bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Hierarchical structure is employed by humans during visual motion perception

Johannes Bill[a,b,2], Hrag Pailian[b], Samuel J Gershman[b,c,1], and Jan Drugowitsch[a,c,1,2]

[a]Department of Neurobiology, Harvard Medical School
[b]Department of Psychology, Harvard University
[c]Center for Brain Science, Harvard University

Abstract

In the real world, complex dynamic scenes often arise from the composition of simpler parts. The visual system exploits this structure by hierarchically decomposing dynamic scenes: when we see a person walking on a train or an animal running in a herd, we recognize the individual's movement as nested within a reference frame that is itself moving. Despite its ubiquity, surprisingly little is understood about the computations underlying hierarchical motion perception. To address this gap, we develop a novel class of stimuli that grant tight control over statistical relations among object velocities in dynamic scenes. We first demonstrate that structured motion stimuli benefit human multiple object tracking performance. This performance advantage is well-explained by a Bayesian observer model that exploits hierarchical structure. A second experiment, using a motion prediction task, reinforced this conclusion and provided fine-grained information about how the visual system flexibly exploits motion structure.

The visual scenes our brains perceive in everyday life are filled with complex dynamics. Information hitting the retina changes not only with every motion in the scene, but also with every head movement and saccade. Noise, occlusions, and ambiguities further make visual information inherently unreliable. In order to maintain stable, coherent percepts in the face of complex and unreliable inputs, our brains exploit the spatially and temporally structured nature of the environment [1].

Motion structure refers to statistical relations of velocities. One form of structure common in natural scenes is motion grouping: when a rigid object is set in motion, all of its visual features move coherently. This structure allows us to infer the existence of objects based on the coherent motion of features—the Gestalt grouping cue known as *common fate* [2]. Grouping based on common fate has been shown to influence our ability to track objects [3–6], to search displays [7], and to store information in short-term memory [8].

However, the strict definition of common fate is too brittle to accommodate natural scenes in which visual features do not move together rigidly and yet are still grouped together. In some cases this is because we perceive objects as deforming non-rigidly. In other cases, we perceive objects that are hierarchically structured [9–11]: the parts of an object move rigidly *relative* to a reference frame (the object), which itself could be a rigidly moving part of another object, and so on. For example, we perceive the motion of hands relative to the motion of arms, and the motion of arms relative to the motion of the torso (**Fig. 1A,B**). The entire body may be moving relative to another reference frame (a train or escalator). The perception of hierarchically organized motion suggests a powerful "divide-and-conquer" strategy for parsing complex dynamic scenes.

Recent work has formalized the representation and discovery of hierarchical motion structures [10]. However, we still know relatively little about if and how the visual system exploits this structure for visual tasks like object tracking. We address this question in two experiments, one testing multiple object tracking (MOT, [12]) and one testing multiple object trajectory prediction. A key innovation of these experiments is the use of tightly controlled stimuli that isolate the signatures of hierarchical structure from other forms of

[1]Equal contribution, [2]Corresponding authors: johannes_bill@hms.harvard.edu and jan_drugowitsch@hms.harvard.edu

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

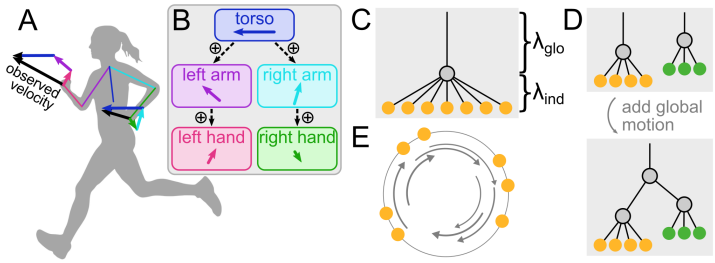

Figure 1. Modular representation of hierarchical motion structure. (*A*) Observed velocity components of a running human. The hands inherit motion from the arms, which inherit motion from the torso. (*B*) Corresponding nested hierarchy of motion relations. Observed velocity is the sum of local motion components. (*C*) Motion graph describing global motion with a strong (strength λ_{glo}) shared motion source (gray node) and weaker (strength λ_{ind}) individual motion sources (orange nodes). Here, the global motion source is not directly observed (i.e., latent) but introduces correlations in the motion of observable objects (orange). (*D*) Two motion clusters (top) can be embedded into a deep hierarchy by adding another latent motion source at the tree's root (bottom). (*E*) Illustration of a stimulus of stochastically rotating dots with global motion structure. These are the class of stimuli used in the experiments.

motion structure. We show that a Bayesian observer model, when equipped with the appropriate structural representation, can qualitatively explain human behavior across a variety of motion structures. In the MOT task, we find that improvement of tracking performance for structured stimuli cannot be explained without exploiting structure knowledge during inference, indicating that humans make use of motion structure knowledge when perceiving dynamic scenes. In the prediction task, we uncover the employed motion structure knowledge of our participants, finding that humans can flexibly recruit different motion relations for scene parsing, including deep motion hierarchies which nest groups of objects within another moving reference frame.

Results

Representation of motion structure

In this section, we introduce a representation of hierarchical motion structure. To motivate this representation, consider a flock of birds. The velocities of individual animals in the flock are naturally described by the sum of a global motion component, which is shared across all birds, and an individual component for each animal. Only the aggregate motion is observed; the underlying motion components are latent variables that help us organize our perception.

To formalize the idea of shared, latent motion components, we assume all velocities in a visual scene to be driven by motion sources (nodes in the graph in **Fig. 1C**) that either represent observable objects (filled nodes) or are latent (unfilled nodes). A motion source inherits velocity from a parent source when connected by an edge, thus supporting tree-like hierarchies. Motion sources can be shared by multiple objects, and the total velocity of an object is the sum of all inherited motions. In addition to the graph connectivity, each motion source has a motion strength λ that determines the source's contribution strength to the speed of dependent observable objects. The simple "global motion" motif in **Fig. 1C** might, for example, to a first order describe the motion structure underlying a flock of birds, where larger motion strengths are illustrated by larger vertical distances (curly braces) between the motion sources.

The separation of motion composition (graph connectivity) and motion strength (vertical node location) gives rise to a flexible, modular representation of motion structure, covering many real-world scenes such as independent motion, clustered motion or deep motion hierarchies. The graph in **Fig. 1D** (top), for example, describes a faster orange and a slower green cluster of otherwise independently moving composed objects (e.g., two non-interacting flocks of birds). A further global motion component, such as an observer moving his head, would introduce another motion source at the root of the tree (see **Fig. 1D** bottom). This illustrates how our motion representation can be used in a modular fashion to describe deep nested motion hierarchies. In fact, any tree-like motion structure built in this way can be represented by a *motion structure matrix* $\mathbf{L}$, accommodating both composition and strengths (see *Material and Methods*), such that we will often refer to

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

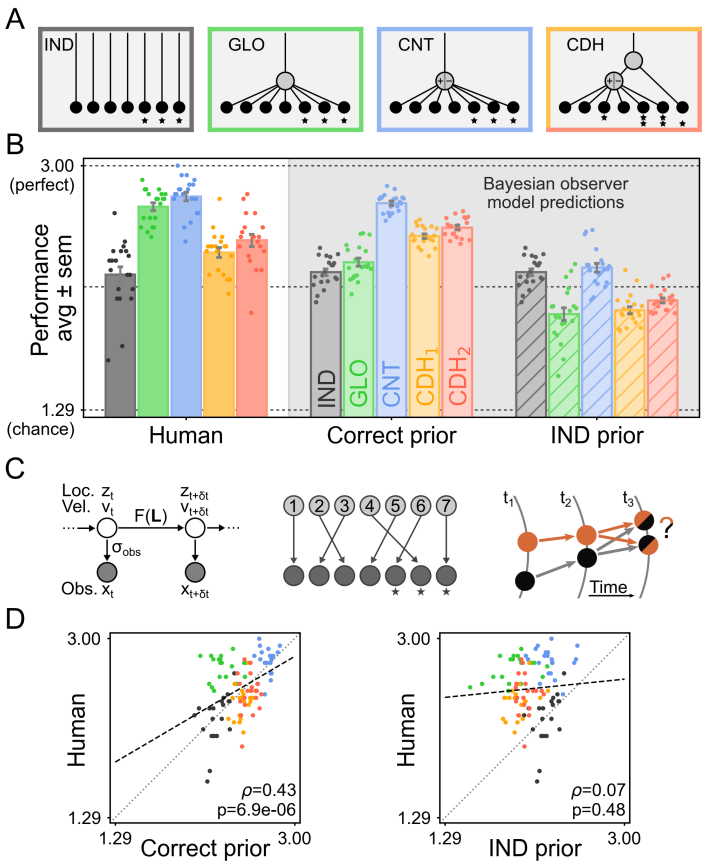

Figure 2. Use of motion structure knowledge during multiple object tracking. (*A*) Tested motion conditions included independent (IND), global (GLO), counter-rotating (CNT), and counter-rotating deep hierarchical (CDH) motion. ⋆'s mark targets. Two different target sets were tested for CDH motion. (*B*) Average performance (number of correctly identified targets) on different motion conditions by human participants, the Bayesian observer model using the correct motion structure, and the Bayesian observer disregarding motion relations (IND prior). Using motion structure during inference is required to explain human performance gains on motion-structured stimuli. (*C*) The Bayesian observer model consists of a Kalman filter with motion structure prior $\boldsymbol{L}$ (left) and a mental assignment of dot identities (center). Perceptual and neural noise can lead to ambiguous assignments and, ultimately, errors in the reported target set (right). (*D*) Scatter plot of the data in (B), one dot per participant and motion condition. The observer model explains human responses well only when using motion knowledge during inference.

specific motion structures by their associated matrices $\boldsymbol{L}$.

This modular representation enabled us to generate motion-structured visual stimuli, which isolated the computational role of motion structure in dynamic visual scenes, and were thus particularly suited for psychophysics experiments. Specifically, we strove to design stimuli in which all object properties and statistical relations among objects were dominated by *structure in velocities*, while keeping other factors like individual object velocities or spatial structure uninformative. To do so, we generated the velocities of observable objects by random draws from a continuous time stochastic process (namely a multivariate Ornstein-Uhlenbeck process [13]; see *Material and Methods* for details) that yielded smooth random trajectories with the desired statistical motion relations imposed by a chosen motion structure $\boldsymbol{L}$. In this process, motion sources play the role of random forces which accelerate or decelerate any dependent object. For example, the latent global source in **Fig. 1C** would induce correlations among the velocities of all observed objects. The resulting trajectories are mathematically tractable and feature real-world properties, such as inertia and friction, without making any assumptions about specific trajectory realizations. To further remove any persistent spatial structure among objects, we put the stimulus on a circle (illustrated in **Fig. 1E**; see *Appendix* for example video), making the locations of objects asymptotically independent. The mathematical tractability of the trajectory-generating process was crucial for precise stimulus control. Knowing the exact joint probability distribution across object velocities allowed us to vary velocity correlations induced by motion structure while keeping the motion velocity statistics of individual objects unchanged, thus making structure in velocities the dominant feature in the presented scenes.

Motion structure improves visual tracking performance

We first asked whether motion structure knowledge benefits humans in a multiple object tracking (MOT) task [12]. Evidence exists that structure among objects, such as grouping [14], symmetry [6] or global translation [4], impacts tracking performance, but hierarchical structure has not been investigated before.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

In the MOT task, $K = 7$ dots rotated about a circle with 3 dots being initially marked as targets. After a few seconds, all dots changed to identical appearance while dot motion continued for another 6 s. After that the dot motion stopped, and participants had to re-identify the initially marked targets. We tested 20 participants on four different blocked motion conditions of 30 trials each, with motion graphs shown in **Fig. 2A** (targets marked by $\star$; see *Appendix* for example trial videos). Independent (IND) motion, the standard MOT task, served as a baseline. In the global (GLO) condition, dot motion is composed of a dominating shared stochastic motion source, as well as small individual per-dot motion sources to break emergent spatial patterns. In counter-rotating (CNT) motion, the latent motion source affects dot velocities in opposite directions: similar to gears, the force accelerating the dots $k = 1,2,3$ in one direction, accelerates dots $k = 4,5,6,7$ (three of which are targets) in the reverse direction. The last motion condition is a counter-rotating deep hierarchy (CDH) in which two 3-dot groups are driven by a counter-rotating motion source that shares a global motion source with the 7th dot. This deep nested structure generates distinguishable velocity patterns that cannot be approximated by any shallow (max. one latent source layer) motion structure. In the CDH condition, we tested two different target sets (CDH_1 and CDH_2) to probe set-specific effects in tracking performance. We measured participant performance as the average number of correctly identified targets within each condition. Since the aim of the experiment was to test how humans can exploit motion structure knowledge rather than learn it, at the beginning of each motion block, participants were explicitly presented with a diagram and three demonstration trials of the motion structure that would be presented in the upcoming block. Overall average stimulus speed was titrated on a per-participant basis to approximately reach a performance of 2.15 (midpoint between chance level ($3 \times 3/7$) and perfect (3)) correctly identified dots in the IND condition. This per-participant speed level was subsequently maintained for the rest of the experiment. During the following data collection, the marginal motion statistics of individual dots were then held constant for each participant across all conditions $\boldsymbol{L}$ and dots k such that conditions only differed in their dot velocity correlations, as determined by $\boldsymbol{L}$. A separate IND condition block (see left-most bar in **Fig. 2B**) that was not used to adjust the stimulus speed confirmed the validity of the adjustment and marks the reference for performance changes in motion-structured stimulus conditions. The performance on the four conditions with motion structure (GLO, CNT, CDH_1 and CDH_2) are shown in **Fig. 2B** (left) next to the IND reference performance. The introduction of structured motion significantly impacted dot tracking performance, ($p \approx 1.03 \times 10^{-14}$, Greenhouse-Geisser corrected repeated measures ANOVA) resulting in a significant performance boost in all motion conditions (one-sided paired t-tests, $p \approx 1.2 \times 10^{-8}$, 2.4×10^{-9}, 7.9×10^{-3}, 2.6×10^{-4}, respectively). In conclusion, motion structure clearly improved human tracking performance.

The computational origin of MOT performance gains

The observed boost in tracking performance for structured stimuli could simply be a byproduct of dot velocity correlations, that is, an intrinsic stimulus property, rather than the result of using motion structure knowledge by the observer. For instance, dots with positively correlated velocities might have less proximity events during a trial, making their confusion less likely [15]. To distinguish the contribution of employed motion structure knowledge from stimulus-intrinsic factors, we extended a Bayesian MOT observer model by Vul and colleagues [16] to incorporate motion structure priors $\boldsymbol{L}$. The deliberately simple model (illustrated in **Fig. 2C**) includes only the core components required to perform the MOT task, making it a minimalistic, normative model of motion-structured MOT performance (see *Material and Methods* for details). Similar to Vul et al. [16], visible dot locations $\boldsymbol{x}_t$ in individual video frames are subject to perceptual and neural noise σ_{obs} while being integrated with mental estimates of location $\boldsymbol{z}_t$ and velocity $\boldsymbol{v}_t$ via a multi-dimensional Kalman filter (**Fig. 2C** left). To capture structured motion, we extended this Kalman filter to incorporate the motion structure matrix $\boldsymbol{L}$ as a Bayesian prior distribution of how dot velocities are expected to evolve over time. A global (GLO) motion prior, for example, would favor positively correlated dot trajectories. The Kalman filter alone, however, assumes that every observed dot location is labeled by the dot's identity, making dot confusions, and therefore imperfect dot tracking, impossible. To produce such confusion, we needed to additionally model the mental assignment of dot identities to the visually identically looking dots on the screen (**Fig. 2C** center, cf. [16]), which is the psychophysical counterpart to the binding problem [17] in neuroscience. Errors in this mental assignment can lead to imperfect re-identification of the target dots at the end of an MOT trial, and arise when dots come close [18] or even cross: uncertainty in internal location

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

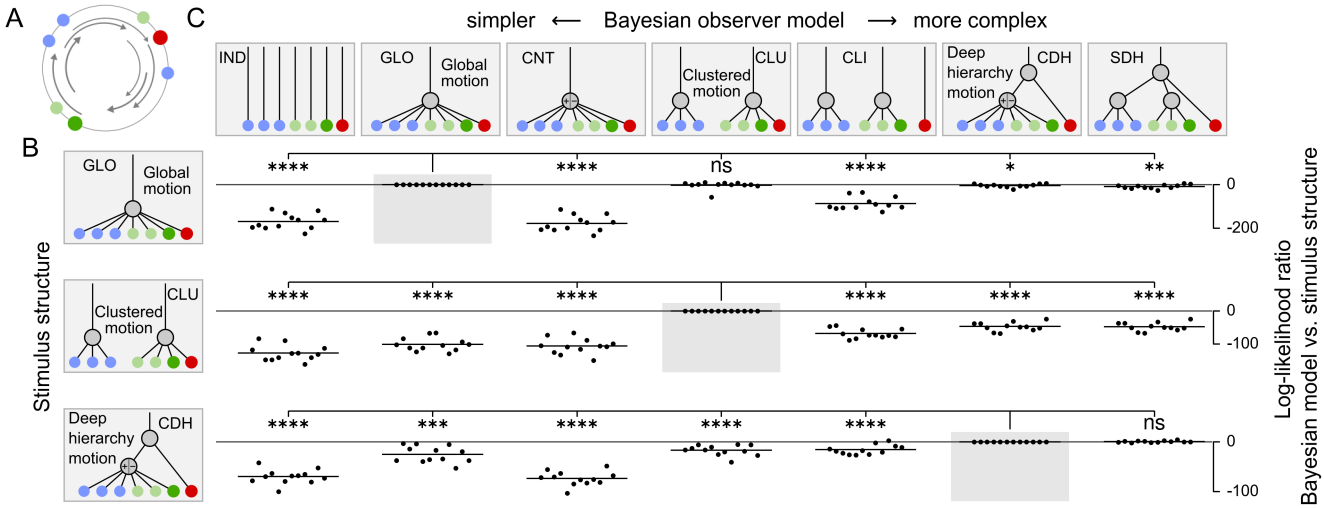

Figure 3. Revealing human motion priors in a multiple object prediction task. (*A*) Illustration of the stimuli. The highlighted green and red dot disappeared after 5 s. Participants had to predict their location at the end of the trial. Dots were color-coded to indicate their role in the motion structure. Here, a global (GLO) stimulus condition is illustrated. (*B*) Tested motion conditions included global (GLO), clustered (CLU) and counter-rotating deep hierarchical (CDH) motion. (*C*) Top: putative motion priors tested for explaining human responses via a Bayesian observer model, ordered by their complexity (see main text for description). Each cell shows per-participant log-likelihood model fit ratio for a particular motion prior, compared to the correct prior underlying the stimulus (indicated by gray background). Negative values indicate that the participant's behavior was better explained by the correct motion prior. One dot per participant, stimulus condition and putative motion prior. Horizontal lines show mean log-likelihood ratios across participants. Humans flexibly employed correctly structured motion priors. $*$'s indicate significance of paired t-tests (p 0.05, 0.01, 10^{-3} and 10^{-4} respectively).

estimates z_t can render multiple mental assignments possible, as exemplified in **Fig. 2C** (right).

We presented the observer model with the exact same trials that had been shown to human participants. The single free parameter of the model, σ_{obs}, was adjusted to match average human performance in the baseline IND motion condition, and was subsequently fixed to predict performance in all other conditions. This resulted in a remarkable qualitative match between human performance and the minimalistic observer model's predictions (**Fig. 2B** left vs. center; see *Appendix* for a discussion of the quantitative mismatch in the GLO condition).

To distinguish if the performance gain for structured motion stimuli was simply due to stimulus-intrinsic factors, we repeated the above procedure with a Bayesian observer that did not exploit complex motion structure knowledge (by making the observer assume an IND motion prior), even when motion structure was present in the stimuli. As **Fig. 2B** (right vs. left) reveals, this modification caused all motion structure-dependent performance gains to vanish. Hence the observed performance gain of human participants cannot be explained by stimulus-intrinsic factors alone. Instead, humans appear to make use of motion structure knowledge when performing the task.

The finding that the Bayesian observer with correct motion prior explains large parts of human performance, while the motion structure-agnostic observer (IND prior) does not, is further supported by scatter plots shown in **Fig. 2D** (replotting the data from panel B; the two Pearson correlation coefficients are significantly different, p 0.001 [19, 20]; additional analysis). This visual impression is corroborated by pre-registered statistical measures: while the cosine similarity between per-participant human and correct-prior observer model performance gains ("structured condition minus IND condition") is highly significant ($p10^{-5}$, one-tailed test against H0: positive cosine similarity is a random effect), the one between per-participant human and IND-prior observer model performance gains is not ($p \approx 0.99$).

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Revealing the structure of human motion priors

The MOT task revealed that humans make use of motion structure knowledge during visual tasks. However, the limited information provided by the participants' responses (three chosen dots per trial) didn't allow us to identify which exact motion features were employed. We therefore developed a second experiment, multiple object prediction, that grants a more fine-scaled insight into the cognitive machinery underlying motion perception.

Multiple object prediction uses partially observed dynamic visual scenes to test human perception in face of uncertainty. As before, in each trial, seven dots rotated stochastically about a circle (see **Fig. 3A**). After 5 s, however, two target dots (the larger green and red dots) became invisible while the motion of the other dots remained visible. After another 1.5 s the scene froze, at which point participants had to predict the location of both target dots with a computer mouse. We used highly volatile motion-structured stimuli with quickly changing dot velocities, such that the use of motion structure information was indispensable for making predictions. We further colored the dots according to their role in the motion structure to prevent dot confusion. In contrast to previous single-dot prediction tasks [11], the rationale for simultaneously predicting two dot locations is that the covariance pattern of errors convey additional information about the motion structure assumptions of our participants. For example, if global motion was assumed, we would expect the participants to jointly over- or underestimate the red and green targets' final locations.

We tested 12 participants on 3 motion conditions each: global motion (GLO), clustered motion (CLU) and counter-rotating deep hierarchical motion (CDH) (motion graphs in **Fig. 3B**; example videos in *Appendix*), with 100 trials per condition. As before, participants were briefly trained on all motion structures and were informed about the specific structure underlying each trial. Participants reported the task to be challenging, but performed reasonably well (average absolute error to optimal predictions: $36.2° \pm 4.0°$, mean $\pm$ SD across participants; additional analysis).

In order to identify the structure of the motion prior employed by each participant in each stimulus condition, we formulated a Bayesian observer model of human responses in the prediction task. This model was the same as our MOT model, but without the dot confusion component, and with location observations assumed to be practically noise-free ($\sigma_{\text{obs}} = 0$, except for the invisible dots for which no further observations were provided). After each trial, motion-structured Kalman filters with different putative motion structure priors $\boldsymbol{L}$ predicted the statistically optimal, most likely location of the target dots based on the motion of the other dots, and the uncertainty in this prediction. Building on [21], we linked this prediction to the locations reported by the participants by assuming additional response variability that scaled with prediction uncertainty (correlated across the two target dots), motor noise (uncorrelated across dots), and the possibility to confuse the red and green dots when reporting their locations (see *Material and Methods* for details). The first two variability sources were modeled as Gaussians, owing to the Gaussianity of observed human response error distributions (see *Appendix*, Fig. S1; additional analysis), whereas the third variability source was implemented by a small swapping probability. Overall, this led to three model parameters that we fit by maximum-likelihood separately for each participant, condition, and putative motion structure prior. Applied to simulated behavior with the same trials, we found that this procedure was able to correctly recover the motion structure underlying simulated responses (*Appendix*, Fig. S2).

To identify the most likely motion structure prior employed by the participants, we compared across 7 structurally different putative motion priors how well the model can explain the participants' trial-by-trial predictions (pre-registered; see *Appendix*, Fig. S3 for additional comparison across a larger range of motion structures, yielding similar results). **Fig. 3C** shows the resulting log-likelihood ratios, with putative motion priors sorted from simpler to more complex motion structures (note that all models have the same number of free parameters). For the global (GLO) motion condition (first row of **Fig. 3C**), human responses are best explained either by the correct global motion prior or by more complex priors which would be similarly suited for solving the task (e.g., because they contain a global motion source). Motion priors without a global motion source, such as IND, CNT, and CLI, resulted in significantly worse model fits. For the clustered (CLU) motion condition (second row), all alternative model priors led to worse model fits. Together, these results establish confidence in the applicability of the Bayesian observer model.

Our most complex motion condition, the counter-rotating deep hierarchy (CDH; **Fig. 3** bottom row), was designed to require inference over both latent motion sources for correct prediction. In other words, none

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

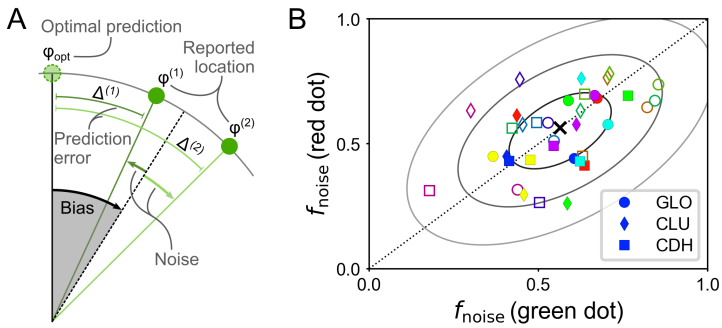

Figure 4. Bias-variance decomposition for the prediction task. (*A*) Human prediction errors are assumed to be the sum of a systematic (bias) and a stochastic (noise) component. The relative contribution of both components can be estimated from repetitions of the same trial, leading to responses $\varphi^{(1)}$ and $\varphi^{(2)}$, and associated errors $\Delta^{(1)}$ and $\Delta^{(2)}$. (*B*) Noise factors for the green and the red dot, one marker per participant (color & filling) and motion condition (shape). A combination of systematic and stochastic errors underlie human sub-optimality. Black cross and ellipsoids indicate mean and iso-density curves (1-3 SDs) of a bivariate Gaussian fitted to all noise factors.

of the tested shallow (non-nested) motion priors can mimic the error covariance pattern expected under a CDH prior. Global (GLO) motion would miss the counter-rotating component of the blue and green groups. Counter-rotating (CNT) motion would miss the global motion component. A clusters-plus-independent (CLI) motion prior (see **Fig. 3C** top) could approximately predict the green dot's location, but lacks the global component for predicting the red dot. The only structures that could approximate a CDH covariance pattern are other deep nested hierarchies, such as a standard deep hierarchy (SDH) that employs a third latent motion source. The result of Bayesian model comparison (bottom row in **Fig. 3C**) clearly favors deep hierarchical motion priors for explaining human predictions. Interestingly, several participants reported that they did not consciously use hierarchical information in the CDH motion condition. The fact that their prediction errors nonetheless indicate a deep hierarchical prior points to the subconscious use of hierarchical motion knowledge in perception by our visual system. Overall, our results strongly suggest that human observers are able to use deep hierarchical motion structure when perceiving dynamic scenes.

Systematic and stochastic errors underlie human sub-optimality

Even though participants employed the qualitatively correct motion structure in the prediction task, they featured additional, suboptimal variability when compared to a statistically optimal observer. This variability could arise from consistent deterministic bias, indicating a systematic mis-integration of the stimuli, or stochastic fluctuations, indicating potentially noisy computations. To determine the contribution of each alternative to overall variability, we repeated each trial twice within each condition, supporting a bias-variance-decomposition [22] of this variability (**Fig. 4A**). Specifically, if all of the variability was due to deterministic biases, participants should feature the exact same, systematic errors, $\Delta^{(1)} = \Delta^{(2)}$, in both trials. If, in contrast, all variability was purely stochastic, the observed errors should be uncorrelated across paired trials. This idea allowed us to quantify the ratio between bias and variance by one noise factor f_{noise} per participant and motion condition, with $f_{\text{noise}} = 0$ and $f_{\text{noise}} = 1$ denoting purely bias-driven and purely noise-driven observers, respectively (see *Material and Methods* for details). We evaluated the noise factor separately for the green and red target dot, with the results shown in **Fig. 4B**. Note that these results are additional analyses. We found that human sub-optimality is a combination of systematic and stochastic errors. The noise factors of both dots were on average equally strong (black $\times$). The positive correlation between red and green dot noise factors ($\rho = 0.51$, Pearson correlation) suggests a participant-dependent level of "noisiness". This impression is further supported by consistently higher or lower noise factors across participants (colors & filling in **Fig. 4B**; p 10^{-4}, one-way Welch's ANOVA).

Discussion

We have shown that humans make use of motion structure knowledge during demanding perceptual tasks, and that they can flexibly employ structured motion priors for the task at hand. Beyond shallow motion motifs, such as global motion or clustered motion, we were able to show that humans can even use deep

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

nested motion hierarchies. Key to revealing the covert motion priors of human participants was the design of analytically tractable stimuli that capitalize on motion relations. Analytical tractability facilitated the development of normative Bayesian observer models, our mathematical scalpel for dissecting the use of motion priors by the participants from stimulus-intrinsic contributions to visual motion perception.

Understanding motion perception as Bayesian inference has a rich scientific history [10, 11, 16, 23–27]. Bayesian models of low-level perception [23–25] were extended to explain human multiple object tracking of independent dot movement [16, 27]. Our work extends this line of research to structured motion, revealing the use of structured motion priors during tracking. Our aim to resolve the fine structure of human motion priors led to the development of a novel probabilistic multiple object prediction task that augments single object prediction [11] with velocity covariances within a tractable protocol. Including posterior covariance matrices in the observer model greatly enhanced the log-likelihood ratios in the prediction task beyond the resolution achievable in single object prediction (see *Appendix*, Fig. S4). While we had originally developed our modular matrix representation $\boldsymbol{L}$ of hierarchical structure as an extension to the motion trees in [10], the proposed decomposition into graph connectivity and motion strengths bears some resemblance to the semantic structure interpretation of singular value decomposition applied in [28] to categorical data and to Pythagorean tree embeddings used in [29] for describing language structure. Flexible structure representations are expected to foster cognitive science research on how humans infer the motion structure of a visual scene—potentially recruiting from a set of motion features—and how such motion features for modular combination could be learned in the first place.

Biologically, the importance of motion for visual scene perception is reflected in the tuning of cells in primate visual areas. Neurons in area MT are frequently tuned to the speed and direction of velocity within their receptive field [30], while downstream areas, like MSTd, encode progressively richer motion primitives such as selective tuning to expansion, rotation and spiraling [31, 32]. This points to a feature repertoire that is tailored to behaviorally relevant stimuli such as radial expansion (the visual pattern on the retina of a forward-moving observer) or rotation (when tilting your head to the left/right). Little, however, is known about how such motion primitives are further recruited in neural circuits for high-level, hierarchical motion processing. Our motion structure representation is compatible with existing neural implementations of Bayesian sensory integration, such as the neural Bayesian filtering model of Beck et al. [33]. In combination with physiology-based models of motion integration [34, 35], this could bring forth normative neural models of structured motion perception in higher visual areas, and guide experiments on the neural code along the visual motion pathway.

Material and Methods

The study has been pre-registered [36] prior to data collection, and all pre-registered analyses are presented in the main text. Any additional analyses are labeled as such.

Motion structure matrix representation

We describe the motion of K visible objects that are driven by M motion sources. Usually, $M\ K$ since, besides shared latent sources, each object can feature individual motion. We represent how motion source m affects object k via a composition matrix of motion motifs $\boldsymbol{C} \in \mathbb{R}^{K\times M}$. If m drives k, we set $C_{km} = 1$; if m does not affect k, we set $C_{km} = 0$; for counter-rotating motion, we set $C_{km} = +1$ and -1 for opposing directions. Thus, each column of $\boldsymbol{C}$ encodes a motion motif (e.g., all 1s for global motion), which is tied to motion source m. Each motion source has an associated real-valued motion strength $\lambda_m \geq 0$, and we define the motion structure matrix $\boldsymbol{L}$ as the product of composition and strengths, $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{\Lambda}$ with $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, .., \lambda_M)$, i.e., strength λ_m scales the m-th motion motif.

Motion-structured stimuli

The stochastic dynamics for location $\boldsymbol{z}_t \in [0, 2\pi)^K$ and velocity $\boldsymbol{v}_t \in \mathbb{R}^K$ are given by

$$\mathrm{d}\boldsymbol{v}_t = -\boldsymbol{v}_t/\tau\,\mathrm{d}t + \boldsymbol{L}\,\mathrm{d}\boldsymbol{W}_t \tag{1}$$

$$\mathrm{d}\boldsymbol{z}_t = \boldsymbol{v}_t\,\mathrm{d}t\ , \tag{2}$$

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

with friction time constant $\tau \in \mathbb{R}^+$ and M independent Wiener processes in vector $\boldsymbol{W}_t$. The values of τ and $\boldsymbol{\Lambda}$ used in each experiment are provided in *Appendix*. For simulations, we employed Euler–Maruyama integration to advance the dynamics, and we map $\boldsymbol{z}_t \mapsto (\boldsymbol{z}_t \bmod 2\pi)$ after each step to keep locations on the circle. The stationary distributions under above dynamics are (see [13]):

$$\lim_{t\to\infty} p(\boldsymbol{v}) = \mathcal{N}(\boldsymbol{v}; \boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{v}}), \quad \lim_{t\to\infty} p(\boldsymbol{z}) = \prod_{k=1}^{K} \text{Uniform}(z_k; 0, 2\pi), \tag{3}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{v}} = \frac{\tau}{2}\boldsymbol{L}\boldsymbol{L}^{\mathsf{T}}$, and the expression for $p(\boldsymbol{z})$ assumes that each object has a non-vanishing independent motion component. Knowing the stationary velocity covariance $\boldsymbol{\Sigma}_{\boldsymbol{v}}$ in closed form gives convenient control over the stimulus. For example, the marginal velocity distribution of the k-th object is $\mathcal{N}(v_k; 0, \sigma^2_{v_k})$ with $\sigma^2_{v_k} = \sum_m \lambda^2_m \cdot C^2_{km}$ which is simply the sum of all parent motion sources' squared strengths. In all experiments, we initialize the stimuli from a sample of their stationary distribution. Then, velocities follow a multivariate Gaussian at any time, and locations become asymptotically independent.

Experiment details: MOT task

Twenty Harvard University undergraduates (Mean age=20.6 years, SD=2.01; 14 females) with normal or corrected-to-normal vision participated in the experiment for course credit. All participants provided informed consent at the beginning of the experiment. The experiment lasted approximately 2 hours and was comprised of a thresholding phase and a testing phase. Across both phases, participants performed a multiple object tracking (MOT) task, in which they were presented with a 2000 ms static display wherein 7 colored dots encircled by white outlines appeared around a larger black ring. All dots proceeded to move around the ring for 5000 ms, following predetermined trajectories. Stimulus motion trajectories were pre-computed using custom Python code and initial values for $\boldsymbol{v}$ and $\boldsymbol{z}$ were drawn from the stimulus motion structure's stationary distributions (3). When pre-computing the trials, we asserted that the final dot locations were non-overlapping to prevent selection ambiguity (trials with overlap were regenerated until they met this criterion). Red square outlines subsequently appeared for 3000 ms around 3 of the moving dots, marking them as to-be-tracked targets. The target cues and disc colors then faded to black during the following 700 ms, after which the white outlines continued to move for 6000 ms until coming to a stop. A random number (1-7) was subsequently superimposed on each dot, prompting participants to report the identities of the perceived targets by making the appropriate keyboard presses. Feedback was provided after every trial. The MOT tasks implemented across both phases were nearly identical, with the following exceptions. Whereas movement trajectories in the thresholding phase followed independent (IND) motion, those implemented in the test phase varied across blocks (Latin-square counterbalanced) based on each motion structure condition (IND, GLO, CNT, CDH1, CDH2). The thresholding phase was used to titrate motion speed by adjusting a per-participant speed factor f_{speed} (see eq. (2) in *Appendix*), such that all participants would achieve a common baseline on IND trials. To this end, participants performed 30 IND trials at an initial speed of $f_{\text{speed}} = 2.0$. Whether performance at this initial speed fell above or below the targeted 2.15 performance threshold, participants completed a subsequent 30 trials of the same task at a faster or slower speed (speed factors varied by 0.25 increments). This staircase thresholding procedure was repeated until each participant's average target identification accuracy approximated 2.15 items per trial (Mean repetitions=2.40, SD=0.68). Once thresholding was complete, participants performed 150 trials of the MOT task at their determined speed. Prior to the onset of each motion structure block, participants were presented with three example trials inherent to the forthcoming block of displays, and were provided with a motion graph that explicitly laid out the motion structure based on the colors of the moving items. Example trials and details of experiment conduction are provided in *Appendix*. No data was excluded from the analysis. The study was approved by the Harvard Institutional Review Board (IRB00000109).

Bayesian observer model: Kalman filtering (both experiments)

We assume independent Gaussian observation noise on the locations of each video frame, $p(\boldsymbol{x}_t \,|\, \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{z}_t, \sigma^2_{\text{obs}}\boldsymbol{I})$. Together with the latent stochastic linear dynamics of (1) + (2), Kalman filters are the statistically ideal observer models for the stimulus. The circular support of $\boldsymbol{z}$, however, renders the normal

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

distribution underlying Kalman filtering into an approximation. To maintain a close approximation to the correct posterior, we ensured that variances of the location posterior $p(\boldsymbol{z}_t \,|\, \boldsymbol{x}_{1:t})$ remained small at all time. For all observable dots, location estimates are much smaller than π anyway; for the unobserved dots in the prediction task, we designed the stimuli such that the correct posterior's standard deviation never exceeded 40°. Thus, most of the probability mass stayed within the (unwrapped) circle at any time, and errors introduced by the Gaussianity assumption are small. The corresponding matrices of the motion-structured Kalman filter are provided in *Appendix*.

Bayesian observer model: MOT task

Above Kalman filter was used for tracking, assuming either the correct motion prior $\boldsymbol{L}^\star$ to underlie the stimulus, or a prior of independent motion with identical marginal velocity distributions $\mathcal{N}\left(v_k; 0, \sigma^2_{v_k}\right)$. Following [16], different possible mental dot assignments γ are scored at each frame by comparing the likelihoods of the current observation $\boldsymbol{x}_t$ under the possible assignments. As in [16], this poses a numerically intractable inference process (per frame, there are $K!$ possible assignments; the total number of assignments per trial thus grows as $K!^{\#\text{frames}}$). We use discrete particle variational inference [37] for approximate inference, with a single particle and the set of all pairwise dot permutations as differential proposals per frame. After each frame, the highest scoring candidate assignment γ is maintained. The assignment γ at the end of the trial determined which three input dots were chosen as targets. In the numerical evaluation, the Kalman filter was presented with noisy observations, $\boldsymbol{x}_t \sim \mathcal{N}\left(\boldsymbol{z}_t, \sigma^2_{\text{obs}}\boldsymbol{I}\right) \bmod 2\pi$, with $\sigma_{\text{obs}} = 0.05$. The value of σ_{obs} was determined via computer simulations to yield on average the target performance (2.15) in the IND condition. For **Fig. 2B**, each trial presented to a human participant was simulated 25 times (with changing noise instantiations), and simulated performance was averaged over all repetitions to reduce the variance of the observer model's predictions (shrinking the error bars in the center and right panel).

Experiment details: prediction task

Twelve adult individuals (Mean age=31 years, SD=9 years; 10 males), participated in exchange for financial compensation ($10 per hour plus a performance-dependent bonus). All participants reported normal or corrected-to-normal vision, and provided informed consent at the beginning of the study. On average, the experiment lasted approximately 80 minutes. Each participant completed 100 trials per motion structure (GLO, CLU, CDH), presented in blocks (50 unique trials per block, each presented twice in randomized order). Block orders were balanced across participants. Participants were informed about the stimulus condition and performed a variable number of training trials prior to each block until they decided to start the experiment block. Dots were color coded as shown in **Fig. 3B**. Trials were composed as follows. Initial values for $\boldsymbol{v}$ and $\boldsymbol{z}$ were drawn from the stimulus motion structure's stationary distributions (3). After a 1000 ms still period, all dots started moving stochastically for 5000 ms according to (1) + (2). During the end of the 5000 ms period, the red and green target dots faded out, and only the remaining five dots were visible for another 1500 ms period after which the scene froze. The green and red dot's locations had to be predicted by directing a green / red mouse cursor to the predicted location on the circle. After each trial, the true dot locations were revealed and participants received points (0 – 20) based on the accuracy of their prediction. The points only served as task engagement and for payment, and played no role for the analysis. Example trials and details of experiment conduction are provided in *Appendix*. No data was excluded from the analysis. The experiment was approved by the Harvard Institutional Review Board (IRB15-2048).

Bayesian observer model: prediction task

Kalman filters with different candidate motion priors $\boldsymbol{L}$ were presented with the same trials that had been shown to human participants. For the observer model, we assume correct dot assignment ($\gamma = \boldsymbol{I}$) and set $\sigma_{\text{obs}} = 0$ for observed dots since dot confusion and observation noise–induced errors are expected to play a negligible role in the prediction task. While invisible, we set $\sigma_{\text{obs}} \to \infty$ for the green and red dots. For candidate prior $\boldsymbol{L}$, human responses $\boldsymbol{\varphi}^{(i)} = (\varphi^{(i)}_{\text{green}}, \varphi^{(i)}_{\text{red}})$ in trial $i = 1,..,100$ are then modeled as

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

$$p(\boldsymbol{\varphi}^{(i)} \mid \boldsymbol{L}, \theta) = (1-\pi_{\mathrm{s}}) \cdot \mathcal{N}\left(\boldsymbol{\varphi}^{(i)}; \boldsymbol{\mu}_{\mathrm{kal}}^{(i)}, a\,\boldsymbol{\Sigma}_{\mathrm{kal}}^{(i)} + b\,\boldsymbol{I}\right) + \pi_{\mathrm{s}} \cdot \mathcal{N}\left(\boldsymbol{S}\boldsymbol{\varphi}^{(i)}; \boldsymbol{\mu}_{\mathrm{kal}}^{(i)}, a\,\boldsymbol{\Sigma}_{\mathrm{kal}}^{(i)} + b\,\boldsymbol{I}\right) \tag{4}$$

where $\boldsymbol{\mu}_{\mathrm{kal}}^{(i)}$ and $\boldsymbol{\Sigma}_{\mathrm{kal}}^{(i)}$ are the (2-dim.) mean and (2×2) covariance matrix of the Kalman filter with motion prior $\boldsymbol{L}$ for the green and red dot at the end of the trial. The matrix $\boldsymbol{S} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ models dot swapping in the response with probability π_{s}, and $a, b \in \mathbb{R}^+$ scale structured inference noise and unstructured motor noise respectively. Eq. 4 describes a *stochastic posterior* observer model (the winning model identified in the systematic comparison of [21]) with lapses and motor noise. The three free parameters $\theta = (\pi_{\mathrm{s}}, a, b)$ were fitted via maximum likelihood for each candidate prior $\boldsymbol{L}$, stimulus condition $\boldsymbol{L}^\star$ and participant. The log-likelihood ratios to the model with correct motion prior,

$$\mathrm{LLR}_{\boldsymbol{L}^\star, \boldsymbol{L}} = \sum_{\text{Trials } i} \log \frac{p\left(\boldsymbol{\varphi}^{(i)} \mid \boldsymbol{L}\ , \theta_{\boldsymbol{L}}^{\mathrm{ML}}\right)}{p\left(\boldsymbol{\varphi}^{(i)} \mid \boldsymbol{L}^\star, \theta_{\boldsymbol{L}^\star}^{\mathrm{ML}}\right)} \tag{5}$$

are plotted in **Fig. 3C**. One-tailed paired t-tests identify if the correct prior $\boldsymbol{L}^\star$ explained human responses better than alternative priors $\boldsymbol{L}$.

Distinguishing systematic and stochastic errors in the prediction task

We evaluated the noise factor f_{noise} separately for the green and red dot per stimulus condition and participant, that is, each noise factor is based on the repetitions $r = 1,2$ of the unique trials $j = 1,..,50$ in one stimulus block. We define the average noise variance

$$\overline{\sigma}_{\mathrm{noise}} = \left\langle \mathrm{Var}\left[\mathrm{d}(\varphi_j^{(r)}, \overline{\varphi}_j)\right]_r \right\rangle_j \tag{6}$$

where $\langle \cdot \rangle_j$ denotes the average over unique trials, $\mathrm{Var}\left[\cdot\right]_r$ is the unbiased estimator of the variance over trial repetitions, $\mathrm{d}(\cdot,\cdot) \in (-\pi, \pi]$ is the (signed) distance on the circle, and $\overline{\varphi}_j$ is the circular mean of the two responses $\varphi_j^{(1)}$ and $\varphi_j^{(2)}$. We further define the total prediction variance

$$\overline{\sigma}_{\mathrm{tot}} = \left\langle \mathrm{d}^2(\varphi^{(i)}, \varphi_{\mathrm{opt}}^{(i)}) \right\rangle_i \tag{7}$$

as the mean squared error made compared to the optimal prediction $\varphi_{\mathrm{opt}}^{(i)} = \boldsymbol{\mu}_{\mathrm{kal}}^{(i)}(\boldsymbol{L}^\star)$. The estimators $\overline{\sigma}_{\mathrm{noise}}$ and $\overline{\sigma}_{\mathrm{tot}} - \overline{\sigma}_{\mathrm{noise}}$ were verified to recover the stochastic and systematic error components of synthetic data (not shown). The noise factor is defined as the ratio $f_{\mathrm{noise}} = \overline{\sigma}_{\mathrm{noise}} / \overline{\sigma}_{\mathrm{tot}}$. Based on synthetic data, the standard deviation of the noise factors in **Fig. 4B** is estimated to be 0.10.

Data and code availability

Computer simulations and data analysis were performed with custom Python code which is available on `https://github.com/DrugowitschLab/motion-structure-used-in-perception`. In this repository we also provide the raw data collected in the experiments.

Acknowledgments

The authors are very thankful to Rick Born, Till Hartmann, Luke Rast, Ariana Sherdil and Valentin Wyart for helpful feedback and discussions, as well as to George Alvarez for allowing us to conduct the MOT experiment in his laboratory. This research was supported by a seed grant from the Harvard Brain Initiative, an Alice and Joseph Brooks Fund Fellowship award (JB), a research fellowship from the Alfred P. Sloan Foundation (SJG), and a James S. McDonnell Foundation scholar award (Grant #220020462, JD). Parts of this research were conducted using the O2 High Performance Compute Cluster at Harvard Medical School.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

[1] D. Kaiser, G. L. Quek, R. M. Cichy, and M. V. Peelen, "Object vision in a structured world", *Trends in cognitive sciences*, 2019.

[2] M. Wertheimer, "Untersuchungen zur lehre von der gestalt", *Psychological Research*, vol. 4, pp. 301–350, 1923.

[3] S. Yantis, "Multielement visual tracking: Attention and perceptual organization", *Cognitive Psychology*, vol. 24, no. 3, pp. 295–340, 1992.

[4] G. Liu, E. L. Austen, K. S. Booth, B. D. Fisher, R. Argue, M. I. Rempel, and J. T. Enns, "Multiple-object tracking is based on scene, not retinal, coordinates.", *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, no. 2, p. 235, 2005.

[5] M. Suganuma and K. Yokosawa, "Grouping and trajectory storage in multiple object tracking: Impairments due to common item motions", *Perception*, vol. 35, no. 4, pp. 483–495, 2006.

[6] C. Wang, X. Zhang, Y. Li, and C. Lyu, "Additivity of feature-based and symmetry-based grouping effects in multiple object tracking", *Frontiers in psychology*, vol. 7, p. 657, 2016.

[7] J. Driver, P. McLeod, and Z. Dienes, "Motion coherence and conjunction search: Implications for guided search theory", *Perception & Psychophysics*, vol. 51, no. 1, pp. 79–85, 1992.

[8] R. Luria and E. K. Vogel, "Come together, right now: Dynamic overwriting of an object's history through common fate", *Journal of Cognitive Neuroscience*, vol. 26, no. 8, pp. 1819–1828, 2014.

[9] G. Johansson, *Configurations in event perception: an experimental study*. Almqvist & Wiksell, 1950.

[10] S. J. Gershman, J. B. Tenenbaum, and F. Jäkel, "Discovering hierarchical motion structure", *Vision research*, vol. 126, pp. 232–241, 2016.

[11] H. Xu, N. Tang, J. Zhou, M. Shen, and T. Gao, "Seeing "what" through "why": Evidence from probing the causal structure of hierarchical motion.", *Journal of Experimental Psychology: General*, vol. 146, no. 6, p. 896, 2017.

[12] Z. W. Pylyshyn and R. W. Storm, "Tracking multiple independent targets: Evidence for a parallel tracking mechanism", *Spatial vision*, vol. 3, no. 3, pp. 179–197, 1988.

[13] C. Gardiner, *Stochastic methods*. Springer Berlin, 2009, vol. 4.

[14] B. J. Scholl, "What have we learned about attention from multiple object tracking (and vice versa)", *Computation, cognition, and Pylyshyn*, pp. 49–78, 2009.

[15] S. Franconeri, S. Jonathan, and J. Scimeca, "Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity", *Psychological science*, vol. 21, no. 7, pp. 920–925, 2010.

[16] E. Vul, G. Alvarez, J. B. Tenenbaum, and M. J. Black, "Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model", in *Advances in neural information processing systems*, 2009, pp. 1955–1963.

[17] A. Thiele and G. Stoner, "Neuronal synchrony does not correlate with motion coherence in cortical area mt", *Nature*, vol. 421, no. 6921, p. 366, 2003.

[18] G. Y. Bae and J. I. Flombaum, "Close encounters of the distracting kind: Identifying the cause of visual tracking errors", *Attention, Perception, & Psychophysics*, vol. 74, no. 4, pp. 703–715, 2012.

[19] J. H. Steiger, "Tests for comparing elements of a correlation matrix.", *Psychological bulletin*, vol. 87, no. 2, p. 245, 1980.

[20] I. Lee and K. Preacher, *Calculation for the test of the difference between two dependent correlations with one variable in common [computer software]*, 2013.

[21] L. Acerbi, S. Vijayakumar, and D. M. Wolpert, "On the origins of suboptimality in human probabilistic inference", *PLoS computational biology*, vol. 10, no. 6, e1003661, 2014.

[22] J. Drugowitsch, V. Wyart, A.-D. Devauchelle, and E. Koechlin, "Computational precision of mental inference as critical source of human choice suboptimality", *Neuron*, vol. 92, no. 6, pp. 1398–1411, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[23] Y. Weiss, E. P. Simoncelli, and E. H. Adelson, "Motion illusions as optimal percepts", *Nature neuroscience*, vol. 5, no. 6, p. 598, 2002.

[24] A. A. Stocker and E. P. Simoncelli, "Sensory adaptation within a bayesian framework for perception", in *Advances in neural information processing systems*, 2006, pp. 1289–1296.

[25] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception", *Nature neuroscience*, vol. 9, no. 4, p. 578, 2006.

[26] J. H. Hedges, A. A. Stocker, and E. P. Simoncelli, "Optimal inference explains the perceptual coherence of visual motion stimuli", *Journal of vision*, vol. 11, no. 6, pp. 14–14, 2011.

[27] N. Srivastava and E. Vul, "Attention modulates spatial precision in multiple-object tracking", *Topics in cognitive science*, vol. 8, no. 1, pp. 335–348, 2016.

[28] A. M. Saxe, J. L. McClelland, and S. Ganguli, "A mathematical theory of semantic development in deep neural networks", *Proceedings of the National Academy of Sciences*, 2019. DOI: 10.1073/pnas.1820226116.

[29] A. Coenen, E. Reif, A. Yuan, B. Kim, A. Pearce, F. Viégas, and M. Wattenberg, "Visualizing and measuring the geometry of bert", *arXiv preprint arXiv:1906.02715*, 2019.

[30] R. T. Born and D. C. Bradley, "Structure and function of visual area mt", *Annu. Rev. Neurosci.*, vol. 28, pp. 157–189, 2005.

[31] M. S. Graziano, R. A. Andersen, and R. J. Snowden, "Tuning of mst neurons to spiral motions", *Journal of Neuroscience*, vol. 14, no. 1, pp. 54–67, 1994.

[32] P. J. Mineault, F. A. Khawaja, D. A. Butts, and C. C. Pack, "Hierarchical processing of complex motion along the primate dorsal visual pathway", *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, E972–E980, 2012.

[33] J. M. Beck, P. E. Latham, and A. Pouget, "Marginalization in neural circuits with divisive normalization", *Journal of Neuroscience*, vol. 31, no. 43, pp. 15 310–15 319, 2011.

[34] O. W. Layton and N. A. Browning, "A unified model of heading and path perception in primate mstd", *PLoS computational biology*, vol. 10, no. 2, e1003476, 2014.

[35] O. W. Layton and B. R. Fajen, "A neural model of mst and mt explains perceived object motion during self-motion", *Journal of Neuroscience*, vol. 36, no. 31, pp. 8093–8102, 2016.

[36] J. Bill, H. Pailian, S. J. Gershman, and J. Drugowitsch, *Do humans employ hierarchical motion structure during visual perception? available at: Osf.io/v8254*, Mar. 2019. [Online]. Available: osf.io/f6gek.

[37] A. Saeedi, T. D. Kulkarni, V. K. Mansinghka, and S. J. Gershman, "Variational particle approximations", *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2328–2356, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Appendix

Details on the stimulus generation and Kalman filters

Equations of motion with volatility and speed factors

For transparent stimulus control, it is convenient to write the stochastic stimulus dynamics as

$$\mathrm{d}\boldsymbol{z}_t = \boldsymbol{v}_t\,\mathrm{d}t\ , \tag{1}$$

$$\mathrm{d}\boldsymbol{v}_t = -\boldsymbol{v}_t/(\tau/\alpha)\,\mathrm{d}t + \sqrt{\alpha}\,f_{\text{speed}}\,\boldsymbol{L}\,\mathrm{d}\boldsymbol{W}_t \tag{2}$$

with *speed factor* f_{speed} and *volatility factor* α. The speed factor scales the absolute value of velocity in the stationary distribution while maintaining the time scale τ of typical changes in the stimulus. The volatility factor, in contrast, adjusts the time scale τ, while keeping the stationary velocity distribution unchanged, because:

$$\lim_{t\to\infty} p(\boldsymbol{v}) = \mathcal{N}\left(\boldsymbol{v}; \boldsymbol{0}, \frac{(\tau/\alpha)}{2}\,\alpha\,f_{\text{speed}}^2\,\boldsymbol{L}\,\boldsymbol{L}^\mathsf{T}\right) = f_{\text{speed}}\,\mathcal{N}\left(\boldsymbol{v}; \boldsymbol{0}, \frac{\tau}{2}\boldsymbol{L}\,\boldsymbol{L}^\mathsf{T}\right)\ . \tag{3}$$

In computer implementations, f_{speed} and α can be absorbed in τ and λ via $\tau \mapsto \tau/\alpha$ and $\lambda \mapsto \sqrt{\alpha}\,f_{\text{speed}}\,\lambda$. For the MOT task, we adjusted f_{speed} in the IND condition in increments of 0.25 until participants performed approximately at the desired target (2.15). After this adjustment, participants covered speed factors between 1.75 and 2.75. All subsequent experiment blocks then used the per-participant speed factor.

Parameters of the motion structures used in the experiments

MOT task. In the MOT task, we set $\tau = 64$s, leading to slowly changing velocities. For $f_{\text{speed}} = 1$, the motion strengths in the stimulus conditions were as follows. For independent (IND) motion, $\lambda_{\text{ind}} = 0.125$ for all seven individual sources. For global (GLO) motion and counter-rotating (CNT) motion, $\lambda_{\text{lat}} = 0.1244$ for the latent source and $\lambda_{\text{ind}} = 0.0125$ for the seven individual sources. For counter-rotating deep hierarchical (CDH) motion, $\lambda_{\text{glo}} = 0.1016$ for the global motion source, $\lambda_{\text{cnt}} = 0.0718$ for the counter-rotating sub-source, $\lambda_{\text{ind}} = 0.0125$ for the individual motion sources of the six dots in the clusters, and $\lambda_{\text{mav}} = 0.0729$ for the seventh dot (the Maverick dot). Note that motion strengths add up quadratically in the marginal dot velocities due to the factor $\boldsymbol{L}\,\boldsymbol{L}^\mathsf{T} = \boldsymbol{C}\cdot\mathrm{diag}(\lambda_1^2..,\lambda_M^2)\cdot\boldsymbol{C}^\mathsf{T}$. At the example of GLO and IND, we have $\sqrt{0.1244^2 + 0.0125^2} = 0.125$ such that the marginal velocities are identical across motion structures. The $\boldsymbol{L}$-matrices of all conditions are provided in **Fig. 5**.

Prediction task. In the prediction task, we set $\tau = 1$s, leading to volatile stimuli. The values of λ for the various motion sources are provided in **Fig. 3**. The $\boldsymbol{L}$-matrices of all conditions and observers are provided in **Fig. 6**.

Discrete time integration and Kalman filter matrices

For numerical integration, we concatenate $(\boldsymbol{z}_t, \boldsymbol{v}_t)$ to a single state vector and advance (1)+(2) of the main paper in small time steps δt:

$$\begin{pmatrix}\boldsymbol{z}_{t+\delta t}\\ \boldsymbol{v}_{t+\delta t}\end{pmatrix} = \underbrace{\begin{pmatrix}\boldsymbol{I}^{(K)} & dt\,\boldsymbol{I}^{(K)}\\ \boldsymbol{0}^{(K)} & (1-\frac{\delta t}{\tau})\,\boldsymbol{I}^{(K)}\end{pmatrix}}_{=:\boldsymbol{F}(\boldsymbol{L})}\begin{pmatrix}\boldsymbol{z}_t\\ \boldsymbol{v}_t\end{pmatrix} + \underbrace{\begin{pmatrix}\boldsymbol{0}^{(K)} & \boldsymbol{0}^{(K\times M)}\\ \boldsymbol{0}^{(K)} & \sqrt{\delta t}\,\boldsymbol{L}\end{pmatrix}}_{=:\sqrt{\boldsymbol{Q}(\boldsymbol{L})}}\mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I}) \tag{4}$$

with $\boldsymbol{\epsilon} = (\epsilon_1,..,\epsilon_{K+M})$ being i.i.d. standard Gaussian noise. The matrices $\boldsymbol{F}$ and $\boldsymbol{Q} = \sqrt{\boldsymbol{Q}}\cdot\sqrt{\boldsymbol{Q}}^\mathsf{T}$ are known as *state transition model* and *process noise covariance* in Kalman filtering.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Computer simulations

Stimulus videos where presented at 60 Hz (MOT task) and 50 Hz (prediction task), thereby defining δt as the inverse of this value during tracking with Kalman filters. For the stimulus generation, we used a 10-times smaller value of δt to ensure smooth trajectories, and video frames were shown every 10th step. Simulations and data analyses were performed with Python 3.6, Matplotlib 3.0, and Scipy 1.2.

Details to the experiments

MOT task

The MOT tasks used in both thresholding and test phases were displayed using MatLab Psychophysics Toolbox on a 21.5" iMac monitor (viewable area: 10.54 cm x 18.74 cm, refresh rate: 60 frames/sec). Viewing distance was unconstrained, but averaged approximately 60 cm. Across both phases, a central fixation cross (0.70° x 0.70° of visual angle) and black ring (diameter: 19.93° of visual angle, thickness: 0.09° of visual angle) remained constant on the screen throughout each trial. Each of the 7 target and non-target items subtended 1.74° of visual angle in diameter, and were encircled by white outlines that had a thickness of 0.09° of visual angle. Red square outlines (2.80° x 2.80° of visual angle, thickness: 0.26° of visual angle) were used to indicate which of these moving items were to be tracked during the trial. The colors of target and non-target items alike were sampled from three equiluminant colors (blue, pink, yellow), such that three items belonged to one color set, three items belonged to another color set, and a singleton was filled with the remaining color. Participants were informed that the colors of the items were uninformative during the thresholding phase, but that they aided the identification of each dot's role within the motion structures during the test phase. After each trial was complete, participants received feedback in the form of red square outlines (2.80° x 2.80° of visual angle, thickness: 0.26° of visual angle) appearing around the correct targets.

Prediction task

In the prediction task, the location of two temporarily occluded target dots (green and red) had to be predicted at the end of the trial, using the motion of five (always visible) non-target dots. Stimuli and graphical user interface were written in custom Python code (using Matplotlib on the Qt5 backend) and were presented on an iMac16,2 (display resolution: 2048 × 1152; frame rate: 50 frames/second) in an anechoic experiment room at Harvard Medical School. Viewing distance was unconstrained, but averaged approximately 60 cm. In each trial, a gray circle (diameter: ~12° of visual angle) remained constant on the screen throughout the trial. On the circle, seven dots (5 non-targets and 2 targets; diameter: 0.5° and 0.7° of visual angle respectively) moved stochastically according to eq. (4). Dots were colored as indicated in **Fig. 3**: light-blue for dots 1, 2 and 3; light-green for dots 4 and 5; green for dot 6 (a target dot); and red for dot 7 (the other target dot). Participants were informed about the meaning of the colors within the motion structure of the respective block. At the trial end, participants predicted the targets' locations by directing a green / red mouse cursor to the predicted location on the circle. The order (green / red) of the predictions was randomized across trials. After each trial, the true dot locations were revealed and participants received points based on the accuracy of their prediction:

$$\text{points} = \sum_{c \in \{\text{green,red}\}} \text{round}\left(10 \times \left[1 - \frac{\mathrm{d}(\varphi_c^{\text{human}}, \varphi_c^{\text{true}})}{\pi/2}\right]_+\right) \tag{5}$$

where $\mathrm{d}(\cdot,\cdot) \in [0,\pi]$ denotes the (absolute) distance on the circle, and the rectification $[\cdot]_+$ restricts the points to positive values. This is a linear decay, per dot, from 10 points for a perfect prediction down to 0 points for errors of 90° and above. Note that due to the intrinsic stochasticity of the targets' motion, even an ideal observer will usually obtain less than 20 points per trial. Participants received a bonus of \$1 per 500 points (rounded up to the next dollar), resulting in \$7 or \$8 of paid bonus (in addition to the \$10 per hour base pay).

Discussion of the quantitative mismatch in the GLO condition of the MOT task

The Bayesian observer model systematically underestimated human MOT performance in the GLO condition (cf. Fig. 2B in the main paper). In general, it is not surprising that a minimal observer model, that only

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

contains the computational components essential to solving the MOT task, cannot capture all nuances of human tracking behavior. Nonetheless, we asked what is special about the global motion stimulus condition? Most notably, dot proximity events typically last longer for GLO stimuli than for other stimulus structures due to the strong positive velocity correlation. Proximity events that span many video frames (each frame lasts for only 1/60Hz = 17ms) are expected to be particularly susceptible to perceptual noise in our observer model for two reasons. First, the employed DPVI algorithm [1] evaluates (and possibly discards) alternative mental dot assignments on a frame-by-frame basis and, thus, alternative hypotheses may not be maintained throughout a longer proximity event. Humans, in contrast, may reason on longer time scales. Second, the tracking process relies on Kalman filtering (forward inference) rather than smoothing (forward-and-backward inference). Updating past location estimates based on new observations could help to resolve ambiguous mental assignments in the recent past as they occur during proximity events. In addition to these algorithmic reasons, human participants may direct their visual focus and attention specifically to locations of proximity events (proximity events last longer, but occur less frequently in GLO motion than in other conditions) in order to locally reduce perceptual noise [2].

To test for a specific detrimental effect of perceptual noise to tracking of GLO stimuli, we re-evaluated the MOT data by decoupling the noise σ_{obs} assumed by the observer model and the noise σ_x actually corrupting observations $\boldsymbol{x}_t$ in computer simulations (in Fig. 2B of the main paper, we had $\sigma_{\text{obs}} = \sigma_x = 0.05$). As an ad hoc-choice, we assumed that $\sigma_{\text{obs}} = 10\sigma_x$, such that corrupted observations in individual frames have a smaller weight in the Kalman update. Readjustment of the noise level to approximately reach the desired average target performance (2.15) in the IND condition yielded $\sigma_{\text{obs}} = 0.7$ and $\sigma_x = 0.07$. The resulting predictions of the observer model, using the decoupled noise for all stimulus conditions, are shown **Fig. 7** (additional analysis; only one trial repetition simulated). The resulting predictions match human behaviour not only qualitatively, but also quantitatively better than the minimal observer model of the main paper ($\rho = 0.57$ for the correct motion prior in the evaluation of the left panel of Fig. 2D). Yet, even with decoupled perceptual noise, motion structured priors are required to explain human performance ($\rho = 0.02$ for the IND prior in the evaluation of the right panel of Fig. 2D). We emphasize that this additional analysis is only a first step towards understanding the detailed contribution of perceptual noise to multiple object tracking of motion structured stimuli, and additional experiments would be needed to appropriately address this question.

References

[1] A. Saeedi, T. D. Kulkarni, V. K. Mansinghka, and S. J. Gershman, "Variational particle approximations", *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 2328–2356, 2017.

[2] N. Srivastava and E. Vul, "Attention modulates spatial precision in multiple-object tracking", *Topics in cognitive science*, vol. 8, no. 1, pp. 335–348, 2016.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

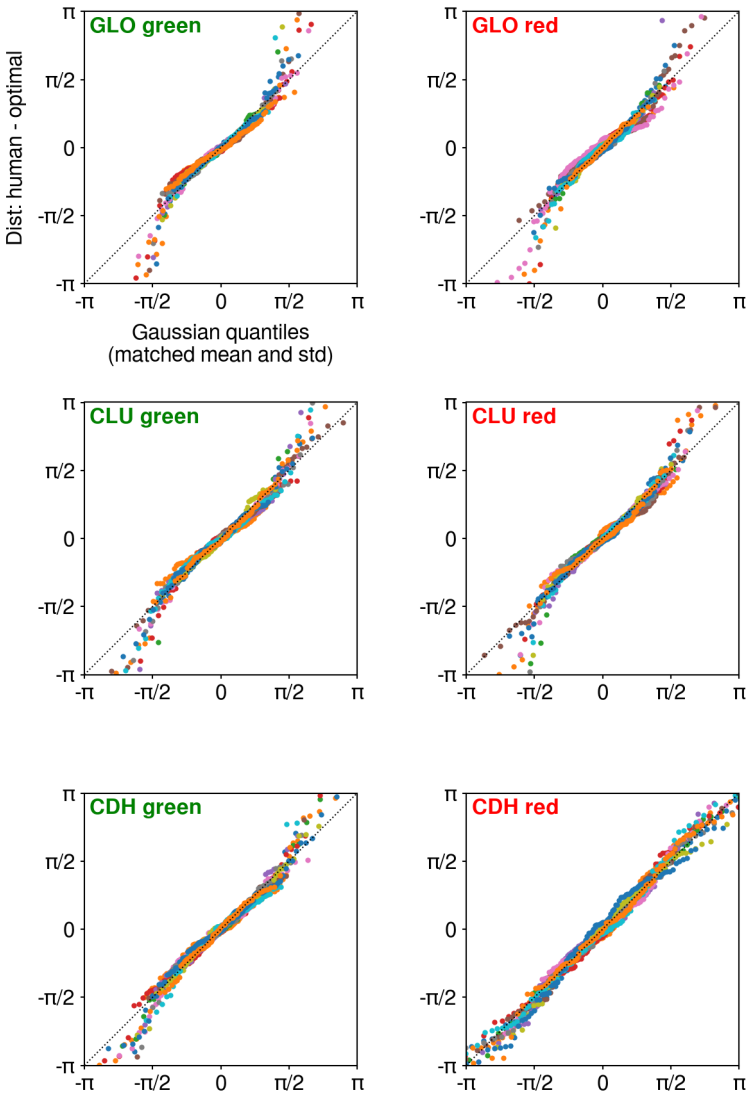

Figure 1. Human prediction errors are normally distributed in the prediction task. Each panel shows a probability plot of prediction errors in a stimulus condition (GLO, CLU, CDH) for the green or red dot (one point per trial; colors=participants). Prediction errors are the circular distance to the optimal prediction which is given by the mean μ_{kal} of a Kalman filter with correct motion structure prior L. Prediction errors (y-axis) are compared against a normal distribution with same mean and variance (x-axis; matched per participant). Most of the responses are along the identity (dotted line) as expected for a normal distribution. The small number of deviations at the tails can be attributed to dot confusion in some trials: swapping the red and green dot leads to uniformly distributed prediction errors, i.e., "heavy tails".

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

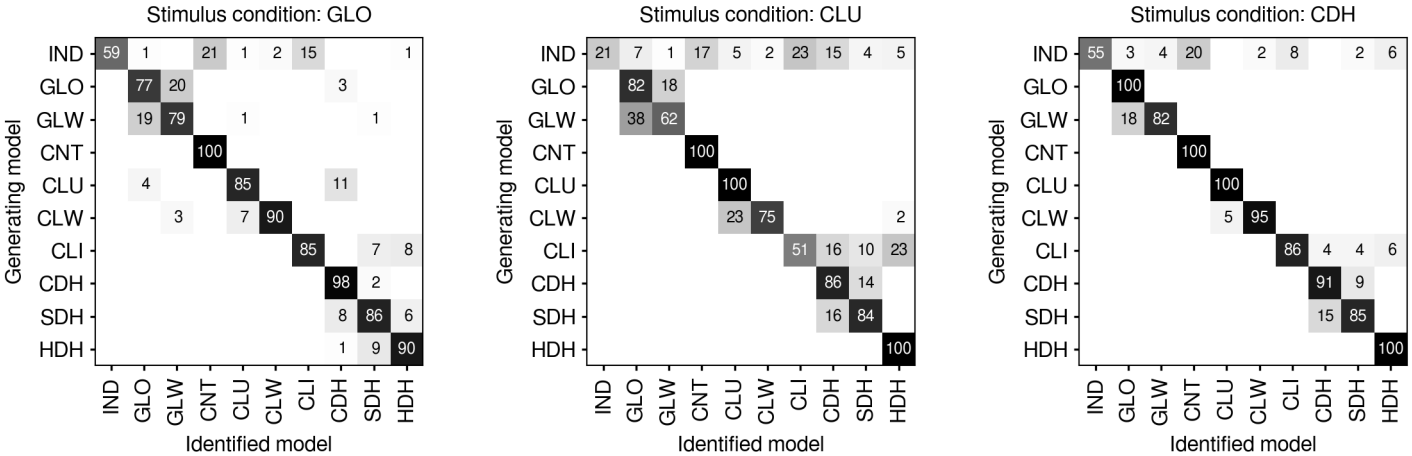

Figure 2. Fitting the Bayesian observer model can re-identify the generating motion prior in the prediction task. Per stimulus condition (GLO: left, CLU: center, CDH: right), we generated artificial responses to all 1200 (= 12×100) trials presented to human participants by drawing samples $(\varphi_{\text{green}}, \varphi_{\text{red}})$ from the observer model presented in *Material and Methods* for each of 10 candidate motion priors L. The 10 priors are the 7 priors from the main paper and 3 additional priors (GLW: weak global, CLW: weak cluster, HDH: half deep hierarchy, see Fig. S3 for their definition). To create "human-like" responses, we used the average of the max. likelihood parameters $\theta = (\pi_{\text{s}}, a, b)$ fitted to human participants under the correct prior $L^{\star}$. Then we fitted all 10 candidate priors to the responses generated from the 10 priors as decribed in *Material and Methods*. This process (response generation and fitting) was repeated 100 times, i.e., in total 3×1200×10×100 = 3,600,000 artificial trials were evaluated. The figure shows the number of times, prior "Identified model" had the highest log-likelihood ratio in explaining the responses generated from prior "Generating model".

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

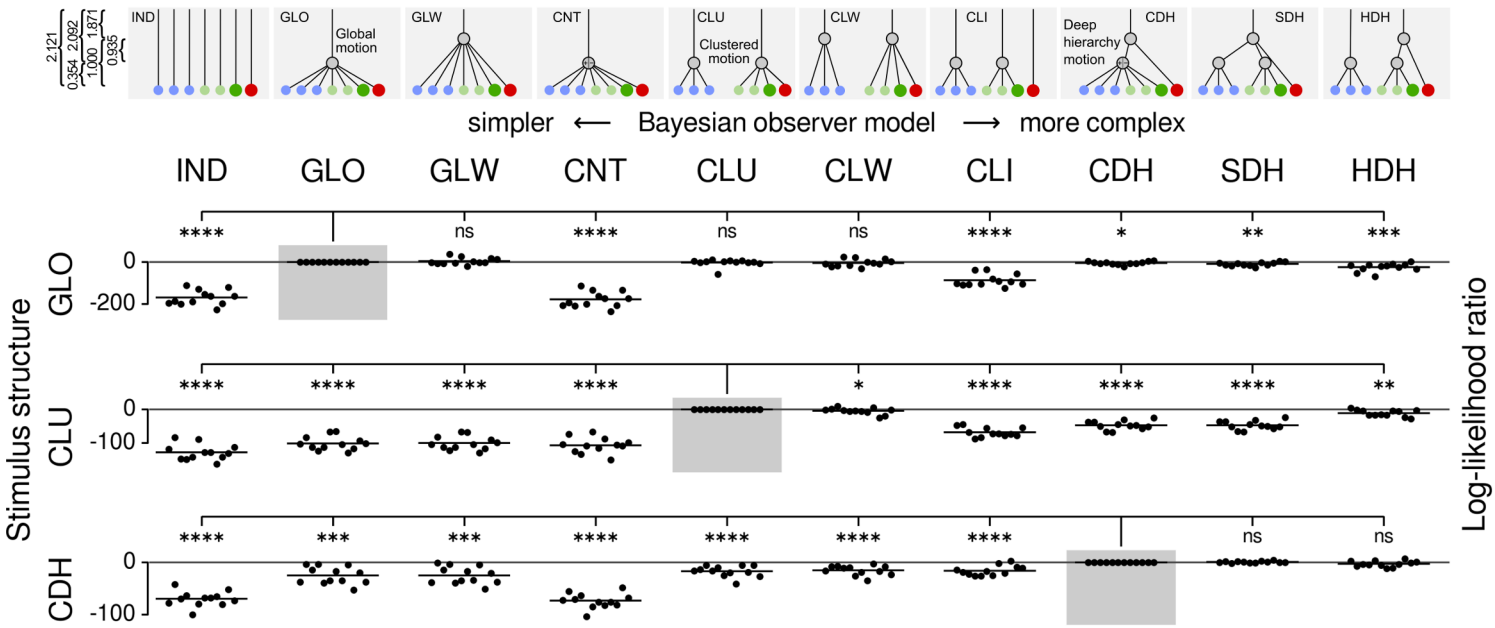

Figure 3. Revealing human motion priors in a multiple object prediction task (more candidate models). Same as Fig. 3 from the main paper, but with 3 additional (non-preregistered) motion structure priors, GLW: weak global, CLW: weak cluster, HDH: half deep hierarchy. These were included as additional, potential strategies to solve the task in the CDH condition. For instance, GLW mirrors the strength of the global motion component in CDH while disregarding the counter-rotating substructure. While GLW and CLW cannot explain human responses in the CDH condition, HDH – another nested structure with 3 latent sources – could explain human responses as well as the correct CDH prior. Due to the similarity of GLO and GLW (or CLU and CLW), these priors could not be distinguished reliably in the GLO (CLU) condition (cf. Fig. S2). Motion strengths λ (vertical extend of edges in the graphs) are not drawn to scale. Numerical values of all motion strengths λ are given in the top-left corner of the figure. For our stimulus design, motion strengths add quadratically, granting control over marginal dot velocities. At the example of GLO motion, $\sqrt{0.354^2 + 2.092^2} = 2.121$ matches the strength of IND.

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

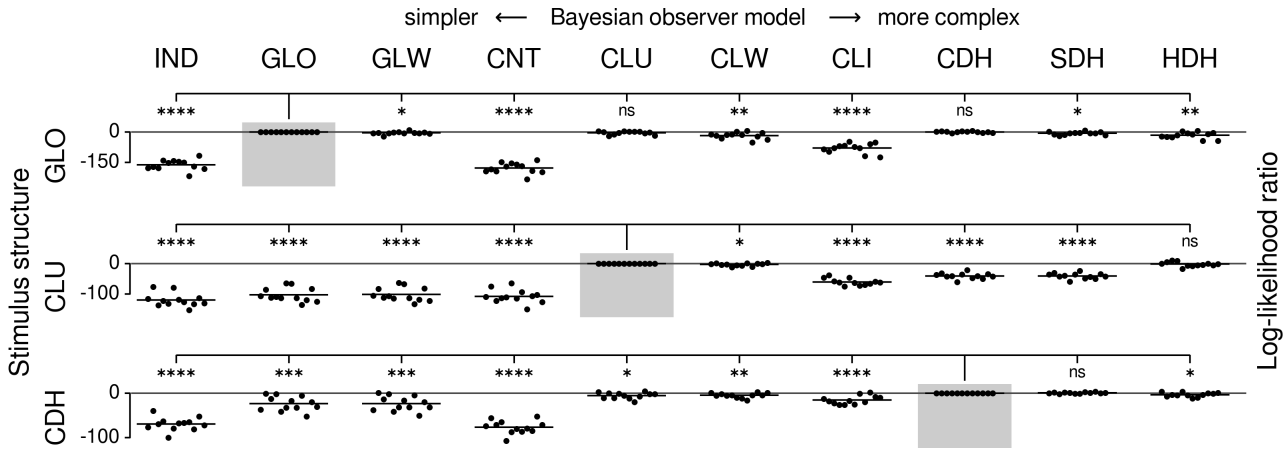

Figure 4. Benefits of multiple object prediction over single object prediction. Same as Fig. 3 from the main paper (and Fig. S3), but with an observer model that disregards the posterior covariance Σ_{kal} by modeling human responses as

$$p(\boldsymbol{\varphi}^{(i)} \mid \boldsymbol{L}, \theta) = (1-\pi_{\text{s}}) \cdot \mathcal{N}\left(\boldsymbol{\varphi}^{(i)}; \boldsymbol{\mu}^{(i)}_{\text{kal}}, a\begin{pmatrix}1&0\\0&0\end{pmatrix} + b\begin{pmatrix}0&0\\0&1\end{pmatrix}\right) + \pi_{\text{s}} \cdot \mathcal{N}\left(\boldsymbol{S}\boldsymbol{\varphi}^{(i)}; \boldsymbol{\mu}^{(i)}_{\text{kal}}, a\begin{pmatrix}1&0\\0&0\end{pmatrix} + b\begin{pmatrix}0&0\\0&1\end{pmatrix}\right),$$

i.e., replacing the structured covariance by unstructured noise for both dots (note that the model still has 3 free parameters). This evaluation resembles the type of data that could be measured in a single object prediction protocol where the green and red dots' locations are predicted in separate trials. Compared to our experiment, it would thus require twice the number of trials. Despite this, a general reduction of contrast in the log-likelihood ratio is observed, when compared to multiple object prediction (Fig. S3). This reduction is especially apparent in the CDH stimulus condition where differentiation of the CLU tracker is reduced to the edge of statistical significance.

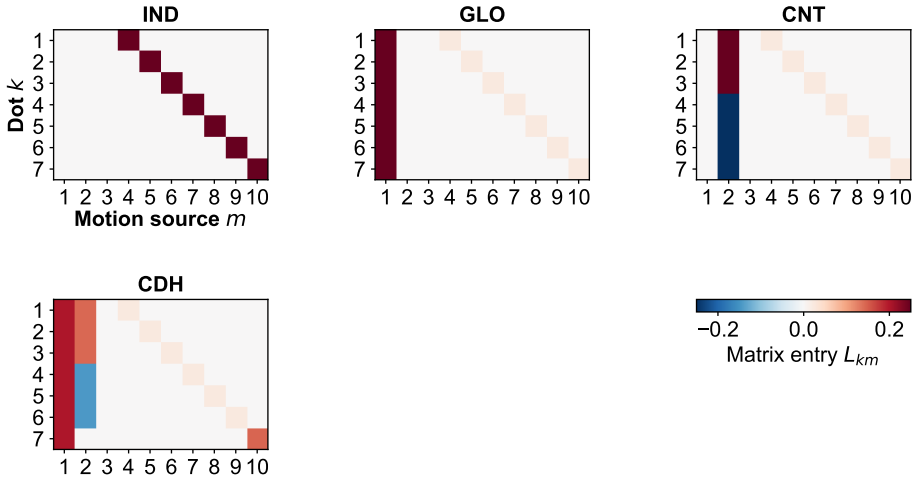

Figure 5. Motion structure matrices of the MOT task. All matrices L lead to the same marginal velocity distributions $p(v_k)$ for all dots k. In this plot, we set $f_{\text{speed}} = 2.0$ (a typical value for human participants).

bioRxiv preprint doi: https://doi.org/10.1101/758573; this version posted September 5, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

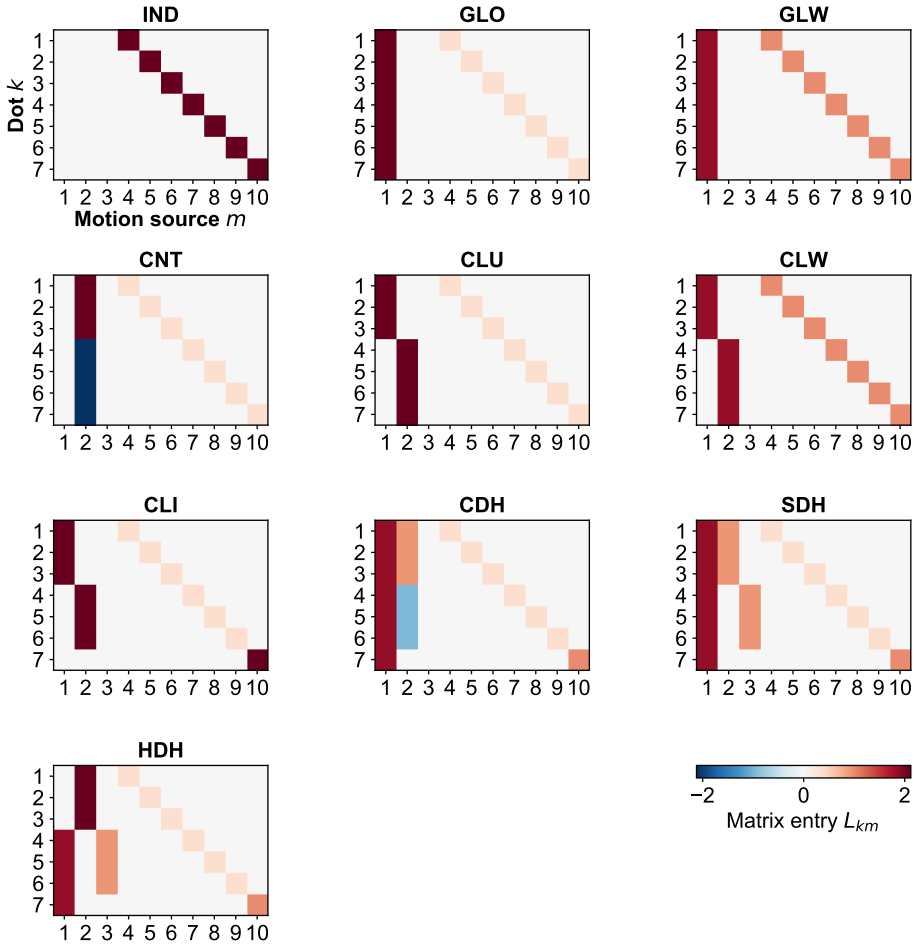

Figure 6. Motion structure matrices of the prediction task. All matrices L lead to the same marginal velocity distributions $p(v_k)$ for all dots k.

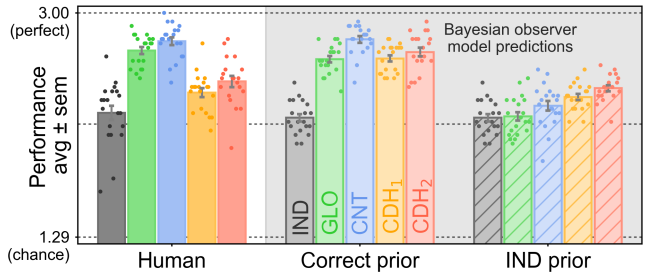

Figure 7. Probing a specific detrimental effect of perceptual noise to tracking of GLO stimuli. Re-evaluation of the MOT data with decoupled perceptual noise (for details, see *Appendix*, section "Discussion of the quantitative mismatch in the GLO condition of the MOT task"). Tracking of GLO stimuli is more affected by noisy observations than other stimulus conditions (especially, IND and CNT), likely due to the longer duration of dot proximity events.