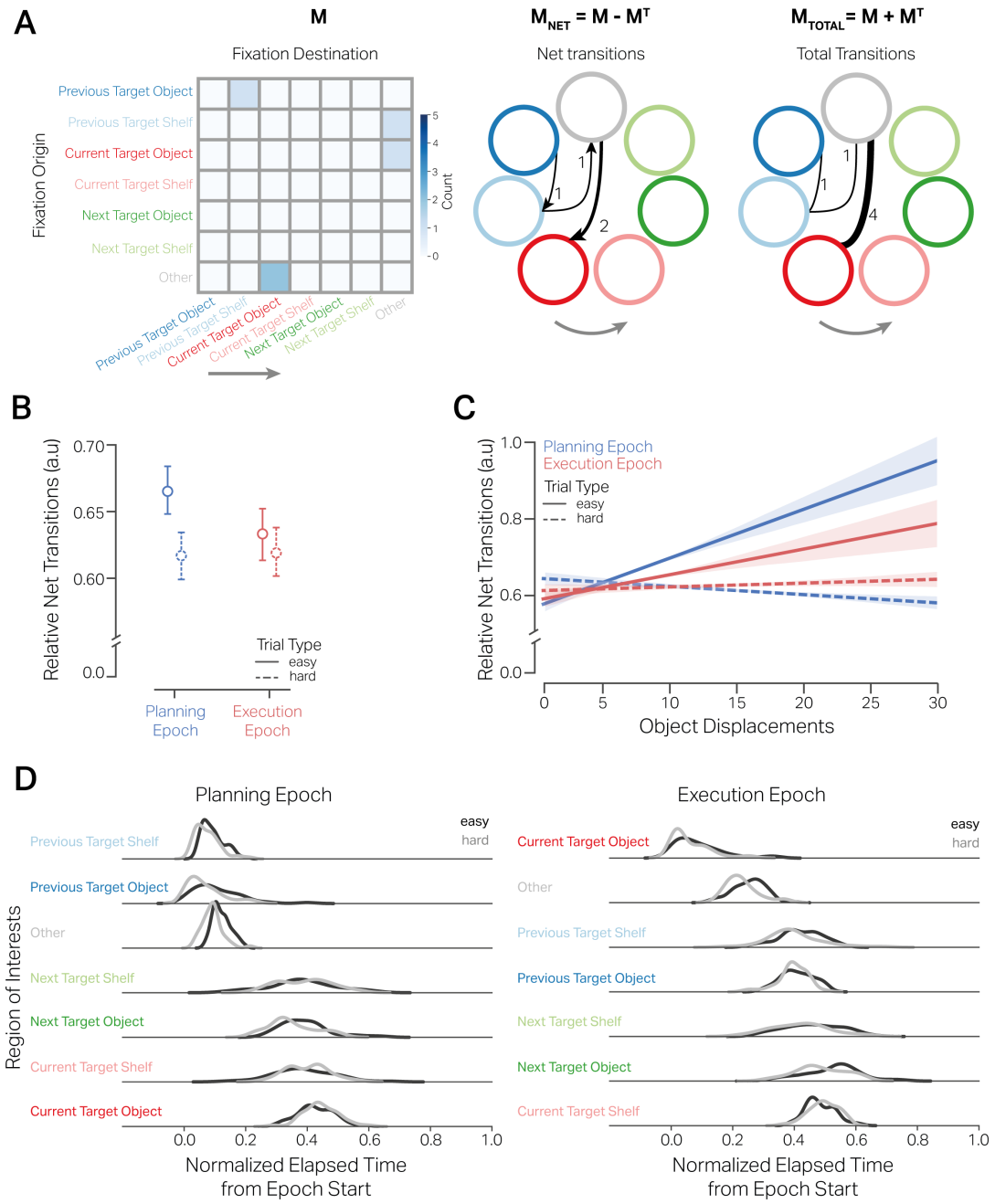

Figure 4: **A.** Exemplar transition matrix denoted as **M** for gaze switching in a planning epoch. The net (M_{net}) and total transitions (M_{total}) are calculated from matrix M. **B.** The differences in relative net transitions (ordinate) for action planning and execution (abscissa) in the two trial types (EASY and HARD). The point estimate denotes the mean relative net transition across subjects and the error bars indicate the 95% confidence interval. **C.** Regression fit over the fixed-effects of trial types (EASY, HARD), epoch types (PLANNING, EXECUTION) and object displacements in the trial on the relative net transitions. The traces denote the regression fit and the shaded region denotes 95% confidence interval. **D.** Distributions of median time to first fixation on the 7 ROIs for the action planning and execution epochs and trial types. The ordinate is sorted in ascending order of latency from epoch start for both planning and execution epochs.

($\rho = 0.09, p = 0.02$). In contrast, for PLANNING epochs in the HARD trials, the relative net transition did not correlate with the number of object displacements ($\rho = -0.07, p = 0.14$) as well as for the EXECUTION epoch ($\rho = 0.03, p = 0.51$).

In order to statistically analyze the differences in mean relative net transitions in a trial we used a linear mixed effects model (section 4.5.4) with relative net transitions in a trial as the dependent variable and trial type (EASY, HARD), epoch type (PLANNING, EXECUTION) and the number of object displacements in the trials. As the independent variables were effect coded, the regression coefficients can be directly interpreted as main effects. Figure 4B illustrates the effects of trial type, epoch type, and number of object displacements on the relative net transitions. There was a significant main effect of factor trial type (HARD - EASY) $\beta = -0.04$ $(95\%CI = [-0.06, -0.03],\ t(87.04) = -5.37,\ p < 0.001,\ \eta^2 = 0.24)$ showing that HARD trials had lower relative net transitions than EASY trials. There was no effect of epoch type (PLANNING - EXECUTION) $\beta = 0.02$ $(95\%CI = [0.00, 0.05],\ t(48.03) = 1.87,\ p = 0.066,\ \eta^2 = 0.05)$ showing that PLANNING epochs did not differ from EXECUTION epochs in terms of relative net transitions. There was a significant main effect of number of object displacements in a trial $\beta = 0.003$ $(95\%CI = [0.001, 0.004],\ t(1448.78) = 3.13,\ p = 0.002)$ showing that a one unit increase in the number of object displacements in a trial led to increase in relative net transitions by a factor of 0.003. There was a significant interaction between trial type and epoch type $\beta = -0.04$ $(95\%CI = [-0.07, -0.01],\ t(1835.94) = -2.66,\ p = 0.008)$ showing that PLANNING epochs in HARD trials had lower relative net transitions. There was a significant interaction between trial type and number of object displacements $\beta = -0.008$ $(95\%CI = [-0.01, -0.004],\ t(770.47) = -4.71),\ p < 0.001)$ showing that one unit increase in the number of object displacements in HARD trials led to decrease in relative net transitions by a factor of -0.009. There was no significant interaction between epoch type and number of object displacements $\beta = 0.001$ $(95\%CI = [-0.002, 0.004],\ t(1573.90) = 0.82,\ p = 0.4)$. Further, there was a significant interaction between trial type, epoch type and number of object displacements $\beta = -0.01$ $(95\%CI = [-0.02, -0.002],\ t(1814.87) = -2.74,\ p = 0.03)$ showing that a one unit increase in the number of object displacements in HARD trials and for PLANNING epochs led to decrease in relative net transitions by a factor of -0.01.

The analysis above lends evidence to differential gaze guidance behavior at the level of action planning and execution with respect to task complexity. The higher relative net transitions in the EASY trials suggest saccades were primarily made towards the current task-relevant objects for directing and guiding the body or hand towards the object of interest. For EASY tasks, we see lower net transitions in the action execution epochs compared to PLANNING epochs, indicating eye movements were engaged in monitoring the task progression, resulting in back and forth saccades between the previous, current, and next target ROIs in the action sequence. The lower relative net transitions in the HARD tasks suggest gaze switching to and from ROIs in search for task-relevant ROIs, especially in action planning epochs and further gaze switches supporting task monitoring in the execution epochs. The overall significant positive correlation between object displacements and the relative net transitions reveal gaze allocation being predominantly made in a just-in-time manner with little overarching planning which also corresponds well with the sub-optimal behavior exhibited by the subjects. In sum, the relative net transitions showed gaze guidance behavior in line with the appropriate selection of action schemas of searching for target object, directing the hand and body to target object, monitoring task progression.

2.3.1 Latency of Fixations on Task-relevant ROIs

As a final measure, to further study the temporal aspects of gaze guidance behavior during action planning and action execution, we were interested in the latency of the first fixations to the 7 ROIs. We used the median time to first fixate on the 7 ROIs per trial as a measure of latency. We hypothesized if the fixations are made to the task-relevant regions in accordance with a set cognitive schema, we would see a strict temporal window where these fixations occur in the planning and execution epochs. On the other hand, if fixations are made haphazardly, the occurrence of fixations would not follow a temporal structure and hence would not belong to specific action schemas. As shown in Figure 4D the distributions of the first fixations on the seven ROIs for the planning and execution epochs show a structured sequence of occurrence of first fixations on the ROIs.

Here, we describe the latency of the first fixations on the 7 ROIs normalized by the epoch duration from the start of the epoch. In the action planning epochs, the first fixation on previous target shelf occurred at 0.09 ($95\%CI = [0.08, 0.10]$) after epoch start in the EASY trials and at 0.07 ($95\%CI = [0.06, 0.08]$) in the HARD trials. Similarly, the first fixation on the previous target object occurred at 0.10 ($95\%CI = [0.08, 0.12]$) in the EASY trials and at 0.06 ($95\%CI = [0.05, 0.08]$) in the HARD trials after the epoch start. The fixations then occurred at 0.12 ($95\%CI = [0.11, 0.13]$) over the 'other' ROI in EASY trials and at 0.09 ($95\%CI = [0.08, 0.10]$) in the HARD trials. Subsequently, fixations occurred on the next target shelf at 0.38 ($95\%CI = [0.0.35, 0.41]$) in the EASY trials and at 0.39 ($95\%CI = [0.36, 0.41]$) in the HARD trials. Similarly, there were first fixations on the next target object at 0.39 ($95\%CI = [0.36, 0.41]$) in the EASY trials and at 0.35 ($95\%CI = [0.32, 0.37]$) in the HARD trials. There were also fixations recorded on the current target shelf at 0.38 ($95\%CI = [0.36, 0.41]$) in EASY trials and at 0.40 ($95\%CI = [0.38, 0.42]$) in the HARD trials. Finally, the first fixations on the current target object were at 0.42 ($95\%CI = [0.41, 0.44]$) in EASY trials and at 0.44 ($95\%CI = [0.43, 0.46]$) in the HARD trials. Table 1 shows the model coefficients of the linear regression model that compared the latency of the first fixation on current target object with the rest of the ROIs and their interactions with respect to the task complexity as described in section 4.5.4. The fixed effects show that the latency of the first fixations on the ROIs were significantly different from the latency on the current target object. Furthermore, the interactions show that task complexity significantly increased the difference in latency of first fixations on previous target object and shelf, other and next target object compared to the current target object. Taken together, irrespective of task complexity, the action planning epochs show a systematic progression of fixations from one ROI to another. This structured temporal sequence of fixations with a defined temporal window shows that the fixations pertaining to the current and future action relevant ROIs are not incidental and part of a cognitive schema to accomplish the task. Moreover, given the task complexity, the temporal profiles of these look-ahead fixations can change and occur slightly earlier. Most importantly, the first fixations on the current target object occur last in sequence after more than 40% of the epoch time has elapsed lending strong evidence to the just-in-time manner in which gaze is allocated to the immediate task requirements.

We similarly analyzed the latency of first fixations on the 7 ROIs during the action execution epochs. Figure 4D summarized the distribution of the latency of first fixations on the different ROIs during the action execution epoch. The first fixation on current target object occurred at 0.09 ($95\%CI = [0.06, 0.11]$) in the EASY trials and at 0.06 ($95\%CI =$

$[0.04, 0.08]$)in the HARD trials. Similarly, the first fixation on the other ROI occurred at 0.25 ($95\%CI = [0.24, 0.27]$) in the EASY trials and at 0.23 ($95\%CI = [0.21, 0.25]$) in the HARD trials after the epoch start. The fixations then occurred at 0.41 ($95\%CI = [0.39, 0.43]$) on the previous target object in EASY trials and at 0.40 ($95\%CI = [0.38, 0.41]$) in the HARD trials. Subsequently, fixations occurred on the previous target shelf at 0.42 ($95\%CI = [0.40, 0.44]$) in the EASY trials and at 0.40 ($95\%CI = [0.37, 0.42]$) in the HARD trials. Similarly, the first fixations on the next target shelf occurred at 0.46 ($95\%CI = [0.44, 0.49]$) in the EASY trials and at 0.42 ($95\%CI = [0.39, 0.45]$) in the HARD trials. Fixations then occurred on the current target shelf at 0.48 ($95\%CI = [0.47, 0.50]$) in EASY trials and at 0.49 ($95\%CI = [0.48, 0.51]$) in the HARD trials. Finally, the first fixations on the next target object were at 0.48 ($95\%CI = [0.46, 0.51]$) in EASY trials and at 0.52 ($95\%CI = [0.50, 0.55]$) in the HARD trials. Table 1 shows the model coefficients of the linear regression model that compared the difference of latency of the first fixation on the current target object with the rest of the ROIs and their interactions with respect to the task complexity in the action execution epoch as described in section 4.5.4. The fixed effects reveal that the difference in the latency of first fixations on all the ROIs was significantly different from the occurrence of the first fixation on the current target object. Furthermore, there is a small interaction of trial type with the difference in latency between the current target shelf and the current target object. In sum, irrespective of the task complexity, there is a systematic sequence of first fixations on the seven ROIs during action execution. The early fixations on the previous target object and shelf indicate these fixations are related to monitoring/checking the current action execution with respect to the previously handled object. Similarly, later fixations on the next target object and shelf indicate fixations used for planning the next action even before the action has been completed. Importantly, the increase in task complexity did not significantly affect the temporal sequence of the first fixations on the task-relevant ROIs during action execution. Here again, the first fixation on the current target shelf, i.e. where the target object will be placed is fixated in the end, and strongly indicates a just-in-time manner of gaze allocation.

In sum, the latency of first fixations on the ROIs revealed the activation of specific cognitive schemas and their relative time of occurrence in the action planning and execution epochs. In both epochs, we found a temporal structure that defined the sequence of occurrence of the first fixations on the task-relevant ROIs. The relative timing of fixations that correspond to the targets in the next action shows that look-ahead fixations occur both during the planning and execution of the immediate action and are less likely to be incidental. Moreover, the relative timing of fixations on the current task-relevant target shows again a just-in-time manner of gaze allocation.

3 Discussion and Conclusion

In the present study, we investigated the spatio-temporal aspects of gaze control while action planning and action execution with varying task complexity. We report five main findings of this study. First, we found subjects used ad-hoc spatial heuristics to reduce the solution search space resulting in an increased number of object displacements as compared to an optimal model. Second, average fixations aligned to the action initiation revealed gaze allocation behavior that differed at specific time periods due to task complexity. However, given the behavioral differences induced

Table 1: Fixed effects coefficients of the linear mixed model detailing estimated latency of first fixations w.r.t the current target object during the action planning and execution

	Estimate	95% CI	t	df_error	p
Action Planning Epochs					
Latency w.r.t Current Target Object					
Previous Target Shelf	-0.35	[-0.36, -0.33]	-37.82	65.95	**< 0.001*** **
Previous Target Object	-0.34	[-0.37, -0.32]	-29.00	49.66	**< 0.001*** **
Other	-0.33	[-0.35, -0.31]	-37.03	72.37	**< 0.001*** **
Next Target Object	-0.06	[-0.08, -0.05]	-6.95	47.36	**< 0.001*** **
Next Target Shelf	-0.04	[-0.07, -0.02]	-3.80	49.24	**< 0.001*** **
Current Target Shelf	-0.04	[-0.06, 0.02]	-3.39	50.16	**0.001*** **
Interactions					
Trial Type : Previous Target Shelf	-0.03	[-0.06, -0.01]	-2.30	403.67	**0.022** **
Trial Type : Previous Target Object	-0.05	[-0.09, -0.02]	-3.10	123.78	**0.002** **
Trial Type : Other	-0.04	[-0.07, -0.02]	-3.23	486.19	**0.001** **
Trial Type : Next Target Object	-0.05	[-0.09, -0.02]	-3.16	83.81	**0.002** **
Trial Type : Next Target Shelf	-0.009	[-0.05, -0.04]	-0.39	52.00	0.69
Trial Type : Current Target Shelf	-0.002	[-0.04, 0.04]	-0.10	55.38	0.91
Action Execution Epochs					
Latency w.r.t Current Target Object					
Other	0.16	[0.13, 0.19]	11.40	50.47	**< 0.001 *** **
Previous Target Object	0.32	[0.30, 0.35]	24.56	61.73	**< 0.001 *** **
Previous Target Shelf	0.33	[0.30, 0.35]	23.59	52.69	**< 0.001 *** **
Next Target Shelf	0.36	[0.33, 0.39]	24.55	58.75	**< 0.001 *** **
Current Target Shelf	0.41	[0.39, 0.44]	32.12	64.73	**< 0.001 *** **
Next Target Object	0.42	[0.39, 0.45]	28.48	54.68	**< 0.001 *** **
Interactions					
Trial Type : Other	0.009	[-0.03, 0.05]	0.39	136.41	0.69
Trial Type : Previous Target Object	0.02	[-0.02, 0.06]	0.95	116.24	0.34
Trial Type : Previous Target Shelf	0.01	[-0.03, 0.05]	0.48	98.32	0.62
Trial Type : Next Target Shelf	-0.01	[-0.06, 0.04]	-0.34	63.52	0.73
Trial Type : Current Target Shelf	0.04	[0.00, 0.09]	2.15	122.09	**0.03 ***
Trial Type : Next Target Object	-0.003	[-0.05, 0.05]	-0.09	60.66	0.92

by task complexity, the average action aligned fixations differed only slightly. Third, based on the scan path transitions, we observed greater relative net transitions to and from ROIs in the EASY trials compared to HARD trials indicating a just-in-time manner of gaze allocation to task-relevant ROIs. The lower proportion of net transitions to and from the ROIs in HARD trials is further evidence of search behavior required to find the target relevant for the immediate action. Moreover, the lower net transitions during action execution suggest gaze switching in line with checking fixations that monitor task progression. Fourth, the relative timing of the first fixations on the task-relevant ROIs in the action planning and execution phase revealed a systematic sequence of fixations leading up to the immediate task-relevant ROI, strongly suggesting a just-in-time strategy of action supporting gaze allocation. Finally, task complexity affected the timing of first fixations in the action planning phase but not in the action execution phase. In sum, our findings reveal a spatio-temporal structure of eye movements corresponding to low-level cognitive schemas that support the production and execution actions that are not necessarily optimal.

The central aim of our study is to generalize the cognitive mechanisms of gaze control for tasks that are not over-learned and routine. Building on the work of Land et al. (1999) where eye movements were studied while preparing tea, the tasks in our study are novel in the sense that they do not have an inherent action sequence attached to them. Our experimental setup provided a way to capture oculomotor behavior for tasks with varying complexity that did not have a strict action sequence. By studying the dispersion of fixations on the previous, current, and next task-relevant ROIs, we show a structured sequence of fixations that may be classified based on previous work. Moreover, we add to the current body of research by providing a data-driven approach to defining look-ahead fixations, directing fixations, guiding fixations, checking fixations, etc. Our results generalize the occurrence of these fixations to the action sequence and the task complexity by doing away with object identity.

Land et al. (1999) proposed the various functions of eye movements from locating, directing, guiding, checking during tea-making. In our study, we also showed the occurrence of these fixations time-locked to the action initiation. Importantly, our study shows that task complexity affects the proportion and timing of fixations for locating/look-ahead as well as for checking the task progression. While look-ahead fixations occurred predominantly before the initiation of the immediate action, the checking fixations occurred in parallel with the action execution. Interestingly, directing and guiding fixations were not affected by task complexity indicating that these fixations are central to the action repertoire.

Sullivan et al. (2021) recently showed the various timescales at which look-ahead fixations can occur while assembling a tent. They reported look-ahead fixations attributed to the current sub-task and within 10 seconds of motor manipulation. In our study, we found look-ahead fixations on the drop-off location of the current object displacement as well as towards the target object of the next displacement action. Our results show that these fixations are more salient in the more complex tasks. Further, we also show the systematic latency of these fixations with respect to the immediate action onset. Hence, the look-ahead fixations described in our study are very likely part of a cognitive schema that accounts for future actions and less so due to incidental causes.

Land and Furneaux (1997) have further elaborated on the schemas that direct both eye movements and motor actions. They proposed a chaining mechanism by which one action leads to another where the eyes supply constant information to the motor system via a buffer. In our analysis, the temporal structure of the first fixations on the ROIs lends further evidence of such a cognitive buffer that facilitates moving from one action to another where eye movements queue in the objects and locations necessary for current and future actions, while also checking with previous action outcomes. Importantly, the occurrence of fixations on the object or location necessary for the immediate action last in the sequence of fixations supports the prevalence of the 'just-in-time' nature of these action-related fixations. Taken together, our study further corroborates the cognitive schemas that sequentially support action planning and execution.

In our sorting tasks, we see a higher number of object displacements with increasing task complexity. To assess the sorting behavior of the participants we compared their object displacement behavior with a greedy depth-first search algorithm which optimizes for the shortest path to the solution. Studies in human performance in reasoning tasks as well as combinatorial optimization problems (MacGregor and Chu, 2011) have revealed that humans solve these tasks

in a self-paced manner rather than being dependent on the size of the problem. Pizlo and Li (2005) found that subjects do not perform an implicit search of the problem state space where they plan the moves without executing them, and where longer task duration would lead to shorter solution paths. Instead, they showed that humans break the problem down into component sub-tasks which gives rise to a low-complexity relationship between the problem size and time to solution. Further, Pizlo and Li (2005) showed that instead of using such an implicit search, subjects use simpler heuristics to decide the next move. To this effect, while subjects in our study are sub-optimal compared to a depth-first search algorithm and use arbitrary spatial heuristics during the task, humans, in general, are prone to use non-complex heuristics that favor limited allocation of resources in the working memory. Hence, the higher trial duration and number of object displacements shown by the participants do not necessarily demonstrate a lack of planning.

From the perspective of embodied cognition (Wilson, 2002; Ballard et al., 2013; Van der Stigchel, 2020), humans off-load cognitive work on the environment and the environment is central to the cognitive system. The behavioral results show that subjects use spatial heuristics to complete the tasks indicating they exploit the external world to reduce the cognitive effort of selecting optimal actions. Further, Droll and Hayhoe (2007) have suggested that a just-in-time strategy is likely used to lower the cognitive cost of encoding objects in the world into the visual working memory. The prevalence of the just-in-time strategy shown by subjects in our study is again evidence of them using the environment as an external memory. In the embodied cognition framework, cognition is situated and for producing actions. With our study, we show that task complexity affected various spatio-temporal features of gaze control in the action planning and execution phase. These findings are also in line with König et al. (2016) illustrating that eye movements reveal much about the cognitive state.

In conclusion, in the present study, we investigated the oculomotor responses to novel task scenarios that varied in complexity. Our results generalize past work on action-oriented eye movements to tasks that are not routine or over-learned such as tea-making, sandwich-making, hand-washing, etc. Our generalized approach eliminates the analysis based on object identity and shows the different gaze control features related to low-level cognitive schemas such as locating task-relevant objects of interest, directing and guiding the body and hands, as well as monitoring the task progression. Furthermore, we show the prevalence of cognitive schemas does not necessarily support optimal actions but rather only the production and execution of actions. More importantly, we show how fixations are tightly coupled with the action sequences pertaining to the tasks and engaged in a just-in-time manner.

4 Methods

4.1 Participants

A total of 55 participants (39 females, mean age = 23.9 ± 4.6 years) were recruited from the University of Osnabrück and the University of Applied Sciences Osnabrück. Participants had normal or corrected-to-normal vision and no history of neurological or psychological impairments. They either received a monetary reward of €7.50 or one participation credit per hour. Before each experimental session, subjects gave their informed consent in writing. They also filled

out a questionnaire regarding their medical history to ascertain they did not suffer from any disorder/impairments which could affect them in the virtual environment. Once we obtained their informed consent, we briefed them on the experimental setup and task. The Ethics Committee of the University of Osnabrück approved the study.

4.2 Apparatus & Procedure

For the experiment, we used an HTC Vive Pro Eye head-mounted display (HMD)(110° field of view, 90Hz, resolution 1080 x 1200 px per eye) with a built-in Tobii eye-tracker[1]. Participants used an HTC Vive controller to manipulate the objects during the experiment with their right hand. The HTC Vive Lighthouse tracking system provided positional and rotational tracking and was calibrated for a 4m x 4m space. For calibration of the gaze parameters, we used the 5-point calibration function provided by the manufacturer. To make sure the calibration error was less than 1°, we performed a 5-point validation after each calibration. Due to the study design, which allowed a lot of natural body movements, the eye tracker was calibrated repeatedly during the experiment after every 3 trials. Furthermore, subjects were fitted with HTC Vive trackers on both ankles, both elbows and, one on the midriff. The body trackers were also calibrated subsequently to give a reliable pose estimation using inverse kinematics of the subject in the virtual environment. We designed the experiment using the Unity3D[2] version 2019.4.4f1 and SteamVR game engine and controlled the eye-tracking data recording using HTC VIVE Eye-Tracking SDK SRanipal[3] (v1.1.0.1).

The experimental setup consisted of 16 different objects placed on a shelf of a 5x5 grid. The objects were differentiated based on two features: color and shape. We used four high-contrast colors (red, blue, green, and yellow) and four 3D shapes (cube, sphere, pyramid, and cylinder). The objects had an average height of 20cm and a width of 20cm. The shelf was designed with a height and width of 2m with 5 rows and columns of equal height, width, and depth. Participants were presented with a display board on the right side of the shelf where the trial instructions were displayed. Subjects were also presented with a red buzzer that they could use to end the trial once they finished the task.

4.3 Experimental Task

Subjects performed two practice trials where they familiarized themselves with handling the VR controller and the experimental setup. In these practice trials, they were free to explore the virtual environment and displace the objects. After the practice trials, subjects were asked to sort objects based on the one and/or two features of the object. There were two types of trials: EASY and HARD. Subjects were not limited by time to complete the task. The participants were not explicitly told about the task category and were told to follow the instructions given to them at the beginning. Each subject performed 24 trials in total with each trial type (as listed below) randomly presented twice throughout the experiment. The experimental setup is illustrated in Figure 1B. The EASY trials instructions were as follows:

1. Sort objects so that each row has the same shape or is empty

[1] https://enterprise.vive.com/us/product/vive-pro-eye-office/
[2] Unity, www.unity.com
[3] SRanipal, developer.vive.com/resources/vive-sense/sdk/vive-eye-tracking-sdk-sranipal/

2. Sort objects so that each row has all unique shapes or is empty
3. Sort objects so that each row has the same color or is empty
4. Sort objects so that each row has all unique colors or is empty
5. Sort objects so that each column has the same shape or is empty
6. Sort objects so that each column has all unique shapes or is empty
7. Sort objects so that each column has the same color or is empty
8. Sort objects so that each column has all unique colors or is empty

The HARD trials instructions were as follows:

1. Sort objects so that each row has all the unique colors and all the unique shapes once
2. Sort objects so that each column has all the unique colors and all the unique shapes once
3. Sort objects so that each row and column has each of the four colors once.
4. Sort objects so that each row and column has each of the four shapes once.

4.4 Data pre-processing

4.4.1 Gaze Data

The data preprocessing steps are illustrated in Figure 1C. As a first step, using the eye-in-head 3d gaze direction vector for the cyclopean eye we calculated the gaze angles for the horizontal θ_{h} and vertical θ_{v} directions. All of the gaze data was sorted by the timestamps of the collected gaze samples. The 3d gaze direction vector of each sample is represented in (x, y, z) coordinates as a unit vector that defines the direction of the gaze in VR world space coordinates. In our setup, the x coordinate corresponds to the left-right direction, y in the up-down direction, z in the forward-backward direction. The formulas used for computing the gaze angles are as follows:

$$\theta_{\mathrm{h}} = \frac{180}{\pi} * \arctan \frac{x}{z}$$
$$\theta_{\mathrm{v}} = \frac{180}{\pi} * \arctan \frac{y}{z}$$

Next, we calculated the angular velocity of the eye in both the horizontal and vertical coordinates by taking a first difference of the gaze angles with respect to time formula below:

$$\omega_{\mathrm{h}} = \Delta\theta_{\mathrm{h}}/\Delta t$$

$$\omega_{\mathrm{v}} = \Delta\theta_{\mathrm{v}}/\Delta t$$

Finally, we calculated the magnitude of the angular velocity (ω) at every timestamp from the horizontal and vertical components using:

$$\omega = \sqrt{\omega_h^2 + \omega_v^2}$$

To filter the samples where gaze was relatively stable, we used an adaptive threshold method for saccade detection described by Voloh et al. (2020). We selected an initial saccade velocity threshold θ_0 of $200°/s$. All eye movement samples with an angular velocity of less than θ_0 were used to compute a new threshold θ_1. θ_1 was three times the median absolute deviation of the filtered samples. If the difference between θ_1 and θ_0 was less than or equal to $1°/s$ θ_1 was selected as the final saccade threshold else, θ_1 was used as the new threshold to further filter samples and the above process was repeated. This was done until the difference between θ_{n} and $\theta_{\mathrm{n+1}}$ was less than or equal to $1°/s$. This way we arrived at the cluster of samples that belonged to fixations and the rest were classified as saccades.

After this, we calculated the duration of the fixations and saccades. To handle minuscule fixations and saccades, we labeled all samples with saccade duration less than 30 ms bracketed by fixations also as a fixation and merged the fixations before and after this interval. Similarly, we also labeled all fixation samples with a duration of less than 50 ms, interspersed between two saccades, also as a saccade. Following this, we recalculated the fixation and saccade durations. This procedure suitably handles misclassification of single samples or pairs of samples into fixations or saccades. Finally, we rejected all fixations with a duration greater than 3.5 times the median absolute deviation of the population fixation duration as well as fixations that were less than 100 ms long.

4.4.2 Grasp data

Subjects used the trigger button of the HTC Vive controller to virtually grasp the objects on the shelf and displace them to other locations. In the data, the trigger was recorded as a boolean which was set to TRUE when a grasp was initiated and was reset to FALSE when the grasp ended. Using the position of the controller in the world space, we determined the locations from the shelf where a grasp was initiated and ended. We also removed trials where the controller data showed implausible locations in the world space. These faulty data can be attributed to the loss of tracking during the experiment. Next, we removed grasping periods where the beginning and final locations of the objects on the shelf were the same. We calculated the interquartile range (IQR) of the participants' number of object displacements for the two trial types (EASY and HARD). To remove the outlying object displacements in the trials, we removed the trials with 1.5 times the IQR of object displacements. We also removed those trials with fewer than three object displacements.

After pre-processing, we were left with data from 48 subjects with 996 trials in total and 20.4 ± 1.8 trials per subject. Furthermore, we had eye-tracking data corresponding to 12,266 object displacements with 12.3 ± 5.11 object displacements per trial.

4.5 Data Analysis

4.5.1 Task-based behavioral differences

As a first step, we evaluated the behavioral differences induced by the task instructions for EASY and HARD trials. First, we evaluated the differences in the time taken to finish the trials between two trial types using two-sample independent t-tests. Secondly, in order to assess the planning behavior of the participants, we determined the optimal object displacements required to accomplish the tasks for the two trial types. To determine the optimal object displacements we designed a depth-first search algorithm that computed the minimum number of displacements required to sort the objects for the 5000 random initial configurations of 16 objects in 25 shelf locations for both the EASY and HARD trials constraints. We compared the mean number of object displacements made by the participants in the EASY and HARD trials with the model-based object displacements using independent t-tests.

4.5.2 Action Locked Gaze Control

In order to examine the spatial and temporal characteristics of eye movements while performing the sorting tasks, we divided the object and shelf locations into 7 regions-of-interest (ROIs) including previous, current, and next target object and target shelf. More specifically, the previous target object refers to the object that was handled in the previous action and the previous target shelf as the shelf where the previous target object was placed. Similarly, the current target object refers to the object that is picked up and placed on the target shelf in the current action and the next target object and next target shelf in the immediately following action in the sequence. All other regions which did not conform to the above 6 ROIs are categorized as 'other' and not relevant to the action sequence. In this format, we could parse the sequence of eye movements on the seven ROIs that are relevant for the planning and execution of the object-related actions.

We then analyzed the average fixation behavior aligned with the action initiation. For each grasp onset in a trial, we chose the time period from 2 seconds before grasp onset and 2 seconds after. We divided this 4 second period into bins of 0.15 seconds and calculated the number of fixations on the seven ROIs described above. For each time bin, we calculated the proportion of fixations on the ROIs per trial type (EASY, HARD). To find the time-points where there were significant differences between EASY and HARD trials for a given ROI, we used the cluster permutation method. Here, we use the t-statistic as a test statistic for each time-bin, where t is defined as:

$$t = \sqrt{N} * \frac{x}{\sigma}$$

where x is the mean difference between the trial types, σ is the standard deviation of the mean, and N is the number of subjects. We used a threshold for t at 2.14 which corresponds to the t-value at which the p-value is 0.05 in the t-distribution. We first found the time-bins where the t-value was greater than the threshold. Then, we computed the

sum of the t-values for these clustered time-bins which gave a single value that represented the mass of the cluster. Next, to assess the significance of the cluster, we permuted all the time-bins across trials and subjects and computed the t-values and cluster mass for 1000 different permutations. This gave us the null distribution over which we compared the cluster mass shown by the real data. To account for the multiple independent comparisons for the seven ROIs, we considered the significant clusters to have a Bonferroni corrected p-value less than 0.007 (0.05/7). In the results, we report the range of the significant time-bins for the seven ROIs for the two trial types the corresponding p-values.

4.5.3 Sequential Gaze Control in Action Planning and Execution

In order to study the function of eye movements for both action planning and execution, we divided each trial into 2 types of epochs. The action execution epoch spanned the time from the start of object displacement to the end. The action planning epochs started from the end of previous object displacement to the start of current object displacement. The schematic of this epoch creation is illustrated in Figure 1D. This division of time within each trial into separate epochs allows us to parse the role of eye movements in the planning and execution of object-related actions separately.

To examine the scan paths within the action planning and execution epochs we created transition matrices that show the origin and destination locations of the fixations on the 7 ROIs. We used the steps described by Hooge and Camps (2013) to first create the scan paths and then the transition matrices. We calculated the transition matrices summarizing gaze transitions from and to the 7 ROIs from the action planning and execution epochs for each object displacement. Using the transition matrices, we calculated the net and total transitions from and to each ROI. For every transition matrix **M** per action planning and execution epoch of object displacement, the net (M_{net}), and total (M_{total}) transition are defined as follows:

$$M_{net} = M - M^T$$

$$M_{total} = M + M^T$$

As discussed in Hooge and Camps (2013), if subjects make equal number of transitions between the ROI pairs in both directions, we can expect no transitions in the net transition matrix and can surmise that gaze was allocated more randomly. Conversely, with strong gaze guidance, we would expect more net transitions. Hence, using the net and total transitions per epoch, we calculated the relative net transitions (R) as:

$$R = \frac{\sum M_{net}}{\sum M_{total}}$$

We then took the mean of the relative net transitions per trial as a measure of gaze guidance behavior in that trial. Higher mean relative transitions would indicate gaze allocated to ROIs in a systematic and planned manner whereas, lower relative transitions would represent a random gaze allocation towards the ROIs.

Further, we also calculated the latency of the first fixation on the 7 ROIs from the start of the planning or execution epochs. We computed the median time to the first fixation per trial that would indicate time to first fixate on the ROIs for 50% of the action planning and action execution epochs. This method was used by Montfoort et al. (2007) and further applied by Hooge and Camps (2013) to capture the gaze attraction power of ROIs. As the action planning and execution epochs varied in duration, we normalized the time points by dividing them by the duration of the epoch. This way, the time elapsed since the start of an epoch is comparable to all epochs across trials and subjects.

4.5.4 Linear Mixed-Effects Models

We modeled the linear relationship of the relative net transitions dependent on the trial type (EASY, HARD), epoch type (planning, execution), and the number of object displacements and their interactions. All within-subject effects were modeled with random intercept and slopes grouped by subject. The categorical variables trial type and epoch type were effect coded (Schad et al., 2020), so that the model coefficients could be interpreted as main effects. The object displacement variable which pertained to the number of object displacements in the trial was coded as a continuous numeric variable and centered on zero mean. The model fit was performed using restricted maximum likelihood (REML) estimation (Corbeil and Searle, 1976) using the lme4 package (v1.1-26) in R 3.6.1. We used the L-BFGS-B optimizer to find the best fit using 20000 iterations. Using the Satterthwaite method (Luke, 2017), we approximated degrees of freedom of the fixed effects. The full model in Wilkinson notation (Wilkinson and Rogers, 1973) is defined as:

$$Relative_{Transitions} \sim 1 + trial_type * epoch_type * object_displacements$$
$$+(1 + trial_type + epoch_type|Subject)$$

We modeled the linear relationship of the median time to first fixation dependent on the trial type (EASY, HARD) and the 7 ROIs and their interactions. We computed two models for the action planning and execution epochs. All within-subject effects were modeled with random intercept and slopes grouped by subject. The categorical variables trial_type and ROI were effect coded as before and the model coefficients could be interpreted as main effects. For both models, we chose the latency of the first fixation on the current target object as the reference factor so that the latency of the first fixation of all other ROIs could be compared to it. The model fit was performed using restricted maximum likelihood (REML) estimation (Corbeil and Searle, 1976) using the lme4 package (v1.1-26) in R 3.6.1. We used the L-BFGS-B optimizer to find the best fit using 20000 iterations. Using the Satterthwaite method (Luke, 2017),

we approximated degrees of freedom of the fixed effects. The full model in Wilkinson notation (Wilkinson and Rogers, 1973) is defined as:

$$Fixation_{time} \sim 1 + trial_type * ROI$$
$$+(1 + trial_type * ROI|Subject)$$

Author Contributions

AK, PK, TS: conceived and designed the study. TS, PK: Procurement of funding. HN, KI: programmed the experiment in virtual reality. programmed the LeapMotion study. AK: data collection. AK, FNN: data preprocessing. AK: data analysis. AK: initial draft of the manuscript. AK, FNN, HN, KI, TS, PK: revision and finalizing the manuscript. All authors contributed to the article and approved the submitted version.

Data availability statement

Acknowledgement

We are grateful for the financial support by the German Federal Ministry of Education and Research for the project ErgoVR (Entwicklung eines Ergonomie-Analyse-Tools in der virtuellen Realität zur Planung von Arbeitsplätzen in der industriellen Fertigung)-16SV8052. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

D. H. Ballard, M. M. Hayhoe, and J. B. Pelz. Memory representations in natural tasks. *Journal of cognitive neuroscience*, 7(1):66–80, 1995. ISSN 0898-929X. doi: 10.1162/jocn.1995.7.1.66. URL `http://dx.doi.org/10.1162/jocn.1995.7.1.66`.

D. H. Ballard, D. Kit, C. A. Rothkopf, and B. Sullivan. A hierarchical modular architecture for embodied cognition. *Multisensory research*, 26(1-2):177–204, 2013. ISSN 2213-4794. doi: 10.1163/22134808-00002414. URL `http://dx.doi.org/10.1163/22134808-00002414`.

G. T. Buswell. How people look at pictures: a study of the psychology and perception in art. *Chicago University Press*, 198, 1935. URL `https://psycnet.apa.org/fulltext/1935-05800-000.pdf`.

V. Clay, P. König, and S. U. König. Eye tracking in virtual reality, 2019. URL `http://dx.doi.org/10.16910/JEMR.12.1.3`.

R. R. Corbeil and S. R. Searle. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics: a journal of statistics for the physical, chemical, and engineering sciences*, 18(1): 31–38, Feb. 1976. ISSN 0040-1706. doi: 10.1080/00401706.1976.10489397. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1976.10489397.

J. A. Droll and M. M. Hayhoe. Trade-offs between gaze and working memory use. *Journal of experimental psychology. Human perception and performance*, 33(6):1352–1365, Dec. 2007. ISSN 0096-1523. doi: 10.1037/0096-1523.33.6.1352. URL http://dx.doi.org/10.1037/0096-1523.33.6.1352.

W. Einhäuser, U. Rutishauser, and C. Koch. Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of vision*, 8(2):2.1–19, Feb. 2008. ISSN 1534-7362. doi: 10.1167/8.2.2. URL http://dx.doi.org/10.1167/8.2.2.

A. K. Engel, A. Maye, M. Kurthen, and P. König. Where's the action? the pragmatic turn in cognitive science. *Trends in cognitive sciences*, 17(5):202–209, May 2013. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2013.03.006. URL http://dx.doi.org/10.1016/j.tics.2013.03.006.

M. M. Hayhoe, A. Shrivastava, R. Mruczek, and J. B. Pelz. Visual memory and motor planning in a natural task. *Journal of vision*, 3(1):49–63, 2003. ISSN 1534-7362. doi: 10.1167/3.1.6. URL http://dx.doi.org/10.1167/3.1.6.

J. M. Henderson and T. R. Hayes. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature human behaviour*, 1(10):743–747, Oct. 2017. ISSN 2397-3374. doi: 10.1038/s41562-017-0208-0. URL http://dx.doi.org/10.1038/s41562-017-0208-0.

I. Hooge and G. Camps. Scan path entropy and arrow plots: capturing scanning behavior of multiple observers. *Frontiers in psychology*, 4:996, Dec. 2013. ISSN 1664-1078. doi: 10.3389/fpsyg.2013.00996. URL http://dx.doi.org/10.3389/fpsyg.2013.00996.

W. James. *The Principles of Psychology*. Cosimo Classics, 2007. ISBN 9781602063143. URL https://play.google.com/store/books/details?id=u4lOzQEACAAJ.

A. Keshava, A. Aumeistere, K. Izdebski, and P. Konig. Decoding task from oculomotor behavior in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications*, number Article 30 in ETRA '20 Short Papers, pages 1–5, New York, NY, USA, June 2020. Association for Computing Machinery. ISBN 9781450371346. doi: 10.1145/3379156.3391338. URL https://doi.org/10.1145/3379156.3391338.

A. Keshava, N. Gottschewsky, S. Balle, F. N. Nezami, T. Schüler, and P. König. ACTION AFFORDANCE AFFECTS PROXIMAL AND DISTAL GOAL-ORIENTED PLANNING. *bioRxiv*, pages 2021–2007, 2021.

S. Kollmorgen, N. Nortmann, S. Schröder, and P. König. Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS computational biology*, 6(5):e1000791, May 2010. ISSN 1553-734X, 1553-7358. doi: 10.1371/journal.pcbi.1000791. URL http://dx.doi.org/10.1371/journal.pcbi.1000791.

P. König, N. Wilming, T. C. Kietzmann, J. P. Ossandón, and others. Eye movements as a window to cognitive processes. *Journal of eye movement research*, 9(5), Dec. 2016. ISSN 1995-8692, 1995-8692. doi: 10.16910/jemr.9.5.3. URL https://pdfs.semanticscholar.org/8cc6/91616b435ed08263ff8007b2d27c0b8c1d27.pdf.

M. Land, N. Mennie, and J. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28(11):1311–1328, 1999. ISSN 0301-0066. doi: 10.1068/p2935. URL http://dx.doi.org/10.1068/p2935.

M. F. Land. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3): 296–324, May 2006. ISSN 1350-9462. doi: 10.1016/j.preteyeres.2006.01.002. URL http://dx.doi.org/10.1016/j.preteyeres.2006.01.002.

M. F. Land and S. Furneaux. The knowledge base of the oculomotor system. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 352(1358):1231–1239, Aug. 1997. ISSN 0962-8436. doi: 10.1098/rstb.1997.0105. URL http://dx.doi.org/10.1098/rstb.1997.0105.

M. F. Land and M. Hayhoe. In what ways do eye movements contribute to everyday activities? *Vision research*, 41 (25-26):3559–3565, 2001. ISSN 0042-6989. doi: 10.1016/s0042-6989(01)00102-x. URL http://dx.doi.org/10.1016/s0042-6989(01)00102-x.

M. F. Land and P. McLeod. From eye movements to actions: how batsmen hit the ball. *Nature neuroscience*, 3(12): 1340–1345, Dec. 2000. ISSN 1097-6256. doi: 10.1038/81887. URL http://dx.doi.org/10.1038/81887.

S. G. Luke. Evaluating significance in linear mixed-effects models in R. *Behavior research methods*, 49(4):1494–1502, Aug. 2017. ISSN 1554-351X, 1554-3528. doi: 10.3758/s13428-016-0809-y. URL http://dx.doi.org/10.3758/s13428-016-0809-y.

J. N. MacGregor and Y. Chu. Human performance on the traveling salesman and related problems: A review. *The Journal of Problem Solving*, 3(2):2, 2011. ISSN 1932-6246. doi: 10.7771/1932-6246.1090. URL https://docs.lib.purdue.edu/jps/vol3/iss2/2/.

D. L. Mann, H. Nakamoto, N. Logt, L. Sikkink, and E. Brenner. Predictive eye movements when hitting a bouncing ball. *Journal of vision*, 19(14):28, Dec. 2019. ISSN 1534-7362. doi: 10.1167/19.14.28. URL http://dx.doi.org/10.1167/19.14.28.

F. Mars and J. Navarro. Where we look when we drive with or without active steering wheel control. *PloS one*, 7(8): e43858, Aug. 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0043858. URL http://dx.doi.org/10.1371/journal.pone.0043858.

J. S. Matthis, J. L. Yates, and M. M. Hayhoe. Gaze and the control of foot placement when walking in natural terrain. *Current biology: CB*, 28(8):1224–1233.e5, Apr. 2018. ISSN 0960-9822, 1879-0445. doi: 10.1016/j.cub.2018.03.008. URL http://dx.doi.org/10.1016/j.cub.2018.03.008.

N. Mennie, M. Hayhoe, and B. Sullivan. Look-ahead fixations: anticipatory eye movements in natural tasks. *Experimental brain research. Experimentelle Hirnforschung. Experimentation cerebrale*, 179(3):427–442, May 2007. ISSN 0014-4819. doi: 10.1007/s00221-006-0804-0. URL http://dx.doi.org/10.1007/s00221-006-0804-0.

I. Montfoort, M. A. Frens, I. T. C. Hooge, G. C. L.-v. Haselen, and J. N. van der Geest. Visual search deficits in Williams-Beuren syndrome. *Neuropsychologia*, 45(5):931–938, Mar. 2007. ISSN 0028-3932. doi: 10.1016/j.neuropsychologia.2006.08.022. URL http://dx.doi.org/10.1016/j.neuropsychologia.2006.08.022.

J. Navarro, E. Hernout, F. Osiurak, and E. Reynaud. On the nature of eye-hand coordination in natural steering behavior. *PloS one*, 15(11):e0242818, Nov. 2020. ISSN 1932-6203. doi: 10.1371/journal.pone.0242818. URL http://dx.doi.org/10.1371/journal.pone.0242818.

D. A. Norman and T. Shallice. Attention to action. In R. J. Davidson, G. E. Schwartz, and D. Shapiro, editors, *Consciousness and Self-Regulation: Advances in Research and Theory Volume 4*, pages 1–18. Springer US, Boston, MA, 1986. ISBN 9781475706291. doi: 10.1007/978-1-4757-0629-1_1. URL https://doi.org/10.1007/978-1-4757-0629-1_1.

F. J. Parada and A. Rossi. Perfect timing: Mobile brain/body imaging scaffolds the 4e-cognition research program. *The European journal of neuroscience*, May 2020. ISSN 0953-816X, 1460-9568. doi: 10.1111/ejn.14783. URL http://dx.doi.org/10.1111/ejn.14783.

J. B. Pelz and R. Canosa. Oculomotor behavior and perceptual strategies in complex tasks. *Vision research*, 41(25-26): 3587–3596, 2001. ISSN 0042-6989. doi: 10.1016/s0042-6989(01)00245-0. URL http://dx.doi.org/10.1016/s0042-6989(01)00245-0.

Z. Pizlo and Z. Li. Solving combinatorial problems: the 15-puzzle. *Memory & cognition*, 33(6):1069–1084, Sept. 2005. ISSN 0090-502X. doi: 10.3758/bf03193214. URL http://dx.doi.org/10.3758/bf03193214.

D. J. Schad, S. Vasishth, S. Hohenstein, and R. Kliegl. How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of memory and language*, 110:104038, Feb. 2020. ISSN 0749-596X. doi: 10.1016/j.jml.2019.104038. URL https://www.sciencedirect.com/science/article/pii/S0749596X19300695.

F. Schumann, W. Einhäuser-Treyer, J. Vockeroth, K. Bartl, E. Schneider, and P. König. Salient features in gaze-aligned recordings of human visual input during free exploration of natural environments. *Journal of vision*, 8(14):12.1–17, Oct. 2008. ISSN 1534-7362. doi: 10.1167/8.14.12. URL http://dx.doi.org/10.1167/8.14.12.

B. Sullivan, C. J. H. Ludwig, D. Damen, W. Mayol-Cuevas, and I. D. Gilchrist. Look-ahead fixations during visuomotor behavior: Evidence from assembling a camping tent. *Journal of vision*, 21(3):13, Mar. 2021. ISSN 1534-7362. doi: 10.1167/jov.21.3.13. URL http://dx.doi.org/10.1167/jov.21.3.13.

B. T. Sullivan, L. Johnson, C. A. Rothkopf, D. Ballard, and M. Hayhoe. The role of uncertainty and reward on eye movements in a virtual driving task. *Journal of vision*, 12(13):19, Dec. 2012. ISSN 1534-7362. doi: 10.1167/12.13.19. URL http://dx.doi.org/10.1167/12.13.19.

S. Van der Stigchel. An embodied account of visual working memory. *Visual cognition*, 28(5-8):414–419, Sept. 2020. ISSN 1350-6285. doi: 10.1080/13506285.2020.1742827. URL https://doi.org/10.1080/13506285.2020.1742827.

B. Voloh, M. R. Watson, S. König, and T. Womelsdorf. MAD saccade: statistically robust saccade threshold estimation via the median absolute deviation. *Journal of eye movement research*, 12(8), May 2020. ISSN 1995-8692. doi: 10.16910/jemr.12.8.3. URL http://dx.doi.org/10.16910/jemr.12.8.3.

G. N. Wilkinson and C. E. Rogers. Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C, Applied statistics*, 22(3):392, 1973. ISSN 0035-9254, 1467-9876. doi: 10.2307/2346786. URL https://www.jstor.org/stable/10.2307/2346786?origin=crossref.

M. Wilson. Six views of embodied cognition. *Psychonomic bulletin & review*, 9(4):625–636, Dec. 2002. ISSN 1069-9384. doi: 10.3758/bf03196322. URL http://dx.doi.org/10.3758/bf03196322.

A. L. Yarbus. *Eye Movements and Vision*. Springer, Nov. 2013. ISBN 9781489953797.