COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 43 (2019) e12775

ISSN: 1551-6709 online
DOI: 10.1111/cogs.12775

The Cognitive Architecture of Perceived Animacy: Intention, Attention, and Memory

Tao Gao,[a] Chris L. Baker,[b] Ning Tang,[a] Haokui Xu,[a]
Joshua B. Tenenbaum[c]

[a]*Departments of Statistics and Communication, University of California, Los Angeles*
[b]*ISEE, Inc.*
[c]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Received 15 November 2016; received in revised form 30 April 2019; accepted 28 May 2019

Abstract

Human vision supports social perception by efficiently detecting agents and extracting rich information about their actions, goals, and intentions. Here, we explore the cognitive architecture of perceived animacy by constructing Bayesian models that integrate domain-specific hypotheses of social agency with domain-general cognitive constraints on sensory, memory, and attentional processing. Our model posits that perceived animacy combines a bottom–up, feature-based, parallel search for goal-directed movements with a top–down selection process for intent inference. The interaction of these architecturally distinct processes makes perceived animacy fast, flexible, and yet cognitively efficient. In the context of chasing, in which a predator (the "wolf") pursues a prey (the "sheep"), our model addresses the computational challenge of identifying target agents among varying numbers of distractor objects, despite a quadratic increase in the number of possible interactions as more objects appear in a scene. By comparing modeling results with human psychophysics in several studies, we show that the effectiveness and efficiency of human perceived animacy can be explained by a Bayesian ideal observer model with realistic cognitive constraints. These results provide an understanding of perceived animacy at the algorithmic level—how it is achieved by cognitive mechanisms such as attention and working memory, and how it can be integrated with higher-level reasoning about social agency.

Keywords: Cognitive modeling; Social perception; Chasing; Intention; Attention; Memory

1. Introduction

Our visual experience is not limited to the perception of physical properties, but also contains rich social content: representations of agents and their properties, including

Correspondence should be sent to Tao Gao, Departments of Statistics and Communication, UCLA, 8117 Math Sciences Bldg, Los Angeles, CA 90095. E-mail: tao.gao@stat.ucla.edu

goals, intentions, abilities, and relationships (Heider & Simmel, 1944). Imagine watching several kids playing on a playground. One will not just see several objects moving independently and randomly (as in the classic multiple object tracking displays of Pylyshyn & Storm, 1988). Instead, our perception can involve various types of social interactions, including chasing, fleeing, blocking, helping, hindering, and so on. On the one hand, these percepts are rapid, automatic, and occur at a glance, from extremely sparse cues—suggestive of bottom–up, intuitive processing. On the other hand, our conscious of social agency incorporates rich, context-dependent concepts and meanings, reflective of top–down, deliberative reasoning. Both aspects of perceived animacy are evident even from extremely sparse displays consisting of just a few lines and moving shapes, as first demonstrated by the classic Heider and Simmel (1944) animation, depicted in Fig. 1A.

Modern research has typically emphasized either the bottom–up or top–down processes involved in social perception. Recent research on bottom–up (visual) social perception has integrated the experimental rigor of psychophysical studies of multiple object tracking (MOT) (e.g., Pylyshyn & Storm, 1988; Scholl, Pylyshyn, & Feldman, 2001) and visual search (e.g., Treisman & Gelade, 1980; Wolfe, 1994) into displays styled after those of Heider and Simmel (1944). The goal is to evoke strong percepts of agency, animacy and intentionality, which vary quantitatively as a function of key stimulus parameters. Fig. 1B

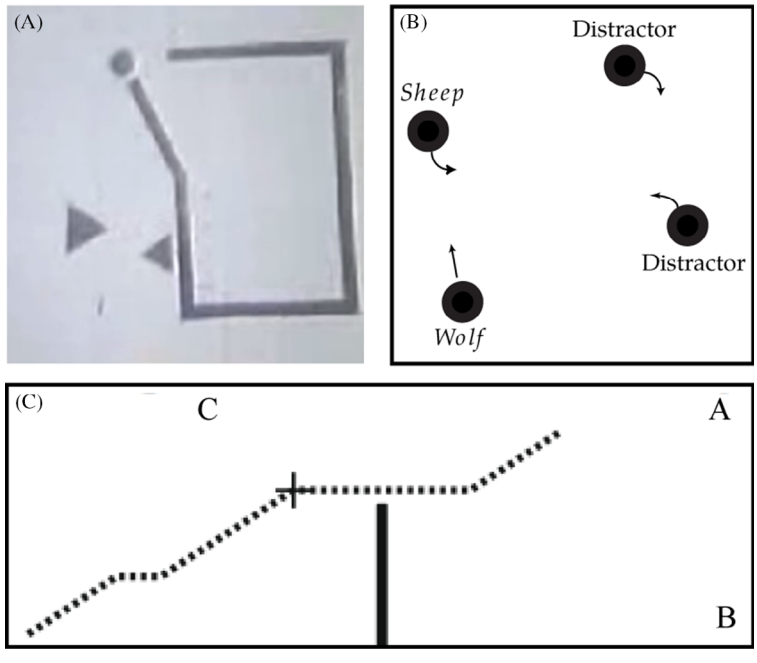

Fig. 1. Stimuli for studying perceived animacy. (A) The original Heider and Simmel (1944) display, which contains multiple agents in a complex environment. (B) The search-for-chasing display from Gao et al. (2009), in which multiple agents interact within a simple environment. (C) The display from Baker et al. (2009), in which a single agent moves toward various goal objects in an environment with obstacles.

shows an illustrative experimental display, used to measure psychophysical performance at detecting a "wolf" chasing a "sheep" as a function of the agents' movement noise and the number of distractors present. These types of displays have been used by many studies to measure the objective efficiency of searching for goal-directed motions (e.g., Gao, Newman, & Scholl, 2009; Gao & Scholl, 2011; Meyerhoff, Huff, & Schwan, 2013; Meyerhoff, Schwan, & Huff, 2014a), as well as the kinds of visual information that trigger various types of social percepts (e.g., Pantelis & Feldman, 2012; Tremoulet & Feldman, 2000). Barrett, Todd, Miller, and Blythe (2005) argue that the capacity to perceive intentional interactions in simple animated displays evolved as part of a biological adaptation to strong selection pressures for rapid detection and categorization of predation threats, mating opportunities, and so on and propose a computational model of these abilities that efficiently maps bottom–up, object-centered motion features to various interaction categories.

Computational accounts of top–down reasoning about intentional agents in simple animated displays have modeled how the structure of agents' actions and of the situational context shape conceptually rich mental state inferences, such as attribution of goals to individual (Baker, Saxe, & Tenenbaum, 2009; see Fig. 1C) and interactive (Baker, Goodman, & Tenenbaum, 2008; Pantelis et al., 2014; Ullman et al., 2010) agents, or attribution of beliefs and desires (Baker, Saxe, & Tenenbaum, 2011). These accounts formalize the "principle of rationality" from philosophy (Dennett, 1987) and developmental psychology (Gergely, Nádasdy, Csibra, & Bíró, 1995)—the assumption that intentional agents will act rationally to achieve their desires, given their beliefs about the world—in terms of probabilistic models of agents' rational belief-, desire-, goal-, and context-dependent action planning, based on accounts of rational utility-theoretic planning from AI and economics. Reasoning about mental states is formalized as Bayesian inference over internal models of rational planning, which can produce precise, accurate fits to quantitative human data across a range of contexts and inferences. However, these techniques may be infeasible under real-world processing constraints; it is unclear whether these computations can be optimized or approximated using efficient, bottom–up processes.

1.1. Reverse-engineering the cognitive architecture of perceived animacy

Here, our aim is to engineer a system that integrates bottom–up and top–down processing to achieve perceived animacy that are both rapid and richly meaningful. This presents an engineering challenge: How can we design a robust system that overcomes the limitations of each process, and builds on their strengths? Bottom–up processing is rapid and parallel, but inflexible; top–down processing is rich and flexible, but slow and expensive. Here, we argue that the function and performance of the system depend on its architecture, and that by attempting to reverse-engineer the cognitive architecture of perceived animacy, we can predict and explain human psychophysical data.

Based on previous work in this area and insights from the study of visual perception and cognition, we propose that the cognitive architecture should observe the following three computational principles: (a) Due to the stochastic nature of the world, agents'

movements over time, and human perception itself, the architecture should support probabilistic representations (Baker, Saxe, & Tenenbaum, 2009; Knill & Richards, 1996). (b) The architecture should provide scaffolding (e.g., data structures and interfaces) to support efficient approximate probabilistic inference, using computational processes analogous to principled inference algorithms from machine learning and statistics (e.g., Bishop, 2006). (c) These computations should be executed under the cognitive constraints (such as attention and memory) revealed by studies of visual cognition (Baddeley, 2003; Chun, Golomb, & Turk-Browne, 2011; Luck & Vogel, 1997), so that the model is cost-sensitive and neurologically and cognitively realistic.

Our architecture performs probabilistic computations within both bottom–up and top–down processes. The outputs of the bottom–up process can be accessed by top-down inference only through selective attention. Therefore, the connection between these two processes is low-bandwidth, limited by attentional foci. Different constraints apply to both processes. Bottom–up processes are constrained by the precision of parallel perception and the fidelity of iconic memory. Top–down processes are constrained by the precision of attentive perception (e.g., Srivastava & Vul, 2015) and the capacity of working memory. Our approximate inference algorithm given the outputs of the bottom–up and top–down processes is roughly inspired by particle filtering (Gordon, Salmond, & Smith, 1993), where each particle represents a particular hypothesized social interaction. A schematic of this architecture is shown in Fig. 2.

1.2. Cognitive constraints on perceived animacy

Understanding the cognitive constraints on perceptual processes is critical for modeling the architecture of perceived animacy, for two reasons. First, these constraints interact with the architecture in complex ways, requiring careful design to overcome. Second, they produce a unique, quantitative behavioral signature, which requires formal models to predict and explain. Here, we consider two types of constraints, on attention and working memory, respectively. Attention and working memory are core cognitive resources (for reviews, see Baddeley, 2003; Chun, Golomb, & Turk-Browne, 2011). While their engagement in the processing of physical objects has been extensively investigated, it is largely unclear how these fundamental cognitive processes support and constrain perceived animacy.

1.2.1. Selective attention

The presence of even a few agents can create a large space of possible social interactions. Recent psychophysical studies have shown that the performance of chasing detection drops as the number of distractors increases (Meyerhoff et al., 2014a), suggesting a key role for attentive processing in perceiving chasing. This finding is consistent with the fact that humans can only selectively track a few moving objects (e.g., Pylyshyn & Storm, 1988; Scholl, Pylyshyn, & Feldman, 2001). However, recent studies have shown that human detection of goal-directed and intentional actions also involves a pre-attentive process that can attract attention. Animate agents can automatically capture attention in

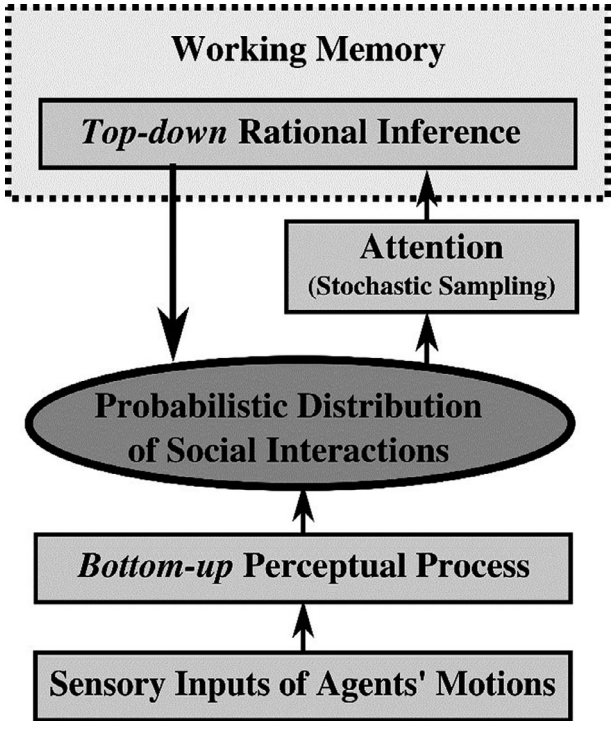

Fig. 2. A cognitive architecture for understanding social interactions. This architecture combines bottom–up perception with top–down inference, and it is constrained by limited cognitive resources, including attention and working memory.

both static (New, Cosmides, & Tooby, 2007) and dynamic displays (Pratt, Radulescu, Guo, & Abrams, 2010). In addition to capturing attention, perceived intentional motions can further facilitate the perception of low-level visual features such as orientation (van Buren & Scholl, 2014). In this study, we do not treat these results as contradictory findings that are intrinsically opposed. Nor do we seek a straightforward answer regarding whether the detection of perceived animacy is pre-attentive or attentive. Both processes are important and carry distinct functions. The challenge is to reveal how they work together to achieve efficient understanding of intentional action. Our architecture provides a natural solution by integrating pre-attentive, bottom–up information with an attentive, top–down selection process.

1.2.2. Decaying memory

Perceiving social interactions can also be constrained by the capacity of memory. Intuitively, any social interaction must last at least a few hundred milliseconds to be perceived as meaningful. Episodes shorter than this may not quite be informative for several reasons. First, social interactions require time to execute. An agent may move toward or away from another agent in a short period just by chance. Only a real predator will

pursue another agent persistently over time (Gao & Scholl, 2011). Therefore, relying on information presented in a very short period of time can produce many false alarms. Second, an agent's goal-directed motion in a single frame can be noisy, due to imperfect motion control or distractions from the environment and other agents. Third, our perception may not be able to precisely represent an agent's motion in a very brief period. For these reasons, accumulating agents' motion trajectories in memory for a certain amount of time is critical for filtering out the noise and extracting real goal-directed motion signals.

In the area of visual working memory, although there has been a large amount of work on the short-term storage of static features and objects, only a few studies have explored how working memory maintains objects' motion trajectories (e.g., Papenmeier, Huff, & Schwan, 2012; Sun et al., 2015). Consistent with discoveries of visual working memory, it has been shown that performance of identifying an animate display based on a narrative drops most dramatically between 3 and 4 items in the display (Wick, Alaoui Soce, Garg, Grace, & Wolfe, 2019). In this study, we systemically vary how rapidly motion information decays in the working memory component of our model and analyze how this memory rate impacts the perception of goal-directed motion. Top–down processing is assumed to have greater memory capacity than bottom–up processing. Similar to our strategy for studying attention, we test the memory capacity required within various architectures to capture human psychophysical performance, and assess whether or not these requirements are cognitively realistic.

1.3. The current study: Bayesian modeling with cognitive constraints

We construct models to reveal the computational processes underlying the perception of animacy and intention. Due to the complexity of human on-line perception and the interplay between attention, memory, and social perception, the space of possible model architectures is quite extensive. We devised the following approach in order to obtain a good match between model and human results while avoiding over-fitting. (a) We started from building an ideal observer model, which simply inverts the generative process of a chasing display by following the Bayes' rule. This process has no free parameter to fit as it is strictly constrained by algorithms generating the chasing displays. (b) Parameters of modeling human cognitive capacities (e.g., precision of attention, decaying rate of working memory) are all constrained by empirical Cognitive Science studies. For example, to model attentive tracking, we assume that attention can track 3–5 object and maintain motion information between 500 ms and 2 s. These parameters are not entirely free in the sense that they are limited to ranges that are cognitively realistic based on well-established facts of attention and working memory. (c) A set of parameters are then fitted by minimizing the human-model root-mean-square (RMS) error, including the precision and the memory rate parameters of the pre-attentive and attentive processes. This is the only "fitting" process in our approach. In addition, it was executed only in Experiment 1. (d) To avoid overfitting, all parameters are fixed after Experiment 1, leaving no free parameters to fit in Experiment 2 and Experiment 3. The results showed that parameters fitted in

Experiment 1 can be well generalized to Experiment 2 (with a new performance metric) and Experiment 3 (with a new set of experimental conditions). (e) As we argued in the introduction, a robust cognitive model should match human performance well even when its parameters deviate slightly from their optimal values. Supporting this argument, supplementary figures show that our models can capture patterns of human performance with parameters varying across the entire ranges that are cognitively realistic. The robustness of our models indicates that the success of our models lies in their cognitive architecture, rather than a set of fine-tuned parameters.

We focus on three models that have different cognitive architectures. (1) Bayesian Ideal Observer Model. This model has perfect precision and unlimited capacity in tracking and storing the agents' movements. By using Bayes' rule, the model rationally infers the agents' intentions that best explain the observed trajectories. Clearly, the unlimited capacity of this model is cognitively unrealistic. Nevertheless, it is necessary to construct this model first, as an objective measure of the absolute difficulty of the task. In a dynamic display with many objects constantly changing direction, it may be very challenging to detect a noisy chasing signal within a few seconds, even for an ideal observer model with unlimited resources. Therefore, testing this model against human performance provides a natural starting point for our investigation.

(2) Pure Attention (Serial) Model. This model assumes that a capacity-constrained attentive process, and this process alone, can explain human psychophysics. Like the ideal observer model, it also infers agents' intentions using Bayes' rule. However, its processing of the environment is incomplete as it can only "perceive" the agents that are currently tracked by its attention and store their movements for a limited amount of time.

(3) Hybrid Model. This model assumes that the psychophysics of chasing reflects the interaction between bottom–up and top–down process, which are connected through attention. Unlike the Pure Attention Model, it can perceive agents outside its attention in parallel, albeit with very limited precision and memory. This parallel process itself is insufficient for detecting chasing, but it can be used as "informed guess" to guide top–down attention.

In the next few sections, we first describe the psychophysical task we choose to test these models. We then introduce the specific implementations of these three models in the context of this task.

2. The psychophysics of chasing: Measuring the objective accuracy of perceived animacy

To study a cognitive architecture which integrates bottom–up and top–down perception, we need a task in which both types of process contribute in distinct ways. Here, we select a visual search task involving the detection of chasing (e.g., Dittrich & Lea, 1994; Gao, Newman, & Scholl, 2009; Meyerhoff, Huff, & Schwan, 2013), in which observers must identify a pair of agents respectively chasing and fleeing one another (a "wolf" and

a "sheep"; see Fig. 1B) from among a field of distractor objects. In this obstacle-free display, chasing is defined as direct heat-seeking, which makes the wolf move directly toward the sheep's position. Though simple, such stimuli reliably evoke the automatic perception of chasing. In addition, neuroscientific evidence shows that they strongly activate posterior temporal sulcus (pSTS), a critical brain region engaged in social perception (e.g., Gao, Scholl, & McCarthy, 2012).

We manipulate two key factors of our displays that demand bottom–up and top–down integration: the number of distractor agents present ("set size"), and the (in)directness with which the wolf chases the sheep ("chasing subtlety"). The number of possible pairs of social interactions is a quadratic function of the number of agents, inducing a large hypothesis space with even a few items (e.g., for set size N, the number of hypotheses is $N \times (N - 1)$). Chasing subtlety controls the maximum deviation from the heat-seeking direction (see Fig. 3, adapted from Gao et al., 2009). Every 200 ms, the wolf randomly selects a motion direction with deviation from the heat-seeking direction (clockwise or counter-clockwise) lower than the subtlety value. The objective efficiency of chasing drops as the subtlety value increases. However, as long as the subtlety value is smaller than 180°, there will be a correlation between the wolf and the sheep's motion, and the wolf can gradually get closer and closer to the sheep over time.

Two interesting patterns of behavioral results are observed. First, only small subtlety values (e.g., 0°, 30°) can be readily perceived as chasing, and the performance of detecting chasing among distractors quickly drops to near chance level with subtlety values larger than 90°. Second, unperceived chasing matters. There is a range of subtlety values with which chasers can escape detection while still reliably approaching the sheep—a type of "stalking" which exploits the limitations of the perceptual system. In an interactive experiment in which participants control the sheep themselves, they are most likely to be caught by unperceived chasers, with relatively high subtlety, and thus low objective efficiency, but high relative success (Gao et al., 2009). These results demonstrate a dissociation between successful chasing strategies and the subjective perception of chasing. Thus, the cognitive constraints that produce this dissociation have important implications for human perceptual performance.

2.1. General method: Search-for-chasing

Participants viewed a display (12° × 12°) containing moving discs (0.2° in radius) (see Fig. 1B). The number of discs was manipulated within and across experiments. The color of each item was randomly selected without replacement from a set of nine colors, with RGB values (250, 0, 0), (0, 200, 0), (20, 20, 255), (220, 220, 220), (255, 255, 0), (255, 0, 255), (0, 255, 255), (255, 100, 100), (255, 0, 50). (These colors can be viewed in an on-line demo at https://www.youtube.com/watch?v=-1vsICTGCng.) Every 200 ms each disc changed its direction. Most of the items moved randomly. However, one disc (the "wolf") chased another disc (the "sheep") with a randomly selected chasing subtlety value. The sheep randomly sampled 40 motion directions and selected the one that maximized its distance to the wolf. In practice, this made the sheep "flee" the wolf.

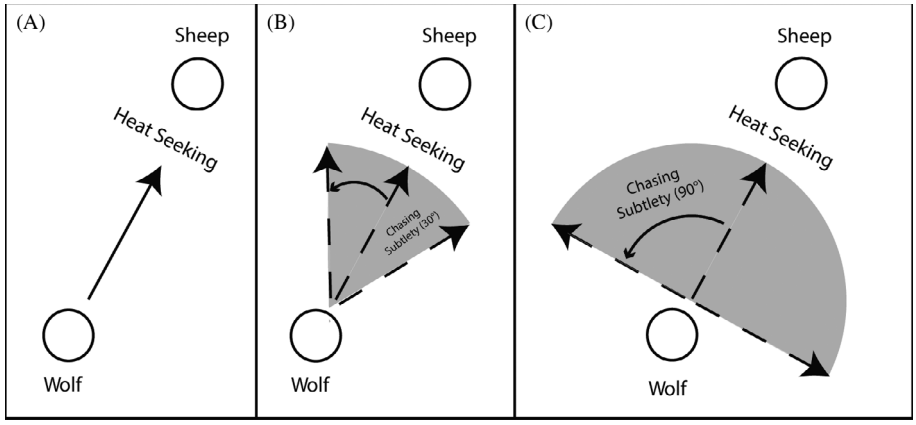

Fig. 3. An illustration of the chasing subtlety manipulation used in Experiments 1, 2, and 3. (A) When the chasing subtlety is 0, the wolf always heads directly toward the (moving) sheep, in a "heat-seeking" manner. (B) When the chasing subtlety is 30, the wolf is always heading in the general direction of the sheep, but it is not perfectly heat-seeking; instead, it can move in any direction within a 60 window, with the window always centered on the (moving) sheep. (C) When the chasing subtlety is 90, the wolf's direction of movement is even less constrained: Now the wolf may head in an orthogonal direction to the (moving) sheep, but it can still never be heading away from it. The shaded areas in (B) and (C) indicate the angular zone which constrains the wolf's direction of movement on that given frame of motion.

Each trial lasted 8 s. Participants were asked to press the "space" bar to stop the experiment as soon as they detected the wolf and sheep. They then needed to left-click the mouse to select the wolf and sheep in order. The size of the selected item would increase by 10% and a letter "W" or "S" would be presented in the center of the object. They could also right-click to cancel the selection of an item. Both accuracy and reaction time (RT) were recorded. Accuracy was defined as identifying both the "wolf" and "sheep" correctly. RT was defined by the duration between the start of a trial and the press of the "space" bar.

2.2. The ideal observer model of chasing-detection

We first construct an ideal observer model, which is a pure top–down causal model with unlimited cognitive capacity, allowing the model to sense, attend to, and retain all agents' interactions perfectly. This model allows us to evaluate the objective computational challenge of the task. It does not guarantee perfect performance, but sets an upper bound on the performance attainable by a model. The mathematical equations for this and other models are all listed in Supplementary Material file.

The structure of the ideal observer model is depicted in Fig. 4. It is based on the causal processes generating the wolf, sheep, and distractors' trajectories. The model assumes that the wolf-sheep identity and the chasing subtlety are latent variables that must be inferred. At the highest level, the model assumes that two items are randomly selected as the wolf and sheep from the total of N items. As discussed earlier, there will be

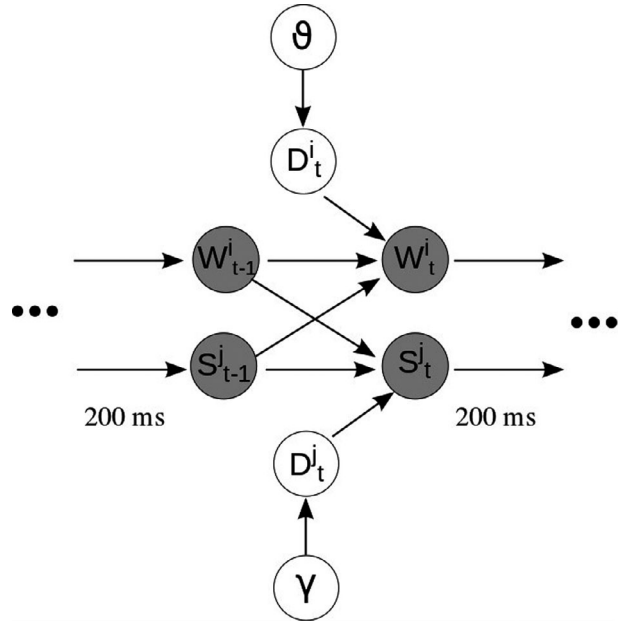

Fig. 4. The ideal observer model. θ is the chasing-subtlety value, which is sampled uniformly from the subtlety values used in an experiment. γ is the escaping subtlety. It is fixed at 1.92, which is empirically estimated by using the variance of the stochastically generated sheep trajectories. W^i and S^j form a wolf–sheep pair, which is randomly selected from $N \times (N-1)$ pairs, where N is the set size. W_t^i and S_t^j are the positions of the hypothesized wolf and sheep at time t. D_t^i and D_t^j are W_t^i and S_t^j's deviations from the heat-seeking chasing and escaping directions. W_t^i and S_t^j are then determined by their previous positions W_{t-1}^i and S_{t-1}^j and D_t^i and D_t^j.

$N \times (N-1)$ different wolf-sheep combinations. The wolf's chasing subtlety is assumed to be uniformly sampled from the prior over subtlety values, remaining constant across a single trial. At every time interval, the wolf's deviation from perfect heat-seeking is randomly sampled given its chasing subtlety value. The sheep's motion is assumed to be generated in a similar way, but with a fixed distribution that we directly estimate from the sheep's trajectories. This distribution assumes very high variance in the sheep's motion, capturing the assumption that the sheep's motion plays a much weaker role in detecting chasing.

Given these assumptions about the data-generating process, the Bayesian inference procedure is as follows. For each possible wolf–sheep pair, the model takes as input the hypothesized wolf and sheep's chasing and escaping deviation from the heat-seeking direction (Eqs. 1–2 in Supplementary Material file). The joint probability of each hypothesized wolf–sheep pair and the wolf's subtlety value can then be updated given the likelihood that such a combination produced the observed deviations (Eqs. 3–5). As participants were not required to judge the subtlety values, the model integrates over subtlety values to compute the marginal distribution of the wolf–sheep pair (Eq. 6). Note that

this model is a capacity-unlimited model, as it assumes that all items' motion directions are perfectly observed and perfectly integrated over time.

2.3. Pure attention model

The second model we construct is a capacity-limited ideal observer model motivated by the perception and memory constraints discussed in the Introduction. This model is based on several assumptions. (a) Limited attention: Observers cannot track all the wolf-sheep hypotheses but have to select a subset of them. The number of hypotheses that can be selected reflects the "capacity" of attention. (b) No pre-attentive processing: There is no access to information about the untracked hypotheses (this assumption will be relaxed in the third model we construct). (c) Perceptual noise and decaying memory: The function of attention is to process the tracked hypotheses with a certain precision (Eqs. 7–14 and Eq. 17) and to retain the outcome in working memory over time with a certain decay rate (Eqs. 15–16 and Eq. 18). (d) Rational top–down inference: Like the ideal-observer model, instead of detecting a fixed motion pattern, this model rationally infers chasing by integrating over all the possible chasing subtlety values used in an experiment. (e) Attention switching: Each slot of attention can randomly drop the hypothesis is tracking at time t and resample a new hypothesis at time $t + 1$ from the posterior distribution of the wolf–sheep pair, which is simply a categorical distribution. (f) Independent attention selection: Multiple slots of attention can select the same hypothesis, which can further improve the perceptual precision (Eq. 17) and memory of that hypothesis (Eq. 18).

The free parameters in this model include the following: (a) The perceptual noise for estimating the deviation of an item's motion direction relative to the heat-seeking chasing or escaping direction. This noise can come from multiple sources, including estimating the current location and motion direction of an item, and comparing its motion direction with the heat-seeking direction. For simplicity, here we do not model each of these processes separately, but represent their sum as a von Mises distribution, the precision of which reflects the amount of the noise. As can be seen from the results of Experiment 1, while a higher precision can improve the performance of the model, the overall pattern of the model is robust across different values of this parameter. (b) The memory decay rate. This parameter controls the rate of exponential decay of an item's motion trajectory in memory, reflected by the ratio of its contribution to the likelihood of a hypothesis. We fix this parameter at 0.7, which captures our intuition that working memory should store dynamic motion information for around 2 s. Since the motion direction updates every 200 ms, this means that memory of a motion direction almost completely decays after 2 s (during a 2 s period, an agent updates its motion direction 10 times; $0.7^{10} = 0.028$). (c) Each slot's resampling of its tracked hypothesis is modeled as a Poisson distribution with a mean of 600 ms (during which an item updates its motion direction for three times). This value is based on the duration of temporal attention revealed by attentional blink (e.g., Chun & Potter, 1995). Note that this does not mean that each slot will change its tracked hypothesis every 600 ms. If the likelihood of the tracked hypothesis is high, it can be more likely selected again by this slot (and other slots). Indeed, in many trials, once the correct hypothesis is tracked by attention, its posterior probability becomes high,

and it will be persistently selected by the same slot and attract more slots onto it. As a result, in the end of a trial, all attention may focus on a single hypothesis.

2.4. Hybrid model combining pre-attentive and attentive process

This model is in part motivated by Wolfe's guided-search model (e.g., Wolfe, 1994) for searching static displays. It relaxes the Pure Attention model's assumption that there is no access to the untracked hypotheses. Instead, it assumes that before selected by attention, all hypothesis can be processed by a parallel process, which is both quantitatively and qualitatively different from the attentive process. This process has the following features: (a) Cue-based search: as a bottom–up process, pre-attentive processing detects heat-seeking motion direction as a fixed cue, instead of integrating over all possible subtlety values in an experiment. (b) Poor quality: compared with the attentive process, the perceptual noise is much larger and the memory decays much faster, which is probably based on iconic memory that lasts <1 s (Dick, 1974; Sperling, 1960). For a more comprehensive discussion of the iconic memory of motion direction, see Shooner, Tripathy, Bedell, and Ogmen (2010). (c) No concentration of resources: While more attentional resources can be focused on the same hypothesis to further boost perception and memory through a resampling process, there is no such process for parallel processing.

The function of the parallel process is to guide the selection of attention. For the pure attentive model, all untracked hypotheses have the same likelihood, as they evenly split the probability mass left by the sum of the tracked hypotheses. For the hybrid model, these untracked hypotheses will have different likelihood, due to the parallel processing. Therefore, the switch of attention during the resampling process will be impacted by the outcome of the pre-attentive process, as the hypotheses favored by pre-attentive process will be more likely selected by attention for further processing (Eqs. 20–21). With this model, we can demonstrate how a cognitively reasonable parallel process can dramatically decrease the number of attention slots required for achieving human-level performance.

2.5. Models versus human

We evaluated multiple models by comparing their results with human performance at two levels: experimental condition and individual trial. Each condition of human experiment was defined by the combination of two variables: Chasing Subtlety and Set Size. Within each condition, multiple trajectories were generated. Every trajectory was then presented to human observers and models as an individual trial.

At the condition level, the mean accuracy of multiple trials over each condition was calculated for both humans and the model. Human-model similarity was measured by the root-mean-square (RMS) error, which is a widely accepted metric for evaluating the differences between values predicted by a model and values observed. In cognitive modeling, the values observed were human data. We want to emphasize that our experimental conditions were not chosen arbitrarily. Instead, they were carefully designed so that different models could produce dramatically different patterns of accuracy across conditions. In this approach, modeling is not just a tool for fitting human results. Instead, it can actively guide the design of future psychophysical experiments. We view this approach

as a good demonstration of the power of combining psychophysics with Bayesian modeling.

At the trial level, chasing detection of each individual trial was analyzed separately, instead of aggregated into the condition mean. Within each Subtlety-Set Size condition, the difficulty of each trial still varied, due to the stochastic processes of generating these trials. Compared with variances across conditions by explicit manipulations, these within-condition trial-by-trial variances were much more subtle, as even experiment designers could not control them. Therefore, significant human-model correlation at the trial-by-trial level has been taken as a strong evidence supporting the validity of a cognitive model (e.g., Brady & Tenenbaum, 2013).

3. Experiment 1: Modeling the interaction of subtlety and set size

3.1. Method

In this experiment, we manipulated both the number of items (3, 4, 6, 9) in the display and the chasing subtlety values (5°, 30°, 60°, 90°, 120°, 150°, 180°). There were 224 trials in total, 8 trials for each condition. Before the formal experiments, there were 12 practice trials, results from which were not recorded. The whole experiment lasted about 50 min. Twelve undergraduate students at Zhejiang University participated in this experiment for payment.

3.2. Human results

The overall accuracy of chasing detection (i.e., identifying both the wolf and sheep correctly) was 43.9%. The accuracy of identifying the wolf was 52.8%. The accuracy of identifying sheep was 51.7%. As the interpretation of just identifying wolf or sheep was unclear, here we focused on chasing detection.

The accuracy of chasing detection and reaction time (RT) of correct detection as a function of set size and chasing subtlety were shown in Fig. 5A and 5B, respectively. For both Accuracy and RT, the main effects of subtlety value, the number of items, and their interactions were significant (see Table 1). These results showed that search for chasing becomes more challenging with larger set size, which was consistent with previous findings (Meyerhoff et al., 2013, 2014a). In addition, there was an interesting Subtlety-Set Size tradeoff: To achieve ~50% accuracy, the subtlety values were 120°, 90°, 60°, 30° for 3, 4, 6, 9 items, respectively. (The chance levels of these fourconditions are 16.7%, 8.3%, 2.4%, and 1.4%, respectively). We would further explore these four conditions in Experiment 2.

3.3. Ideal observer model results

The accuracy and RT of the ideal observer model are depicted in Fig. 6A and 6B, respectively. Clearly, given perfect perception, memory, and the capacity to track all

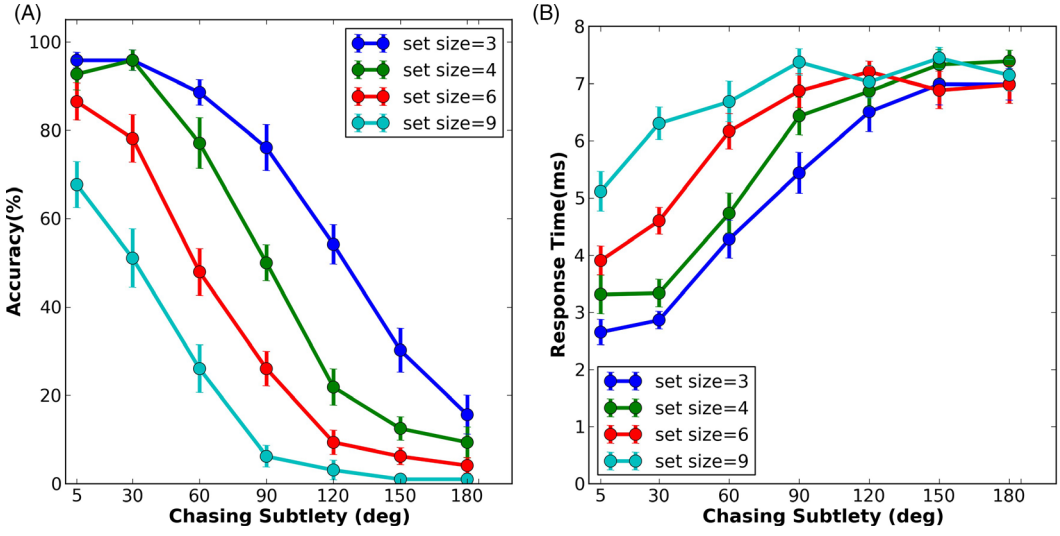

Fig. 5. Human accuracy results from Experiment 1 as a function of chasing subtlety and set size. (A) The accuracy (% correct) of identifying both the wolf and sheep correctly. (B) The reaction time (s) of correctly identifying the wolf and sheep.

Table 1
The main effects of subtlety, set size, and their interactions in Experiment 1

	Subtlety	Set Size	Subtlety × Set Size
Accuracy	$F(6, 66) = 275.3$; $p < .001$	$F(3, 33) = 189.197$; $p < .001$	$F(18, 198) = 8.842$; $p < .001$
RT	$F(6, 66) = 154.689$; $p < .001$	$F(3, 33) = 72.848$; $p < .001$	$F(18, 198) = 12.588$; $p < .001$

hypotheses simultaneously, the ideal observer model could achieve much higher performance than humans. We employed the root-mean-square (RMS) error as the index for measuring the human-model difference. The RMS errors for accuracy and RT were 43% and 2.2 s, respectively. This large discrepancy ruled out the hypothesis that the drop in human performance for large subtlety values is simply due to the intrinsic difficulty of the task. Although the ideal observer model was an unrealistic model of human perceptual processing, it provided a basis for our subsequent investigation of more cognitively realistic models. In these models, we shifted our focus from the generative process of the chasing display to the cognitive architecture and constraints of human perception.

3.4. Pure attention model results

The accuracy and RT with precision 8 and capacity 8 are shown in Fig. 7A and 7B, respectively. The results of every combination of precision (4, 6, 8) and capacity (2, 4, 8, 12, 20 slots) are shown in supplementary figures, Fig. 1A and 1B. While increasing the precision of each slot could increase the model's performance, the overall pattern of

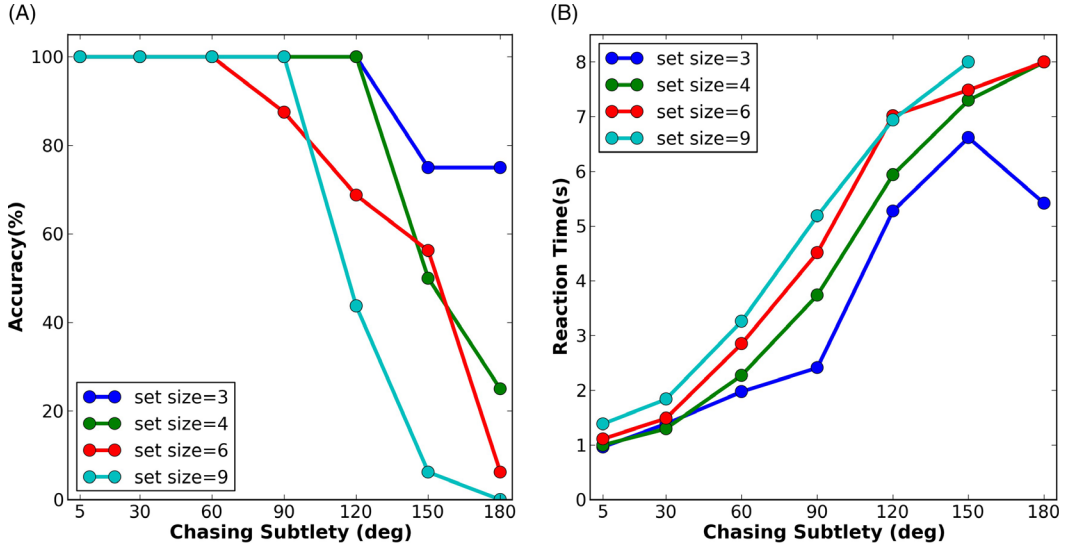

Fig. 6. Ideal observer model results from Experiment 1 as a function of chasing subtlety and set size. (A) The accuracy (% correct) of identifying both the wolf and sheep correctly. (B) The reaction time (s) of correctly identifying the wolf and sheep.

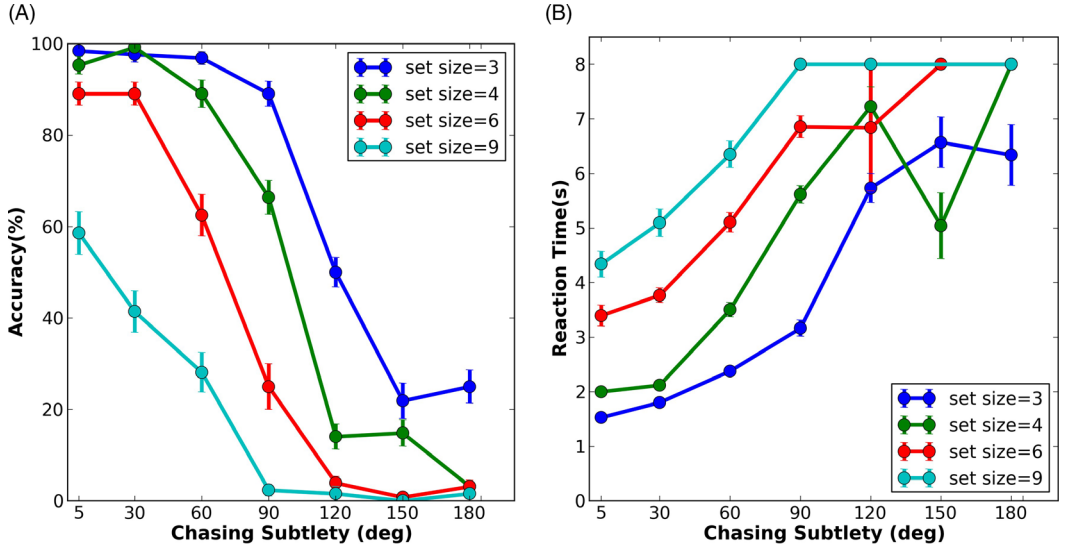

Fig. 7. Pure attention model results from Experiment 1 with 8 attention slots and precision 8. (A) The accuracy (% correct) of identifying both the wolf and sheep correctly. (B) The reaction time (s) of correctly identifying the wolf and sheep. The large SE for large subtlety values is due to the small number of correct trials from these conditions.

results was robust across different precision values. Eight slots fitted the human performance best. The Accuracy RMS errors for precision 4, 6, 8 were 7.8%, 7.8%, 6.7%, respectively; the corresponding RT RMS errors were 1.18, 1.10, 1.21 s.

This model demonstrated that adding attention and memory constraints to an ideal observer model can dramatically shape the model's results to a pattern much closer to human performance. It indicates that cognitive capacity limitations are critical factors determining the psychophysics of chasing. However, this model is still not satisfying. Since each hypothesis consists of a wolf and a sheep, and their relative motion directions, tracking eight hypotheses requires working memory to maintain 16 motion directions for 2 s. This requirement seems to be cognitively unrealistic given the capacity limitation of visual working memory (Sun et al., 2015). Our next model tested whether comparable performance can be achieved with lower attentional resources by leveraging pre-attentive processing.

3.5. Hybrid model results

For the hybrid model, we only allowed the model to store four relative motion directions, which corresponded to tracking only two wolf-sheep hypotheses. The primary motivation was to explore whether the model can reach human-level performance by combining a highly limited attention and memory capacity with cognitively realistic pre-attentive parallel processing. The precision (8) and memory rate (0.7) were identical to those in the Pure Attention Model. In contrast, the precision and memory rate for the parallel processing were much lower, as we expect that the quality of pre-attentive processing should be much worse than that of attentive processing. Pre-attentive precision was selected from (2, 2.5, 3), while the memory rate was selected from (0.4, 0.45, 0.5).

The accuracy and RT of all combinations of parallel precision and memory rates are shown in supplementary figures, Fig. 2A and 2B. This model generally fitted human results well. The mean RMS errors of nine conditions were 8% and 0.85 s for accuracy and RT, respectively. The small RMS error across several conditions indicated that the Hybrid model's good fit to human performance was largely due to its architecture, and it did not require a careful selection of the combination of different parameters. The smallest RMS error (Accuracy: 6%; RT: 0.77 s) occurred when the parallel processing had a 2.5 precision and 0.45 memory rate. The performance as a function of set size and subtlety values with these two parameters isshown in Fig. 8A and 8B, respectively. To avoid over-fitting, in the rest of the paper, we would "freeze" the hybrid model with these parameters to explore how well it can model the results of different experiments. However, we want to emphasize here that the major focus of the current study is not to reveal the exact precision and memory rate of the visual system. Doing so may require another series of psychophysical experiments dedicated for this purpose. Here, we are satisfied to observe that a range of parameters can yield good fits to human performance, which reduces the risk of over-fitting and demonstrates the contribution of the cognitive architecture.

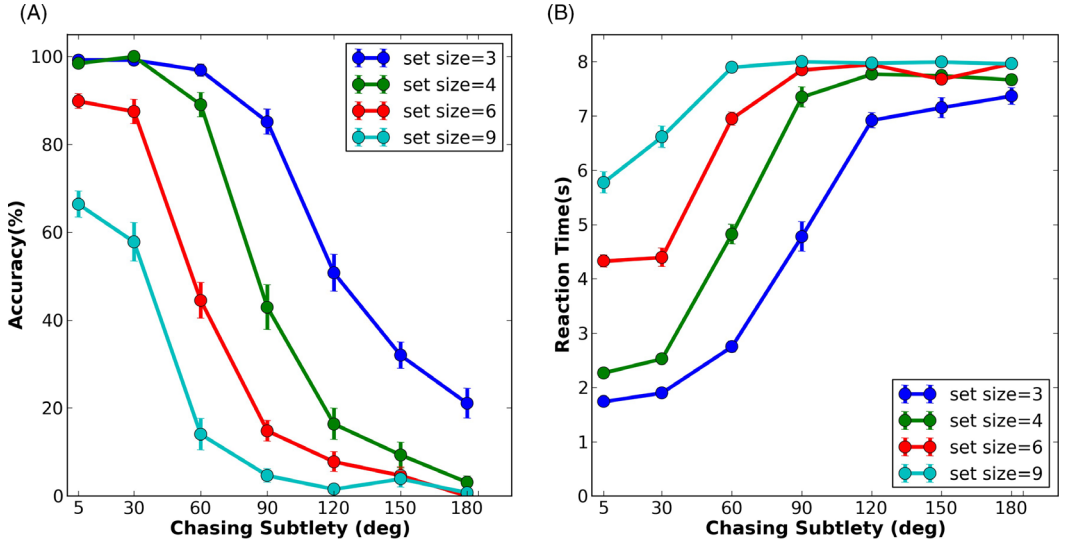

Fig. 8. Hybrid model results from Experiment 1 with 2 attention slots. The perceptual precision and memory rate for the pre-attentive process is 2.5 and 0.45, respectively. (A) The accuracy (% correct) of identifying both the wolf and sheep correctly. (B) The reaction time (s) of correctly identifying the wolf and sheep.

3.6. Discussion

Experiment 1 compares human performance against three different models. While a model with unlimited capacity outperforms humans, both the Pure Attention Model and the Hybrid model can match human performance across various combinations of set sizes and subtlety values. These results collectively demonstrate that the psychophysics of chasing can be characterized using a capacity-limited Bayesian ideal observer model. However, it is difficult to distinguish the Pure Attention Model and the Hybrid model with the current experiment. The Pure Attention Model is more parsimonious, as it does not assume a parallel pre-attentive process, which requires two free parameters. However, the model needs to track at least 8 hypotheses (16 relative motion directions) to reach human performance. This requirement seems to be unrealistic, given existing studies on visual working memory. For the hybrid model, the number of tracked hypotheses is fixed at 2, given an a priori assumption that humans can only track four relative motion directions. While adding a parallel process introduces two more free parameters, their exact values are not critical, as a range of parameters produce a similar pattern of results. In addition, the existence of a parallel, pre-attentive process is also supported by existing behavioral studies (Pratt, Radulescu, Guo, & Abrams, 2010). Taking these considerations together, we prefer the Hybrid model over the Pure Attention Model, although a conclusion cannot be reached given the current experiment alone.

4. Experiment 2: Trial-by-trial correlation of subtlety-set size tradeoff

In Experiment 1, we explored the breadth of the Pure Attention and Hybrid models by showing their fits to a wide range of Subtlety-Set Size combinations. The human-model comparison was conducted by averaging performance of trials within each Subtlety-Set Size combination. In Experiment 2, we further evaluated the "depth" of these two models by investigating whether their results correlate with human performance at a trial-by-trial level within each condition. Since each trajectory is stochastically generated, even with the same combination of set size and subtlety, some trials will be more challenging than others.

4.1. Method

Given the results of Experiment 1, we selected four Subtlety-Set Size combinations for evaluating trial-by-trial correlation: (120°, 3), (90°, 4), (60°, 6), (30°, 9), in which the first value represented the subtlety value and the second value represented the number of items. Experiment 1 showed that these four conditions gave roughly the same performance around 50%, which provided a good opportunity for exploring trial-by-trial variance. Since these four conditions were very challenging, accuracy, rather than RT, was emphasized. We did not include the Capacity-Unlimited Ideal Observer model in our analysis, as the performance will be close to 100%, indicating very little trial-by-trial variance. There were 30 trials for each of the four conditions; 16 undergraduates at Zhejiang University participated for cash.

4.2. Results

The overall accuracy of chasing detection was 57.8%. The accuracy of identifying the wolf was 65.6%. The accuracy of identifying sheep was 62.6%. As in Experiment 1, we focused on chasing detection. The averaged accuracy of chasing detection as a function of Subtlety-Set Size is depicted in Fig. 9. The pattern of accuracy across humans, and the pure attention (serial) and hybrid models were consistent with the results of Experiment 1 in the same Subtlety-Set Size conditions.

Our primary focus here was the trial-by-trial correlation within each condition. We first explored whether there was such a correlation among human participants, by conducting a bootstrapped "split-halves" analysis on the human results. Over 1,000 iterations, we randomly split participants into two halves, and computed the mean accuracy for each trial for both halves. Then, we computed the trial-by-trial correlations within each Subtlety-Set Size combination between the two random halves. The mean correlation coefficients and mean p-values are shown on top of each bar in Fig. 10. The first three Subtlety-Set Size conditions showed reliable correlations in the trial-by-trial accuracy, but not the fourth. Overall, these results indicate that for small set sizes, human subjects consistently found some trials more difficult than the others. In the fourth condition (9 items), the large Set Size made search difficult, but this difficulty was relatively constant across trials.

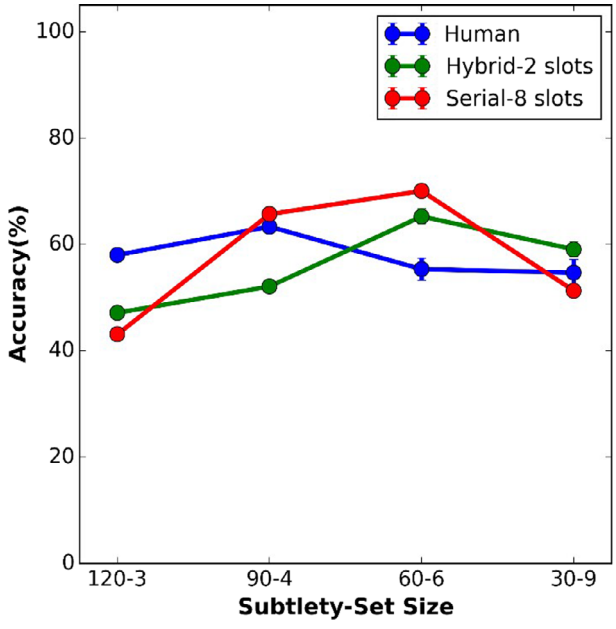

Fig. 9. The accuracy of chasing detection from four Subtlety-Set Size conditions in Experiment 2. The relative flat curves show a Subtlety-Set Size tradeoff.

We expected our models to correlate with human trial-by-trial accuracy in the first three conditions, but not the fourth. This hypothesis was confirmed by the results, showing that the accuracy of the Serial and Hybrid models correlated with human performance with set size 3, 4, and 6. There was no significant correlation with set size 9. These results showed that the capacity-limited models can explain human performance at a trial-by-trial level. However, similar to Experiment 1, the results of Experiment 2 did not distinguish which of the hybrid and pure attention model better explains human performance.

Both the Serial and Hybrid models correlated with trial-by-trial human performance significantly. In addition, Fig. 10 did not show a clear pattern of how these correlation coefficients varied across models. Therefore, we designed Experiment 3 in which the Serial and Hybrid models make different predictions of human performance.

5. Experiment 3: Search chasing in large set sizes

In this experiment, we further tested the two capacity-limited models by using much larger set sizes, including 12, 16, and 20 items. The number of possible wolf–sheep pairs is a quadratic function of the number of items—with 20 items, there are 380 possible hypotheses. By using these large sets, this experiment pushed human participants as well as the two models to their limits. The results should produce a clear distinction between

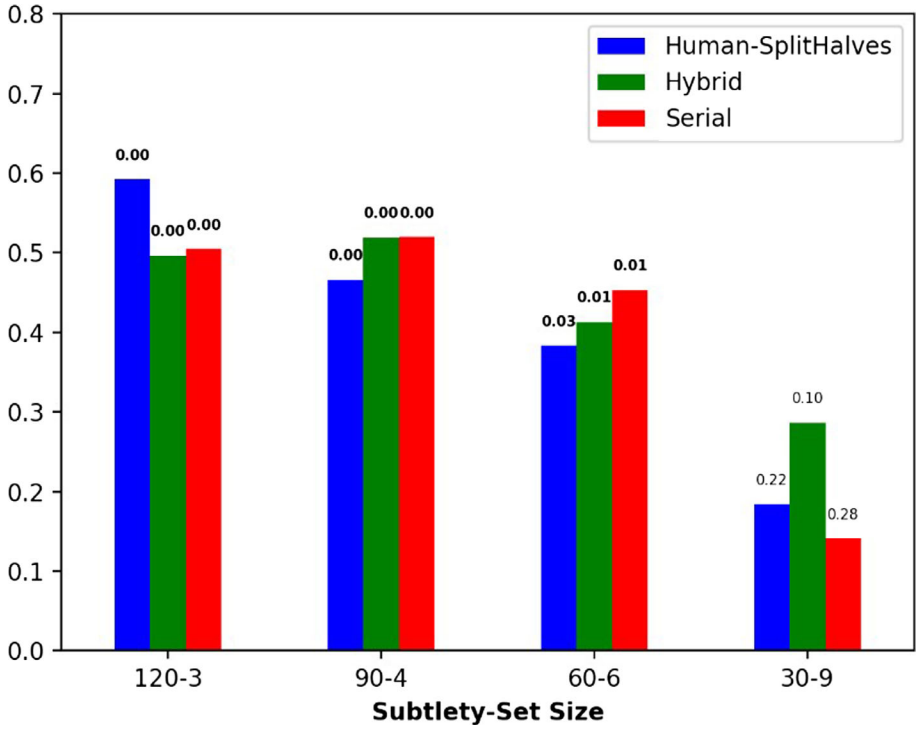

Fig. 10. Trial-by-trial correlations within each Subtlety-Set Size condition. The height of each bar represents the correlation value. The digit on top of each bar represents the p value.

the ability of the two models, Pure-Attention or Hybrid, to capture human performance. Since the results of Experiment 1 already showed that for 9 items, subtlety values larger than 30° become very difficult to detect, here only subtlety 5° was used. Since the results of Experiment 2 showed that a trial-by-trial correlation dropped with larger set sizes, we did not include such analyses in the current experiment. Instead, the dependent measurements were averaged Accuracy and RT (the same as Experiment 1).

5.1. Method

Eight undergraduates at Zhejiang University participated in this experiment for cash. There were 40 trials for set size 6, 12, and 20. The RGB values of each item's color were uniformly sampled from 0 to 255. The other aspects of this experiment were identical to those of Experiment 1.

5.2. Results

The overall accuracy of chasing detection was 49.5%. The accuracy of identifying the wolf was 52.9%. The accuracy of identifying sheep was 52.5%. Accuracy and RT of

chasing detection from humans, Pure-Attention (Serial) model with different slots and Hybrid models are depicted in Fig. 11A and 11B, respectively. The RMS error of the Pure-Attention and Hybrid models compared with human performance is depicted in Fig. 12.

There are several interesting findings. First, the ideal observer model reached ~100% accuracy, again indicating that cognitive constraints must be modeled to explain human performance. The Pure-Attention model with 8 slots performed much worse than humans. Additional simulations showed that to reach human performance with small RMS errors, this model required 20 slots. In contrast, the same Hybrid model could still match human performance well, suggesting that this model can explain human psychophysical data across a wide range of subtlety values and set sizes.

We conducted additional simulations and analyses to further reveal the computational efficiency of different components (attentive vs. pre-attentive) of the Hybrid model. The attentive process can be isolated by removing the pre-attentive process. In practice, this can be easily done by setting both the perceptual precision and memory duration of the pre-attentive process to 0, making the Hybrid model degenerate to a Pure-Attention model with two slots. The pre-attentive process can be isolated by setting the number of the attention slots to 0. The performance of the two isolated components is shown in Fig. 13. The most notable result was that the isolated processes perform strikingly poorly, with accuracy lower than 10%. These results indicated that integrating these two processes in the Hybrid model produced a super-additive effect. It was also interesting to notice that the RT of the parallel components was always 8 s, which was the maximum

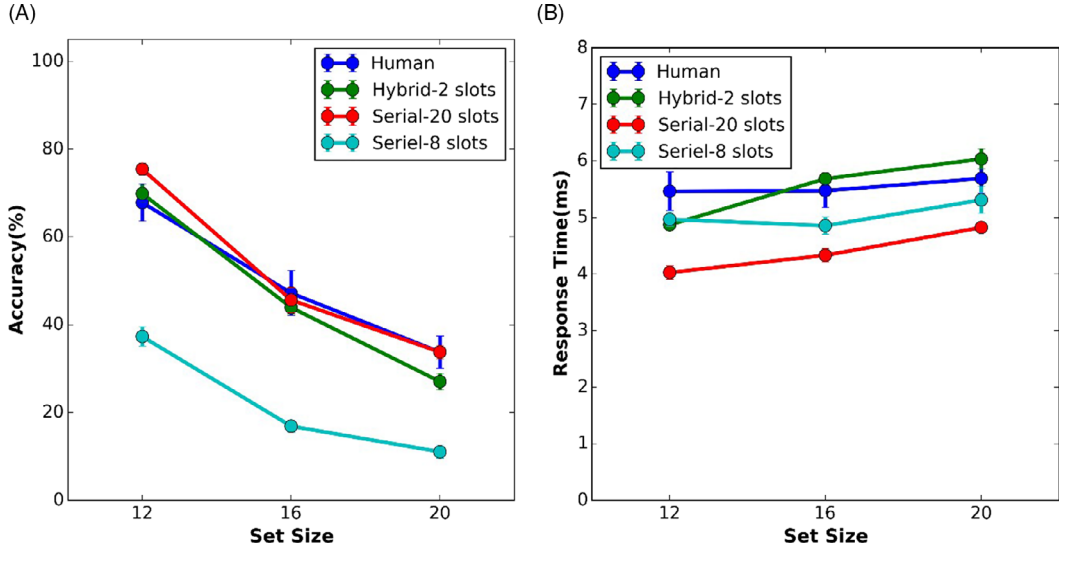

Fig. 11. Experiment 3 performance of humans, the Serial model with 8 and 20 slots, and the Hybrid model. (A) The accuracy (% correct) of identifying both the wolf and sheep correctly. (B) The reaction time (s) of correctly identifying the wolf and sheep.

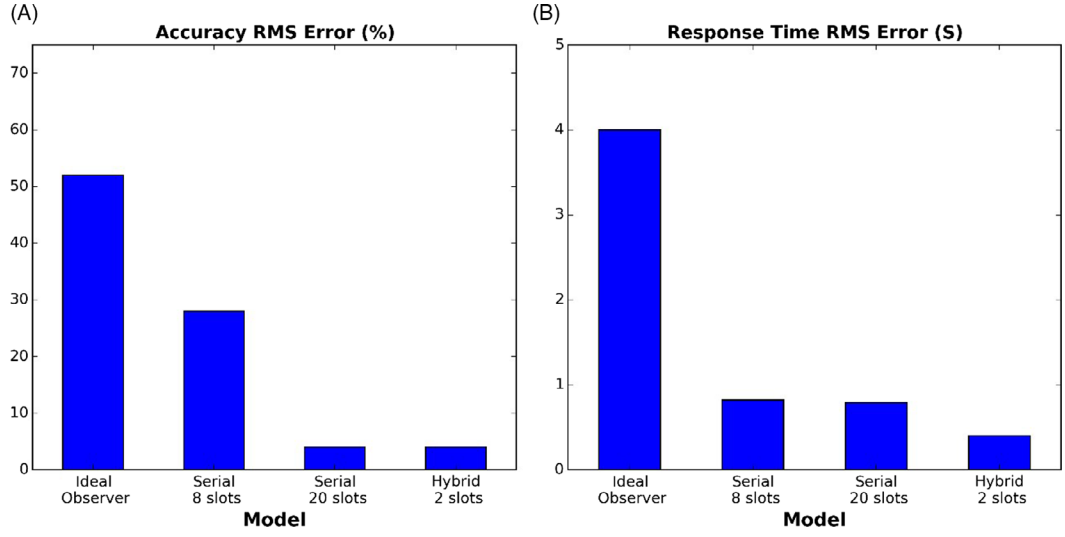

Fig. 12. The RMS errors of the Serial and Hybrid models, compared with human accuracy (A) and RT (B). To match human performance, the Hybrid model needs only 2 slots, while the Pure-Attention model needs 20 slots.

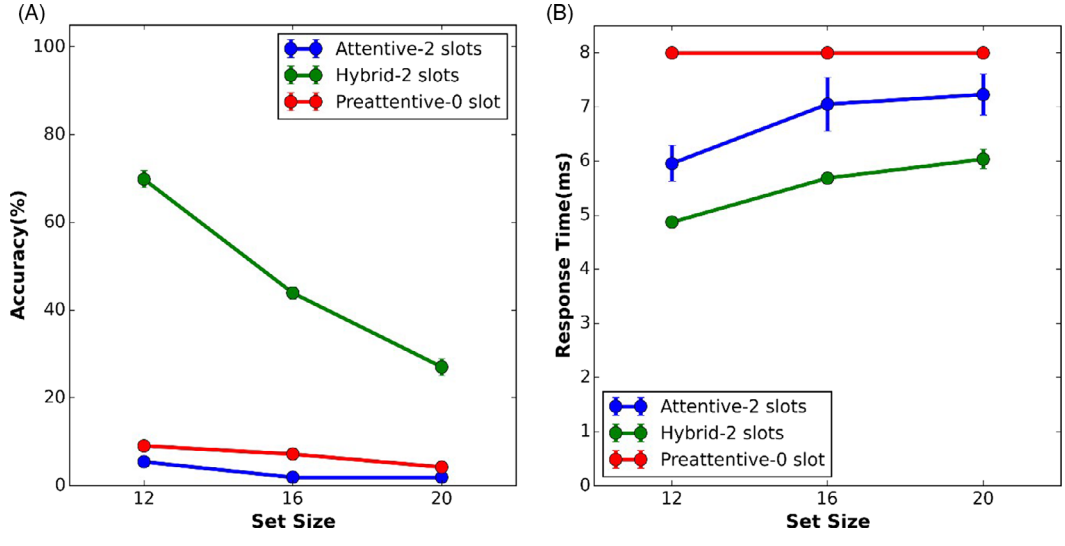

Fig. 13. Accuracy (A) and RT (B) of the isolated Pre-attentive and Attentive processes, compared with the Hybrid model.

duration of a trial. This indicated that while the parallel process could guide the attentive process, it alone could never make the model "confident" enough to initiate a response. Implications of these results were considered further in the General discussion.

6. General discussion

We proposed a cognitive architecture for perceived animacy, in which bottom–up and top–down processes are integrated with efficient usage of limited attention and memory. This architecture is supported by the following psychophysical and modeling results: (a) Human results indicate a tradeoff between set-size and subtlety values, in which chasing detection performance drops while either set-size or subtlety increases. (b) A pure top–down Bayesian ideal observer model with unlimited cognitive capacity significantly outperforms human results and does not show a similar tradeoff. (c) Human results can be best explained by constraining the Bayesian ideal observer model with attention and memory limitations, in combination with a parallel inflexible bottom–up process. (d) Top–down and bottom–up processes are integrated by an efficient "scheduling" mechanism, which allocates limited cognitive resources by evaluating the outputs from both bottom–up and top–down processes. These results collectively reveal a cognitive architecture, in which top–down and bottom–up processes are integrated subject to limited cognitive capacity in a probabilistic inference framework. The properties of these two processes are summarized in Table 2.

6.1. The integration of bottom–up and top–down processes

The cognitive architecture we explored not only explains human performance, but also has properties that are desirable from a computational efficiency standpoint. Top–down and bottom–up interactions are not a new topic in vision science. Such interactions are typically demonstrated by behavioral performance (e.g., Wolfe, 1994), visual illusions (e.g., Di Lollo, Enns, & Rensink, 2000), or neuroscientific methods, such as identifying back projections from higher-level visual cortex to primary visual cortex (e.g., Ahissar & Hochstein, 2004). In contrast, there are fewer studies directly analyzing the computational efficiency of such interactions, how they allow human vision to understand scenes that would be difficult or impossible to evaluate by each process alone.

Table 2
Contrasting bottom–up and top–down processes

Bottom–up, Pre-attentive Process	Top–down, Attentive Process
High perceptual noise	Low perceptual noise
Very short memory (iconic memory)	Relative long memory (Working Memory)
All hypothesis simultaneously	Only a highly limited set of hypotheses
Heuristic search for heat-seeking direction	Top–down rational inference
Null	Optimally switching the tracked hypotheses given the posterior of the hypotheses
Null	Concentrating more resource on the same hypothesis to further boost precision and memory
. . . cannot evoke a certain response by assigning a high posterior probability to a hypothesis	. . .slow, but can confirm a hypothesis with a high posterior probability

We provided direct evidence for a super-additive effect of integrating bottom–up and top–down processes with an efficient "scheduling" mechanism. Simulation results of Experiment 3 show that information extracted by the pre-attentive, bottom–up process alone is weak, as it can only achieve accuracy below 10%. Similarly, an attentive process tracking only four agents (two wolf–sheep pairs) is likewise incapable of searching through a large hypothesis space, also yielding accuracy below 10%. Nevertheless, the combination of these two processes produces accuracy four times higher than the sum of each process alone. Such high performance is primarily due to the complementary strengths of these two processes. The bottom-up process is inflexible and noisy, but parallel and fast. The attentional process is flexible, rationally integrating the wolf's possible strategies, but is serial, capacity-limited and computationally expensive. By a highly efficient mechanism for allocating attention, the strength of these two processes can be combined to overcome their individual weaknesses, making perception both fast and flexible.

Comparing human results with simulation results from various models was critical for us to reach the above conclusions. It is interesting to note that our behavioral results are consistent with recent studies (Meyerhoff et al., 2013, 2014a), showing that search performance drops as set-size increases. However, the theoretical focus of the current study is largely different. These results suggest that attention is necessary for perceived animacy. In contrast, here we argue that attention alone is not sufficient, and show how it can be supported by a noisy yet effective pre-attentive process. This conclusion is in line with the goal of this study, which is not to seek for a straightforward answer regarding whether or not attention is required for perceived animacy. The importance of our work is that we propose a model of attention, showing how it works by interacting with pre-attentive process and working memory at the algorithmic level.

6.2. Scheduling limited cognitive resources

The super-additive effect we observed here sheds light on a solution to a paradox revealed by decades of studies of human vision: on the one hand, human vision is remarkably fast and rich, including deep understanding of objects, causality, physics and animacy. On the other hand, it is also highly capacity-limited, constrained by the amount of information that can be processed by attention and working memory. To some extent, revealing these capacity constraints makes an algorithmic understanding of human vision even more challenging. The model not only needs to produce outputs as rich as human perception, but it must do so with very limited computational budget. Therefore, investigation of limited cognitive resources should not only focus on the "capacity limitation" itself but should ultimately lead to a "cost-sensitive" model that can achieve high-performance with a minimized budget for computation.

Our modeling results indicate that at least part of the solution to the paradox is to integrate top–down and bottom–up processes with a scheduling mechanism that can efficiently allocate limited attention. One should not take bottom–up and top–down processes' super-additive effect for granted. It critically depends on how attention is allocated among competing hypotheses. A sub-optimal scheduling mechanism can easily

make a model both inflexible and computationally expensive. To enhance the model's performance, we design a scheduling mechanism based on a stochastic sampling algorithm. This process is optimal in the sense that the probability of selecting a wolf–sheep pair is determined by the posterior probability of that hypothesis. This idea is inspired by the particle filter method for modeling time series data (Gordon, Salmond, & Smith, 1993). However, in traditional particle filtering, the perceptual input is independent of the hypotheses tracked by the particles. The key difference of our model is the two-way interaction between perceptual inputs and the attentively tracked hypotheses. The perceived motion directions impact the wolf-sheep posterior, from which the tracked hypotheses are sampled. With the attentional resources, the tracked hypotheses are then processed with higher perceptual precision and longer memory, which allows the model to quickly accept or reject the tracked hypotheses. Once the posterior of a tracked hypothesis becomes higher, it will attract more resources, further improving the perceptual quality. This recursive process can lead all attentional resources to quickly focus on a single hypothesis. One interesting property of this process is that the concentration of attentional resources emerges from a simple resampling rule without explicitly distinguishing different stages of attention, for example, spreading versus focal attention.

6.3. Attention as a window to higher-level inference

We argue attention research can be divided into two categories. One is how to schedule attention, as we have discussed in the previous section. The other part is what computation is executed once attention is allocated. In our computational architecture, the computation executed with attention is qualitatively different from pre-attention, which is based on the processing of fixed features. The attention process is much more flexible, reflecting a rational process that can integrate contextual information, adapting to agent motions that can vary across scenes. In our experiment, we modeled the precision of the wolf's motion with the chasing subtlety. The cognitive architecture proposed here can certainly allow us to expand the top–down components to model other possible variations of an agent's motion, without changing the cognitive architecture at all. Other possible sources of the agent's flexible behavior include obstacles in the environment (Baker et al., 2009; Gergely et al., 1995), what it can and cannot perceive, how energetic it is, and even how "smart" it is. All these factors can make the goal-directed behavior vary across agents and across scenes.

6.4. Integrating object-based and relation-based processes

Recent studies suggested that the search-for-chase is object-based (e.g., Meyerhoff et al., 2014a). In addition, in perceived chasing, only an individual object can be perceived as a social agent (van Buren, Gao, & Scholl, 2014). This object-based assumption is an important component of the cognitive architecture modeled in our study. First, in our models, the hypothesis space of chasing is built upon *individual objects*. It assumes that if a patch of visual stimuli cannot be perceived as an individual object, it won't even

be included in the space of possible social interactions. Second, when tracking a specific wolf-sheep hypothesis, our models need to "ground" that hypothesis into the visual display by tracking two individual objects. Third, our models assume that the fidelity of perception and memory can be further enhanced by deploying more attention on the same pair of objects. This quantitative model of attention is largely inspired by studies of multiple-object tracking (e.g., Franconeri, Jonathan, & Scimeca, 2010) and visual working memory of objects (e.g., Zhang & Luck, 2008).

In the meanwhile, we suggest that an exclusive object-based process is insufficient to support the perception of social interactions. Imagine how to build a chasing detection model with just object-based representations, such as "object files" (Kahneman, Treisman, & Gibbs, 1992). To evaluate the hypothesis that one object (A) is the wolf, the model needs to encode its motion relative to the position of a possible sheep (B). But how to represent B's position and motion? Given the principle of object-based processing, they should be stored in A's object file, as A is the target object the model is evaluating. However, it is obvious that B's position and motions do not belong to A. So the model now faces a dilemma: to enforce an object-based evaluation of chasing, it must store attributes of one object into another object's file. This will cause an inter-dependence between two object files, which are supposed to be independent. This only scratches the surface of the challenge. To evaluate the probability that object A is a wolf, the model actually needs to evaluate the possibilities that A is chasing every possible sheep, and then add these probabilities together. It indicates that A's object file must encode all other objects' positions and motions. As a result, one object's attributes must be copied multiple times, distributed among all other objects. This certainly breaks the "encapsulation" principle that object-based representation is supposed to enforce. Implementing this object-based model will be very complicated. We can hardly believe that it captures the architecture of the human mind.

To address the above challenges, we argue that Heider & Simmel–like social displays can be best encoded as a "social network" including both nodes and edges. In this network, a node represents an individual object, while an "edge" represents a social interaction between two objects connected by that edge. Chasing is simply a special social network limited to just one type of pairwise interaction. These edges are explicitly plotted in our online demos. By pushing all inter-object interactions to edges, our model can create clean "object files" that only contain attributes entirely owned by an object. Therefore, principles of object-based representation are actually better enforced in our model. Computation is also greatly simplified. Attributes of an object are stored only once within each object file. In the meanwhile, each inter-object interaction is also stored only once on each edge. In other words, by using both object-based and relation-based representations, information is much better encapsulated in our model.

6.5. Explaining other phenomena of perceived chasing

Here we discuss implications of our model by connecting it to other psychophysical results of perceived chasing reported in other studies. We focus on two interesting phenomena: search asymmetry and a linear effect of set-size.

6.5.1. Search asymmetry

In visual search, a search asymmetry is said to occur when a search for stimulus A among stimulus B produces different results from a search for B among A (Wolfe, 2001). In the context of perceived chasing, however, it is defined as a search for the wolf given the identify of sheep is more efficient than a search for the sheep given the identity of the wolf (Meyerhoff et al., 2014a).

The search asymmetry is not reflected in our human results, as both the wolf and sheep are unknown in our tasks. However, our model can synthesize a search asymmetry effect by turning on and off the wolf or sheep component in the probability inference. In our model, the likelihood of a chasing hypothesis is the product of likelihoods of both the wolf and sheep's motions given that hypothesis (see Eqs. 3–5 in Supplementary Material file). However, this does not imply that their motions contribute equally to chasing detection. The exact contribution of an agent is determined by the output of its likelihood function, and those of competing hypotheses. To test this prediction, we ran a simulation by manipulating the likelihood function of the Hybrid model. In the "Wolf and Sheep" condition, the chasing likelihood is the product of the wolf likelihood and the sheep likelihood. In the "Only Wolf" condition, the wolf likelihood is used for chasing detection while the sheep's motion was ignored. In the "Only Sheep" condition, only the sheep likelihood is used. In practice, only one line of code corresponding to Eq. 5 was changed.

We tested the above variations of the Hybrid model by using displays from Experiment 3, with subtlety 5° and set size 12. The results were illustrated in Fig. 14. There are two important discoveries. First, as predicted, search based on the wolf's motion (27.4%) was much easier than search based on the sheep's motion (2.6%), demonstrating a type of search asymmetry. Second, accuracy of "Wolf and Sheep" (67.6%) was much higher than the combined accuracy of "Only Wolf" and "Only Sheep." It shows that while it was

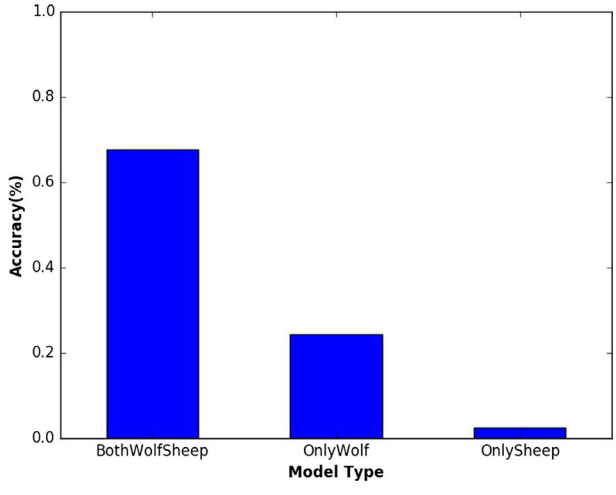

Fig. 14. A simulated search asymmetry. It is easier to detect chasing by using the wolf's motion than using the sheep's motion. In addition, there is a super-additive effect when both the wolf and sheep's motions are evaluated by the model.

almost impossible to detect chasing through the sheep's motion alone, adding it together with the wolf's motion produced a huge super-additive effect. This result can be explained intuitively. In a crowded display, at every moment, there can be several distractors which happen to move toward other distractors. Attention cannot reject these coincidences by just evaluating the wolf's motion direction. When attention jointly evaluates the wolf and sheep's motions, it can only be distracted when a distractor (A) happens to move toward another distractor (B) while B also happens to move away from A. The probability of such joint coincidences drops dramatically. Therefore, attention can quickly reject an incorrect chasing hypothesis and switch to the correct one. To summarize, with different likelihood functions, our model shows (a) how a search asymmetry naturally arises by searching over a space of possible wolf–sheep pairs; and (b) why it is computationally desirable to evaluate both the wolf and sheep, rather than one agent alone.

6.5.2. A linear effect of set-size and beyond

Previous work also demonstrated a linear relationship between set-size and search performance (Meyerhoff et al., 2014a).This result is consistent with predictions of our model. In Experiment 1, when chasing subtlety was 5°, both human results and model results showed a linear effect of set size in accuracy (Figs. 5A and 8A) and RT (Figs. 5B and 8B). In addition, in Experiment 3 with much larger set sizes, human RT was no longer a linear function of set size. Our model captured this non-linear effect of set size as well (Fig. 11B). Based on these results, we argue that a linear relationship between search performance and set size does not necessarily reflect a linear search of each individual object in the search display. Instead, a linear effect with small set sizes may reflect the interaction between a parallel pre-attentive search (which is a constant function of set size) and a serial search of wolf–sheep pair (which is a quadratic function of set size).

In addition to the heat-seeking direction explicitly modeled here, other motion cues may also impact the performance of search-for-chase, such as spatial proximity (e.g., Roux, Passerieux, & Ramus, 2013; Meyerhoff, Schwan & Huff, 2014b). Recent studies have shown that reduced inter-object spacing guides visual attention and eye movements in dynamic multi-objects displays (Galazka & Nyström, 2016; Meyerhoff, Schwan & Huff, 2018; Zelinsky & Todor, 2010). Local density determined by spatial proximity may be a plausible candidate to generate a priority map for serial visual attention in the guided search model.[1] Future research is required to reveal how objects and their interactions are processed by pre-attention and attention based on different types of visual features.

7. Conclusion

We propose a cognitive architecture for perceived animacy, in which a pre-attentive process, attention, and working memory are integrated into a probabilistic inference framework. Our model shows how the psychophysics of chasing can be explained by a Bayesian ideal observer with cognitively realistic capacity constraints. The cognitive architecture reveals how perceived animacy is deeply rooted in core cognitive mechanisms, and how it can be connected to higher-level cognitive inference.

Acknowledgments

This research was supported by ONR MURI N00014-16-1-2007 awarded to Tao Gao.

Note

1. We thank Dr. Meyerhoff for pointing out this alternative account of attention priority map in search-for-chasing.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*, 457–464.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829–839.

Baker, C. L., Goodman, N., & Tenenbaum, J. B. (2008). Theory-based social goal inference. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society* (pp. 1447–1452).

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Baker, C. L., Saxe, R. R., & Tenenbaum, J. B. (2011). Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Thirty-Third Annual Conference of the Cognitive Science Society* (pp. 2469–2474).

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*, 313–331.

Bishop, C. M. (2006). *Machine learning and pattern recognition*. New York: Springer.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*, 85–109.

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, *62*, 73–101.

Chun, M. M., & Potter, M. C. (1995). A two-stage model for multiple target detection in rapid serial visual presentation. *Journal of Experimental psychology: Human Perception and Performance*, *21*, 109–127.

Dennett, D. C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.

Di Lollo, V., Enns, J. T., & Rensink, R. A. (2000). Competition for consciousness among visual events: The psychophysics of reentrant visual processes. *Journal of Experimental Psychology, General*, *129*, 481–507.

Dick, A. O. (1974). Iconic memory and its relation to perceptual processing and other memory mechanisms. *Perception & Psychophysics*, *16*, 575–596.

Dittrich, W. H., & Lea, S. E. (1994). Visual perception of intentional motion. *Perception*, *23*, 253–268.

Franconeri, S. L., Jonathan, S. V., & Scimeca, J. M. (2010). Tracking multiple objects is limited only by object spacing, not by speed, time, or capacity. *Psychological Science*, *21*, 920–925.

Galazka, M., & Nyström, P. (2016). Visual attention to dynamic spatial relations in infants and adults. *Infancy*, *21*, 90–103.

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*, 154–179.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 669–684.

Gao, T., Scholl, B. J., & McCarthy, G. (2012). Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *Journal of Neuroscience*, *32*, 14276–14280.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165–193.

Gordon, N. J., Salmond, D. J., & Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *IEE Proceedings F Radar and Signal Processing*, *140*, 107–113.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*, 243–259.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*, 175–219.

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. New York: Cambridge University Press.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*, 279–281.

Meyerhoff, H. S., Huff, M., & Schwan, S. (2013). Linking perceptual animacy to attention: Evidence from the chasing detection paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 1003–1015.

Meyerhoff, H. S., Schwan, S., & Huff, M. (2014a). Perceptual animacy: Visual search for chasing objects among distractors. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 702–717.

Meyerhoff, H. S., Schwan, S., & Huff, M. (2014b). Interobject spacing explains the attentional bias toward interacting objects. *Psychonomic Bulletin & Review*, *21*, 412–417.

Meyerhoff, H. S., Schwan, S., & Huff, M. (2018). Oculomotion mediates attentional guidance toward temporarily close objects. *Visual Cognition*, *26*, 166–178.

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 16598–16603.

Pantelis, P. C., Baker, C. L., Cholewiak, S. A., Sanik, K., Weinstein, A., Wu, C. C., Tenenbaum, J. B., & Feldman, J. (2014). Inferring the intentional states of autonomous virtual agents. *Cognition*, *130*, 360–379.

Pantelis, P. C., & Feldman, J. (2012). Exploring the mental space of autonomous intentional agents. *Attention, Perception, & Psychophysics*, *74*, 239–249.

Papenmeier, F., Huff, M., & Schwan, S. (2012). Representation of dynamic spatial configurations in visual short-term memory. *Attention, Perception, & Psychophysics*, *74*, 397–415.

Pratt, J., Radulescu, P. V., Guo, R. M., & Abrams, R. A. (2010). It's alive!: Animate motion captures visual attention. *Psychological Science*, *21*, 1724–1730.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*, 179–197.

Roux, P., Passerieux, C., & Ramus, F. (2013). Kinematics matters: A new eye-tracking investigation of animated triangles. *The Quarterly Journal of Experimental Psychology*, *66*, 229–244.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition*, *80*, 159–177.

Shooner, C., Tripathy, S. P., Bedell, H. E., & Ogmen, H. (2010). High-capacity, transient retention of direction-of-motion information for multiple moving objects. *Journal of Vision*, *10*, 8.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, *74*, 1–29.

Srivastava, N., & Vul, E. (2015). Attention dynamics in multiple object tracking. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society* (Vol. 22, pp. 1955–1963).

Sun, Z., Hung, Y., Yu, W., Zhang, M., Shui, R., & Gao, T. (2015). How to break the representation of moving objects? The geometric invariance of dynamic configuration in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *41*, 1247–1259.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*, 97–136.

Tremoulet, P. D., & Feldman, J. (2000). Perception of animacy from the motion of a single object. *Perception*, *29*, 943–951.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., & Tenenbaum, J. B. (2010). Help or hinder: Bayesian models of social goal inference. *Advances in Neural Information Processing Systems*, *22*, 1874–1882.

Van Buren, B., & Scholl, B. J. (2014). Perceived animacy influences other forms of visual processing: Improved sensitivity to the orientations of intentionally moving objects. Poster presented at the annual meeting of the Vision Sciences Society, St. Pete Beach, FL.

Wick, F. A., Alaoui Soce, A., Garg, S., Grace, R. C., & Wolfe, J. M. (2019). Perception in dynamic scenes: What is your Heider capacity? *Journal of Experimental Psychology: General*, *148*, 252–271.

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238.

Wolfe, J. M. (2001). Asymmetries in visual search: An introduction. *Perception & Psychophysics*, *63*, 381–389.

Zelinsky, G. J., & Todor, A. (2010). The role of "rescue saccades" in tracking objects through occlusions. *Journal of Vision*, *10*, 29.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature*, *453*, 233–235.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Fig. S1. (a) The accuracy of the Pure-Attention model with every combination of precision (4, 6, 8) and capacity (2, 4, 8, 12, 20 slots) in Experiment 1. (b) The RT of the Pure-Attention model with every combination of precision (4, 6, 8) and capacity (2, 4, 8, 12, 20 slots) in Experiment 1.

Fig. S2. (a) The accuracy of the Hybrid model with every combination of the parallel precision (2, 2.5, 3.0) and memory decay rate (0.4, 0.45, 0.5) in Experiment 2. (b) The RT of the Hybrid model with every combination of the parallel precision (2, 2.5, 3.0) and memory decay rate (0.4, 0.45, 0.5) in Experiment 2.

Appendix S1. Mathematical formulations of the ideal observer, pure attention, and hybrid models of chasing detection.