Computational Brain & Behavior manuscript no.

Pragmatic inference and visual abstraction enable contextual flexibility during visual communication

Judith E. Fan · Robert D. Hawkins · Mike Wu · Noah D. Goodman

Received: date / Accepted: date

Abstract Visual modes of communication are ubiquitous in modern life — from maps to data plots to political cartoons. Here we investigate drawing, the most basic form of visual communication. Participants were paired in an online environment to play a drawing-based reference game. On each trial, both participants were shown the same four objects, but in different locations. The sketcher's goal was to draw one of these objects so that the viewer could select it from the array. On 'close' trials, objects belonged to the same basic-level category, whereas on 'far' trials objects belonged to different categories. We found that people exploited shared information to efficiently communicate about the target object: on far trials, sketchers achieved high recognition accuracy while applying fewer strokes, using less ink, and spending less time on their drawings than on close trials. We hypothesized that humans succeed in this task by recruiting two core faculties: visual abstraction, the ability to perceive the correspondence between an object and a drawing of it; and pragmatic inference, the ability to judge what information would help a viewer distinguish the target from distractors. To evaluate this hypothesis, we developed a computational model of the sketcher that embodied both faculties, instantiated as a deep convolutional neural network nested within a probabilistic program. We found that this model fit human data well and outperformed lesioned variants. Together, this work provides the first algorithmically explicit theory of how visual perception and social cognition jointly support contextual flexibility in visual communication.

Keywords: drawing; social cognition; perception; deep learning; probabilistic models

Judith Fan
Department of Psychology, University of California San Diego, 9500 Gilman Drive MC 0109, La Jolla, CA 92093, E-mail: jefan@ucsd.edu

Robert Hawkins
Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305

Mike Wu
Department of Computer Science, Stanford University, 353 Serra Mall, Stanford, CA 94305

Noah Goodman
Department of Psychology, Stanford University, 450 Serra Mall, Stanford, CA 94305

Introduction

From ancient etchings on cave walls to modern digital displays, the ability to externalize our thoughts in visual form lies at the heart of key human innovations (e.g., painting, cartography, data visualization), and forms the foundation for the cultural transmission of knowledge (Tomasello, 2009; Donald, 1991). Perhaps the most basic and versatile visualization technique is drawing, the earliest examples of which date to at least 40,000 years ago (Hoffmann, Standish, García-Diez, Pettitt, Milton, Zilhão, Alcolea-González, Cantalejo-Duarte, Collado, De Balbín et al., 2018; Aubert, Brumm, Ramli, Sutikna, Saptomo, Hakim, Morwood, van den Bergh, Kinsley, and Dosseto, 2014), and which can yield images ranging from photorealistic renderings to schematic diagrams. Even in the simple case of sketching an object in the world, there are countless ways of depicting that object, depending on the context. For instance, an automotive engineer formulating a new car design may invest considerable effort to produce detailed drawings that convey fine-grained information about the car's body shape, which impacts how aerodynamic it will be. On the other hand, a cartoonist drawing a street scene may only need a few strokes to sketch a car that conveys the location of the scene. How do drawings, despite spanning such a broad range of appearances, reliably convey their intended meaning?

On the one hand, recent work in computational vision suggests that the identity of an object depicted in a drawing can be derived from its visual properties alone (Fan, Yamins, and Turk-Browne, 2018). These results are consistent with evidence from other domains, including developmental, cross-cultural, and comparative studies of drawing perception. For example, human infants (Hochberg and Brooks, 1962), people living in remote regions without pictorial art traditions and without substantial contact with Western visual media (Kennedy and Ross, 1975), and higher non-human primates (Tanaka, 2007) are able to recognize line drawings of familiar objects, even without prior experience with drawings. Together, these findings suggest that the ability to perceive the correspondence between drawings and real-world objects arises from a general-purpose neural architecture evolved to handle variation in natural visual inputs (Sayim, 2011; Gibson, 1979).

On the other hand, influential work in philosophy has emphasized the role of cultural and social context in determining how drawings denote objects (Goodman, 1976). This perspective is consistent with substantial variation in pictorial art traditions across cultures (Gombrich, 1989, 1969) and the existence of culturally-specific conventions for encoding meaning in pictorial form (Boltz, 1994; Allen, 2000). Further support for the importance of social context has also come from recent laboratory studies of visual communication, which have found that pairs of interacting participants can produce drawings that are referentially meaningful to their partner in context, even when these drawings do not strongly resemble any particular real-world referent out of context (Garrod, Fay, Lee, Oberlander, and MacLeod, 2007; Fay, Garrod, Roberts, and Swoboda, 2010; Galantucci, 2005).

Towards reconciling these perspectives, the current paper explores the hypothesis that visual information and social context jointly determine how drawings convey meaning. To evaluate this hypothesis, we investigated how the drawings people produce varied across communicative contexts, and found that people adapted their drawings accordingly, producing detailed drawings when necessary, but simpler drawings when sufficient. To explain these findings, we developed a computational model of visual communication that embodied two core faculties: visual abstraction, the capacity to judge the correspondence between a real-world object and a drawing of it; and pragmatic inference, the ability to judge what information is not only *valid* to include in a drawing, but also *relevant* in context (Goodman and Frank, 2016; Grice, Cole, and Morgan, 1975; Abell, 2009). This model was instantiated as a deep convolutional neural network visual encoder nested within a probabilistic program that inferred which drawings would be most informative in context. We found that our full model fit the data well and outperformed lesioned

variants, providing a first algorithmically explicit theory of how visual perception and social cognition jointly support contextual flexibility in visual communication.

Results

Effect of context manipulation on communication task performance

To investigate visual communication in a naturalistic yet controlled setting, we employ a drawing-based reference game paradigm. This reference game involves two players: a *sketcher* who aims to help a *viewer* pick out a target object from an array of distractor objects by representing it in a sketch. Such games, which have long provided a source for intuitions in the philosophy of language (Wittgenstein, 1953; Lewis, 1969), have also proven to be a valuable experimental tool for systematically eliciting pragmatic inferences about language use in context (Goodman and Frank, 2016; Kao, Bergen, and Goodman, 2014; Goodman and Stuhlmüller, 2013; Frank and Goodman, 2012), especially the ability of speakers to compose utterances that are informative (Grice, Cole, and Morgan, 1975; Wilson and Sperber, 1986) yet parsimonious (Zipf, 1936) during verbal communication. Here we generalize this methodology to understand the role of pragmatic inference during visual commmunication.

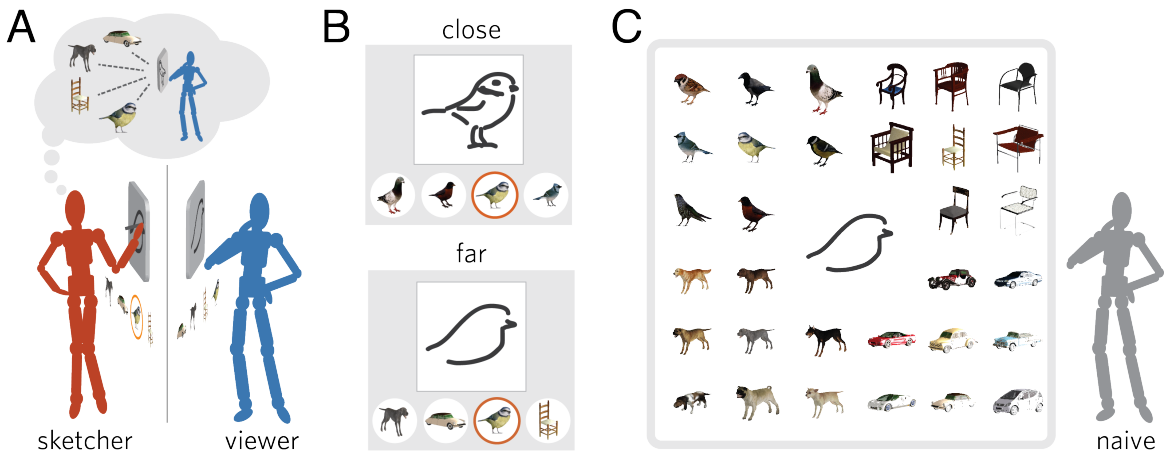

Fig. 1 (A) Communication task. Participants were paired in an online environment to play a drawing-based reference game and assigned the roles of sketcher and viewer. On each trial, the sketcher's goal was to draw one of these objects so that the viewer could distinguish the target from three distractor objects. (B) Context manipulation. Distractor similarity to target was manipulated across two context conditions: in close contexts, the target and distractors belonged to the same basic-level category, while in far contexts, the target and distractors all belonged to different basic-level categories. (C) Recognition task. Naive participants were presented with a randomly sampled sketch from the communication experiment and an array containing all 32 objects used in the experiment, and were instructed to identify the best-matching object.

In our experiment, participants (N=192) were paired in an online environment and communicated with their partner only via a drawing canvas (Fig. 1A). Each trial, both participants were shown a set of four real-world objects, but object locations were randomized for each participant so that they could not use object location information to solve the task. The sketcher's goal on each trial was to draw one of these objects — the target — so that the viewer could pick it out from the array. There were 32 objects in total belonging to four basic-level categories (i.e., bird, car, chair, dog), that were rendered in the same three-quarter pose, under identical illumination, and on a gray background, so participants could not use pose, illumination, or background information to distinguish them. Each object was randomly assigned to exactly one set of four objects, and each set of four objects was presented four times each, such that each object served as the target exactly once. Across trials, the similarity of the

distractors to the target was manipulated, yielding two types of communicative context that appeared in a randomly interleaved order: close contexts, where the target and distractors all belonged to the same basic-level category, and far contexts, where the target and distractors belonged to different basic-level categories (Fig. 1B). We predicted that while sketchers would be generally successful at conveying the identity of the target, their sketching behavior would systematically differ between the two contexts. Specifically, we predicted that sketchers would invest more time and ink in producing their sketches in close contexts, but still produce sufficiently informative sketches with less time and ink in far contexts (Fig. 2).

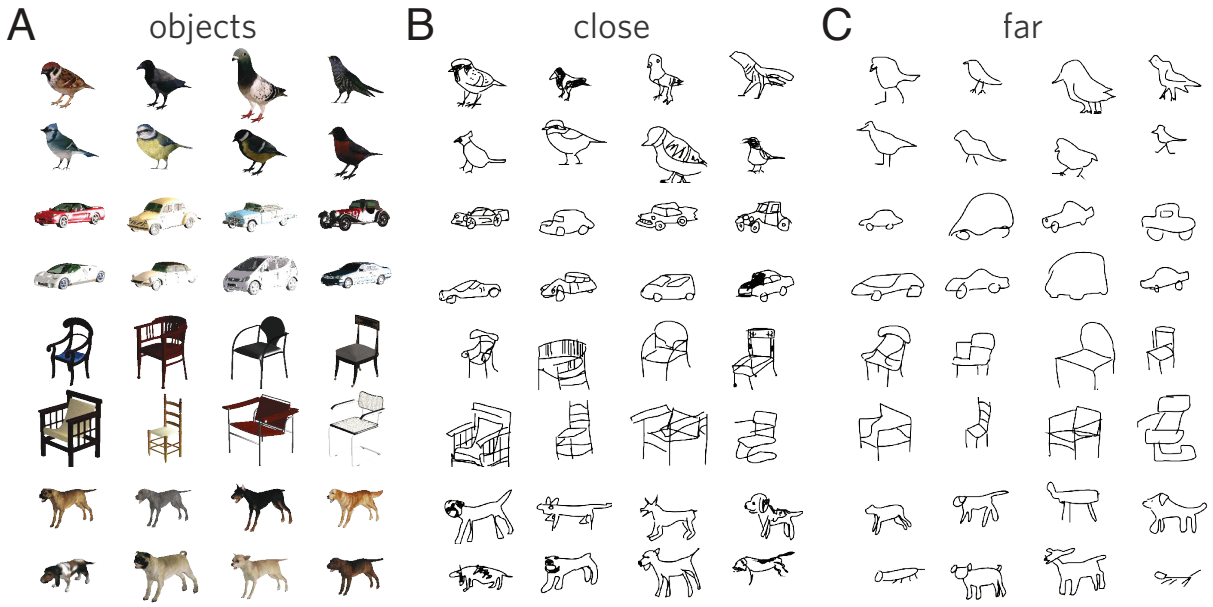

Fig. 2 (A) Object stimuli. (B) Example sketches produced in close context condition. (C) Example sketches produced in far context condition.

Consistent with our prediction, we found that viewers were highly accurate overall at identifying the target from the sketches produced (proportion correct: 93.8%, 95% CI: [92.7%, 94.8%], estimated by bootstrap resampling participants). Moreover, we found that sketchers spent less time (far: 13.7s, close: 30.3s, $p<0.001$), applied fewer strokes (far: 13.5, close: 8.03, 95% CI for difference: [3.75, 7.90], $p<0.001$), and used less ink (proportion of canvas filled; far: 0.042, close: 0.054, 95% CI for difference: [0.01, 0.014], $p<0.001$) to produce their sketches in the far condition than in the close condition (Fig. 3A-C). Despite the relative sparsity of sketches in the far condition, viewers were near ceiling at identifying the target on these trials (far: 99.7%, 95% CI: [0.993, 0.999]; close: 87.9%, 95% CI: [0.858, 0.899], Fig. 3D), and took less time to make these decisions than on close trials (far: 6.32s, close: 8.32s; 95% CI for difference: [1.251, 2.748], Fig. 3E).

Effect of context manipulation on sketch recognizability

A natural explanation for these findings is that the two context conditions differed in how much information was required to identify the target. Specifically, sketchers invested greater time and ink in close contexts to strengthen the correspondence between their sketch and the target object, out of necessity, while they could still succeed in far contexts with sketches that were less costly to produce. To evaluate this possibility, we recruited another group of naive participants (N=112) to perform a sketch recognition task that yielded estimates of how strongly each sketch corresponded to every object in the communication

experiment. On each trial of this recognition experiment, participants were presented with a sketch and an array containing all 32 objects, and were instructed to identify the object that best matched each sketch from the array (Fig. 1C). Across trials, sketches were randomly sampled from the original communication experiment such that no two sketches produced by the same participant appeared in a single recognition experimental session. Consistent with our hypothesis, we found that close sketches were matched with their corresponding target object more consistently than far sketches were (close: 54.2%; far: 37.5%; Z=14.1, p <0.001), although sketches from both context conditions were successfully matched at rates greatly exceeding chance (ps < 0.001).

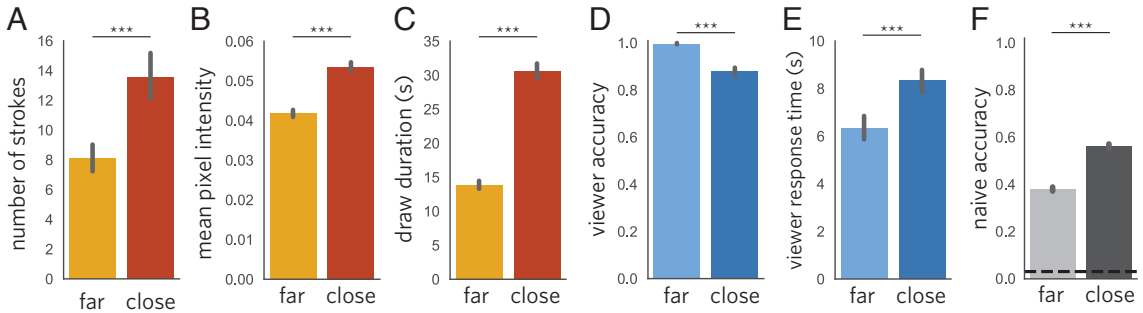

Fig. 3 (A-C) Mean number of strokes, amount of ink, and time spent producing sketches in each context condition. (D-E) Target identification performance in context during communication task. (F) Target identification performance out of context during recognition task. Error bars reflect 95% confidence intervals.

Computational model of contextual flexibility in visual communication

Our empirical findings suggest that sketchers spontaneously modulate the amount of information they convey about the target object according to the communicative context. Such contextual flexibility argues against the notion that visual communication is constrained exclusively by the appearance of the target object, but instead that it is systematically influenced by contextual information that is shared between the sketcher and viewer. Moreover, it suggests an analogy to how shared context influences what people choose to say during verbal communication, a key target of recent advances in computational models of pragmatic inference in language use (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Franke and Jäger, 2016; Bergen, Levy, and Goodman, 2016). Leveraging these advances, we propose that human sketchers determine what kind of sketch to produce in context by deploying two main faculties: *visual abstraction*, which here refers to the ability to judge how well a sketch evokes a real object, and *pragmatic inference*, which here refers to the ability to judge which sketches will be sufficiently detailed to be informative about the target object in context, but no more detailed than necessary. To test this proposal, we developed a computational model of the sketcher that embodies both visual abstraction and pragmatic inference, and was instantiated as a deep convolutional neural network nested within a probabilistic program. Constructing such a model allowed us to use formal model comparison to evaluate the contribution of each component for explaining our empirical findings, as well as make quantitative predictions about visual communication behavior in novel contexts.

Defining communicative utility of sketches

We define the sketcher, $\mathcal{S}$, to be a decision-theoretic agent that produces sketches, s, of the target proportional to their communicative utility, which is a function of a sketch and a context: $U(s,O)$. In our experiment, a context is defined as: $O = \{t, D\}$, where t is the target object and D is a set of three distractor objects, $D = \{d_1, d_2, d_3\}$. When deciding which sketch to produce, the utilities of each sketch are assumed to be normalized over the set of producible sketches via the softmax function:

$$\mathcal{S}(s|O) = \frac{\exp[U(s,O)]}{\sum_i \exp[U(s_i,O)]} \tag{1}$$

In principle, the space of all producible sketches is infinite and continuous, leading to an intractable sum. In practice, we assume that the sketcher model chooses among a large but finite set of sketches: those actually produced by participants in our experiment.

We first introduce the utility function for our proposed pragmatic sketcher, S_{prag}, and then consider lesioned variants for comparison. This utility function formalizes the notion of pragmatic inference as a balance between how informative a more detailed sketch would be in context with how costly it would be to produce such a detailed sketch. It consists of two terms: an informativity term and a cost term.

The sketcher judges a sketch's informativity to be a mixture of two quantities: one reflecting its absolute *resemblance* to the target and the other its relative *diagnosticity* in the presence of particular distractors (see Fig. 4B). Resemblance is determined by how strongly a sketch s corresponds to the target object t, i.e. $\text{sim}(s,t)$, which inherently relies on some form of visual abstraction in order to compute. In our first set of modeling experiments, we estimate $\text{sim}(s,t)$ empirically as the proportion of trials in the recognition experiment on which the target object was matched to the sketch. Later, we present a model of visual abstraction, instantiated as a neural network capable of predicting $\text{sim}(s,t)$ on heldout data.

Diagnosticity is determined by how strongly a sketch s corresponds to the target object t, *relative* to the other objects in context. Owing to its dependence upon contextual information, diagnosticity relies on both pragmatic inference and visual abstraction to compute. Formally, diagnosticity is defined by the natural log probability that a simulated viewer agent, $\mathcal{V}$, would select the target object given the sketch and all objects in context, $\ln \mathcal{V}(t|s,O)$. The simulated viewer $\mathcal{V}$, in turn, is assumed to select the target object proportional to the correspondence between the sketch and the target, $\text{sim}(s,t)$, normalized by the sum of correspondences between the sketch and all four objects in context, again via the softmax function:

$$\mathcal{V}(t|s,O) \propto \frac{\exp\{\alpha \cdot \text{sim(s, t)}\}}{\sum_{i=1}^{4} \exp\{\alpha \cdot \text{sim}(s,o_i)\}} \tag{2}$$

Here, i indexes each object $o \in O$, and α is a scaling parameter that determines how strongly the simulated viewer's decision policy favors the highest-utility sketch. As $\alpha \to \infty$, the simulated viewer is more likely to choose the object with highest perceptual correspondence to the sketch. Intuitively, this means that the viewer is more likely to pick the correct object when the sketch corresponds more strongly to the target than to the distractors.

To combine the resemblance and diagnosticity terms into a single informativity value, we introduce a weight parameter, w_d, that interpolates between them:

$$I(s,O) = w_d \cdot \ln \mathcal{V}(t|s,O) + (1 - w_d) \cdot \text{sim}(s,t). \tag{3}$$

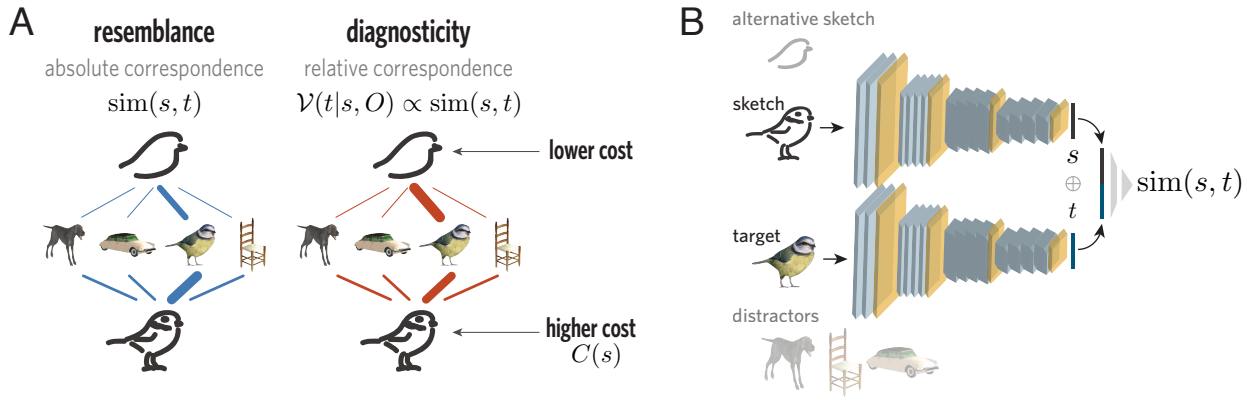

Fig. 4 (A) Schematic containing an example context and two candidate sketches under consideration by the sketcher, $\mathcal{S}$. The thickness of each blue line reflects the absolute strength of the correspondence between a candidate sketch and object in context (resemblance). The thickness of each red line reflects the relative strength of the correspondence between a candidate sketch and each object, compared to its correspondence to the other objects in the context (diagnosticity). A sketch's informativity was hypothesized to depend on both its resemblance and diagnosticity. The sketcher expects the viewer, $\mathcal{V}$, given the sketch and context, to select the target object proportional to the strength of the correspondence between the sketch and target object. All else equal, the sketcher is assumed to prefer sketches that are less costly to produce. (B) Architecture of the visual encoder used to predict the correspondence between sketches and objects, which consists of a base convolutional neural network and fully connected "adaptor" neural network. The parameters of the base neural network are trained on separate data and frozen, whereas the parameters of the adaptor neural network are trained on subsets of our experimental data. First, two identical branches of the base neural network are applied to a sketch and object to extract a feature vector for each image. Next, these feature vectors are concatenated and passed through the adaptor neural network to yield a sketch-object correspondence score.

Combining these terms captures the intuition that a communicative sketcher seeks to produce a sketch that both resembles the target object and distinguishes the target from the distractors.

Finally, we define a sketch's cost, $C(s)$, to be a monotonic function of the amount of time taken to produce it, linearly transformed to lie in the range $[0,1]$. Putting these terms together, we have the full utility:

$$U_{S_{prag}}(s,O) = w_i \cdot I(s,O) - w_c \cdot C(s) \tag{4}$$

where w_i and w_c are independent, nonnegative scaling parameters that are applied to the informativity and cost terms, respectively. These parameters determine how strongly each term contributes to the overall utility of the sketch. This model contains four latent parameters: one each on informativity (w_i) and cost (w_c), one that balances between diagnosticity and resemblance within the informativity term (w_d), and one that tracks the optimality of the simulated viewer's decision policy (α). We inferred these parameters from our data via Bayesian data analysis (see Materials and Methods).

Evaluating contribution of pragmatic inference

We hypothesized that a pragmatic sketcher model that is sensitive to both context and cost would provide a strong fit to human sketch production behavior, as well as outperform lesioned alternatives lacking either component. To test this hypothesis, we compare the full pragmatic model, (S_{prag}), with two nested variants with different utility functions: a *context-insensitive* sketcher, S_{sim}, in which the diagnosticity term is removed (i.e., w_d=0), leaving only the resemblance component in the informativity term; and a *cost-insensitive* sketcher, S_{nocost}, in which the cost term is removed (i.e., $w_c = 0$), leaving only the full informativity term.

Our goal was to evaluate how well each model could produce informative sketches and appropriately modulate its behavior according to the context condition, and not necessarily to reproduce exactly the same sketch a particular participant had on a specific trial. As such, we collapsed across all sketches of a given object produced in a given context condition, yielding

64 'prototype' sketches for each object-context category characterized by the average cost and sketch-object correspondence values in that category. For example, the prototypical 'close basset' sketch is characterized by the average cost and object correspondence values across all basset sketches produced in close contexts. Decisions by the sketcher model were generated at the same level of granularity, in the form of a probability distribution over these 64 prototype sketches. To generate these decisions, first we employed Bayesian data analysis to infer a posterior distribution over the four latent parameters in the model (w_i, w_c, w_d, α). Next, we presented each model with exactly the same set of contexts that were presented to human sketchers in the communication experiment, and evaluated the posterior predictive probabilities that each model assigned to sketches in each object-context category, marginalizing over the posterior distribution over latent parameters. We conducted these inference and evaluation steps independently on five balanced splits of the dataset, providing an estimate of reliability and permitting side-by-side comparison with subsequent modeling results using the same splits for crossvalidation (see *Evaluating contribution of visual abstraction*).

We found that the full model, S_{prag}, provided a much better fit to human behavior than the context-insensitive variant, S_{sim} (median log Bayes Factor [BF] across crossvalidation folds = 16.1; see Table 1), and the cost-insensitive variant, S_{nocost} (BF = 9.54). To gain further insight into the functional consequences of each lesion, we investigated three aspects of each model's behavior: (a) *sketch retrieval*: the ability to assign a high absolute rank to the target sketch category in context, out of the 64 object-context alternatives; (b) *context congruity*: the ability to consistently assign a higher rank to the context-congruent version of the target object over the context-incongruent version; and (c) *cost modulation*: how consistently a model produced costlier sketches than average in the close condition, and less costly sketches than average in the far condition, mirroring human behavior.

We found that in general, sketch retrieval performance was high for all three model variants (target rank 95% CI: pragmatic = $[1.43, 1.50]$, context-insensitive = $[1.54, 1.60]$, cost-insensitive = $[1.55, 1.60]$) (Fig. 5A, left). This shows that all three model variants were highly accurate at retrieving a sketch of the correct *object*, with both close and far versions providing a better match than any of the other sketches. However, only the pragmatic sketcher was able to reliably produce the sketch appropriate for the context condition more frequently than would be predicted by chance (95% CI proportion: $[0.571, 0.620]$; Fig. 5B, left); neither the context-insensitive nor the cost-insensitive variants displayed this context congruity (95% CI: context-insensitive = $[0.478, 0.525]$, cost-insensitive = $[0.498, 0.501]$). We observed that the lack of context congruity in the lesioned variants was attributable to an overall bias towards close sketches, which are highly informative in absolute terms, and thus higher in communicative utility if the distractors or sketch cost is ignored.

Moreover, only the pragmatic sketcher produced costlier sketches than average in the close condition (95% CI normalized cost: $[0.205, 0.218]$ vs. grand mean cost = 0.196; Fig. 5C, left), and less costly sketches than average in the far condition (95%CI: $[0.175, 0.180]$). The context-insensitive variant is inherently unable to modulate the cost of the sketches it produces by context condition, and thus was no more or less likely to select a costlier, more diagnostic sketch on a close trial (95% CI: $[0.187, 0.194]$) than a far trial (95% CI: $[0.187, 0.192]$), and preferred slightly less costly sketches overall. While the cost-insensitive variant did exhibit cost modulation by context, because it ignores their cost, it preferred costlier sketches overall in both close (95%CI: $[0.229, 0.241]$) and far contexts (95%CI: $[0.214, 0.220]$). Nevertheless, this cost-insensitive variant still produced consistently costlier sketches in close contexts than in far contexts. This can be understood as having been driven by the remaining diagnosticity component of the informativity term. Because the cost-insensitive variant still places a higher utility on sketches that are highly diagnostic, it is still biased to produce costly but diagnostic sketches in close contexts. By contrast, in far contexts, close and far sketches may be similarly diagnostic, and thus the model produces a mixture of these sketch types.

Together, these results suggest that both context and cost sensitivity are critical for capturing key aspects of contextual flexibility in human visual communication.

Evaluating contribution of visual abstraction

Having established the importance of pragmatic inference, we next sought to evaluate the contribution of visual abstraction. Such evaluation requires an encoding model that captures how abstract perceptual information is extracted from raw visual inputs across successive stages of visual processing. Our approach to modeling visual abstraction is grounded in the neural computations that support robust visual object recognition in higher primates. These computations are carried out by a set of hierarchically organized brain regions known as the ventral visual stream (Malach, Levy, and Hasson, 2002; Rolls, 2001). Across the ventral stream, simple visual features are transformed across successive levels of the hierarchy to support readout of more abstract visual properties (e.g., object identity). Recent work has found deep convolutional neural networks (DCNN), optimized to perform challenging object recognition tasks, to provide a strong computational framework for modeling these computations (Yamins, Hong, Cadieu, Solomon, Seibert, and DiCarlo, 2014; Güçlü and van Gerven, 2015). Specifically, model activations in successive layers of DCNN models have been found to be quantitatively predictive of neural firing patterns in successive regions along the ventral stream (Yamins, Hong, Cadieu, Solomon, Seibert, and DiCarlo, 2014). Especially relevant to the current study, higher-layer representations have also been found to capture more abstract perceptual information in drawings (i.e., intended category) than lower-layer representations (Fan, Yamins, and Turk-Browne, 2018).

Informed by this prior work, we hypothesized that higher-layer representations of a DCNN would provide a stronger basis for predicting human judgments of the perceptual correspondence between a sketch and object than would representations in lower layers. To evaluate this hypothesis, we compared DCNN-based encoding models that varied only in depth, and thus the degree of visual abstraction they achieve prior to the final step of predicting the perceptual correspondence between a sketch and object.

Each encoder variant consists of two functional components: a base visual encoder network, B, and an adaptor network, A: $\mathrm{sim}(s,o) = A(B(s,o))$. For the base encoder, we employ a widely used and high-performing deep convolutional neural network architecture, VGG-19, pretrained to recognize objects from the Imagenet database, whose parameters remain frozen (Simonyan and Zisserman, 2014; Deng, Dong, Socher, Li, Li, and Fei-Fei, 2009). We then augment the pretrained feature representation of the base encoder with a shallow adaptor network, which is trained to predict the perceptual correspondence between specific sketch-object pairs. The reason we train an adaptor network is that although prior work has shown that representation of object *categories* converges for sketches and photos at higher layers in DCNN models trained only on photos (Fan, Yamins, and Turk-Browne, 2018), additional supervision can substantially improve the accuracy of predictions involving comparisons between sketches and photos at the *instance* level (Sangkloy, Burnell, Ham, and Hays, 2016).

complex transformations applied by successive layers of VGG-19 are required to capture human behavior To evaluate the importance of the greater visual abstraction available at higher layers of VGG-19, we compare adaptor networks that intercept VGG-19 image representations at a lower, middle, and higher layer.

Each adaptor network was trained to predict the empirical estimates of sketch-object correspondence from the recognition experiment, and evaluated on held out data in a 5-fold crossvalidated manner using the same splits as in the previous section.

Consistent with our hypothesis, we found that a pragmatic sketcher model employing high-level features provided a substantially better fit to the data than one using mid-level features (high vs. mid BF: 94.8) or low-level features (high vs. low BF: 257). These results show that making fuller use of the depth of VGG to compute the perceptual correspondence between a sketch and object yields a stronger basis for explaining human visual communication behavior. Unsurprisingly, this pragmatic sketcher model employing high-level features did not fit the data as well as the pragmatic sketcher model that could directly access human recognition data (BF = 105.71). However, a major advantage of incorporating a visual encoder is the capacity to generalize to novel sketches without requiring the collection of additional recognition data for each new sketch.

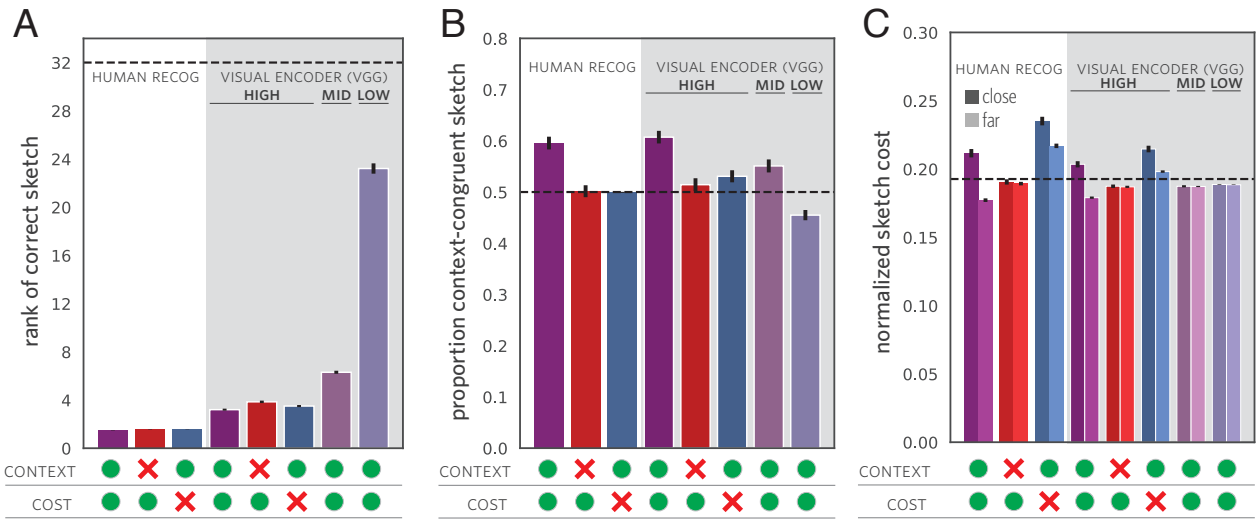

Fig. 5 Sketch production behavior by model variant. A green disc indicates that a given model is context/cost sensitive; a red 'X' indicates the lack of context/cost sensitivity. Results in lefthand region of each panel (white background) reflect model predictions when using empirical estimates of $sim(s,o)$ based on human sketch recognition behavior. Results in the righthand region (gray background) reflect model predictions when using variants of the visual encoder that represented sketches and objects at varying levels of visual abstraction (i.e., high, mid, low). All results reflect average model behavior on test data across five crossvalidation folds. Error bars represent 1 s.e. for this average estimate, found by applying inverse-variance weighting on individual confidence intervals from each train-test split. A: Rank of target sketch in list of 64 object-context categories, ordered by the probability assigned by each model. Dashed line reflects expected target rank under uniform guessing. Distribution of target rank scores across models suggest that high-quality estimates of $sim(s,o)$ are critical for strong performance. B: Proportion of trials on which each model assigned a higher rank to the context-congruent sketch of the correct object than the context-incongruent version of the correct object. Dashed line reflects expected behavior under indifference between the two versions of the sketch. Only models above this line show consistent and appropriate modulation of sketch producton by context. C: Normalized time cost of sketches produced by each model. Predicted sketch cost on each trial computed by marginalizing over probabilities assigned to each sketch category. Darker bars reflect behavior in the close condition; lighter bars the far condition. Dashed line indicates the average cost of sketches in the full dataset; bars below this line reflect a preference for sketches that are less costly than average, bars above this line for sketches that are costlier than average. Only models that span this dashed line match the pattern of contextual modulation of sketch cost displayed by human sketchers.

To further probe the functional consequences of decreasing the capacity for visual abstraction, we investigated each model variant on sketch retrieval, context congruity, and cost modulation. Critically, we found that high-level features supported strong performance on sketch retrieval (95% CI target rank: $[3.03, 3.37]$, Fig. 5A), compared to mid-level features (target rank: $[6.05, 6.56]$) and low-level features (target rank: $[22.4, 24.1]$). These results show that without a high-performing visual encoder, the model is much less likely to produce sketches of the correct object, a basic prerequisite for successful visual communication even in the absence of contextual variability.

Moreover, the pragmatic sketcher model using high-level features also displayed context congruity (95% CI: $[0.583, 0.632]$, Fig. 5B), comparable in degree to the best-performing pragmatic model that operated directly on empirical estimates of sketch-object correspondence, showing that our full sketcher model displayed this signature of contextual flexibility for novel commu-

nicative contexts and sketches. The variant using mid-level features also displayed context congruity to a weaker extent (95% CI: $[0.526, 0.576]$), suggesting that an intermediate level of visual abstraction is sufficient to achieve an intermediate degree of context congruity. By contrast, the variant using low-level features failed to prefer the context-congruent sketch category (95% CI: $[0.435, 0.475]$), providing a lower bound on the level of visual abstraction required in the underlying encoder to support flexible visual communication behavior.

Again, only the pragmatic sketcher model using high-level features displayed the same qualitative pattern of cost modulation as people did (95%CI: close = $[0.199, 0.208]$, far = $[0.178, 0.181]$, Fig. 5C), while both of the other variants using mid-level and low-level features failed to do so (95%CI: mid-level: close = $[0.186, 0.189]$, far = $[0.186, 0.188]$; low-level: close = $[0.188, 0.189]$, far = $[0.188, 0.189]$).

These results so far show the best-performing visual encoder to be the one making fuller use of the depth of the base visual encoder to extract more abstract perceptual properties, providing strong evidence for the importance of a high degree of visual abstraction for explaining our empirical findings. Next, we performed the same context and cost sensitivity lesion experiments as before in order to evaluate the contribution of pragmatic inference in our full sketcher model. Again, we found that the pragmatic sketcher provided a stronger overall fit to human behavior than the context-insensitive variant (BF = 28.1; see Table 1), and a modestly better fit than the cost-insensitive variant (BF = 1.98). Critically, we found that removing context and cost sensitivity diminished the ability of this model to produce the context-congruent sketch of the correct object (context-insensitive 95% CI: $[0.489, 0.539]$; cost-insensitive 95% CI: $[0.507, 0.554]$; Fig. 5B), and appropriately modulate the cost of the sketches it produced (context-insensitive 95% CI: close = $[0.185, 0.190]$, far = $[0.185, 0.189]$; cost-insensitive 95% CI: close = $[0.210, 0.219]$, far = $[0.196, 0.200]$; Fig. 5C). By contrast, these lesions led to only modest decrements in overall sketch retrieval performance (95% CI target rank: context-insensitive = $[3.65, 4.05]$, cost-insensitive = $[3.33, 3.67]$; Fig. 5A), suggesting that the visual encoder itself is a major determinant of the ability to produce sketches of the correct *object*, even if not the context-congruent version. These results converge with those of the lesion experiments conducted on the pragmatic sketcher model lacking a visual encoder, and together provide strong evidence for the importance of both visual abstraction and pragmatic inference for explaining contextual flexibility in human visual communication.

	human recog		visual encoder			
split	**context vs. no-context**	**cost vs. no-cost**	**context vs. no-context**	**cost vs. no-cost**	**high vs. mid**	**high vs. low**
1	18.0	11.9	44.5	2.70	105	282
2	8.46	9.89	20.9	-0.33	92.5	242
3	19.2	8.95	31.9	1.98	94.8	257
4	13.4	9.54	8.35	-0.67	93.4	248
5	16.1	7.92	28.1	5.99	114	269
median	**16.1**	**9.54**	**28.1**	**1.98**	**94.8**	**257**

Table 1 Log Bayes Factors (BF) for comparisons between full and lesioned model variants (columns) for each crossvalidation fold (rows). Log-BFs>0 indicate greater evidence for the full model than the lesioned variant. Columns under the human recog header contain comparisons between model variants that used empirical estimates of perceptual correspondence based on human sketch recognition behavior. Columns under the visual encoder header contain comparisons between model variants that used a deep convolutional neural network visual encoder, trained in a five-fold crossvalidated manner using human sketch recognition behavior. The context vs. no-context columns includes comparisons between context-sensitive and context-insensitive variant; the cost vs. no-cost columns includes comparisons between cost-sensitive and cost-insensitive variant; the high vs. mid column includes comparisons between model variants using a high adaptor vs. mid adaptor in a context/cost-sensitive model; and the high vs. low column includes comparisons between between model variants using a high adaptor vs. low adaptor in context/cost-sensitive model.

Discussion

The present study examined how communicative context influences visual communication behavior in a drawing-based reference game. We explored the hypothesis that people spontaneously account for information in common ground with their communication partner to produce drawings that are diagnostic of the target relative to the alternatives, while not being too costly to produce. We found that people spontaneously modulate how much time they invest in their drawings according to how similar the distractors are to the target, spending more time to produce more informative sketches when the alternatives were highly similar, but getting away with spending less time and producing less informative drawings when the alternatives were highly distinct. Observing such contextual flexibility provides strong evidence that visual communication about an object is not constrained exclusively by the visual properties of that object alone. Rather, our findings expose a critical role for pragmatic inference — the ability to infer what information would not only be true, but be *relevant* to communicate in context. To test this hypothesis, we developed a computational model that embodied both pragmatic inference and visual abstraction, and found that it predicted human communication behavior well, and outperformed variants of the model lacking either component. Together, this paper presents a first algorithmically explicit theory of how visual perception and social cognition support contextual flexibility during visual communication.

There are deep similarities between the computations performed by the visual encoder in our model and those posited by classic exemplar theories of categorization (Shepard, 1958; Medin and Schaffer, 1978; Nosofsky, 1988). For instance, our model encodes all objects and sketches as vectors embedded in a high-dimensional feature space, and learns (via the adaptor network) a similarity function that computes the correspondence between sketches and objects. Unlike the settings in which classic categorization models have typically been applied, our visual encoder operates directly on image inputs in order to compute similarity relations between instances from distinct visual modalities (i.e., sketch and 3D rendering). Although DCNN representations have recently been applied to explain human similarity judgments about images (Peterson, Abbott, and Griffiths, 2018; Kubilius, Bracci, and de Beeck, 2016), ours is among the first cognitive models to focus on learning instance-level mappings *between* image domains. In developing the visual encoder, we discovered that simple distance metrics (e.g., cosine or euclidean) applied to DCNN feature vectors were insufficient to accurately capture the human judgments of the correspondence between individual sketches and objects. As a consequence, we developed a custom nonlinear similarity function, parameterized by a shallow 'adaptor' neural network, in order to predict these correspondence relationships. More broadly, training adaptor neural networks to read out psychological quantities of interest from generic DCNN feature representations may be a promising approach to modeling how context and learning adapts perceptual representations for various downstream behaviors (Nosofsky, 2011; Medin and Schaffer, 1978).

This work generalizes the Rational Speech Act (RSA) modeling framework, originally developed to explain contextual effects in verbal communication (Frank and Goodman, 2012; Goodman and Stuhlmüller, 2013; Franke and Jäger, 2016; Bergen, Levy, and Goodman, 2016), to the domain of visual communication. RSA models take inspiration from the insights of Paul Grice (Grice, Cole, and Morgan, 1975), and incorporate ideas from decision theory, probabilistic models of cognition, bounded rationality, and linguistics, to understand how substantial variance in natural language use can be explained by general principles of social cognition. They have been shown to capture key patterns of natural language use (Goodman and Stuhlmüller, 2013), achieve good quantitative fits with experimental data (Kao, Bergen, and Goodman, 2014), and enhance the ability of artificial agents to produce informative language in reference game tasks (Monroe, Hawkins, Goodman, and Potts, 2017; Cohn-Gordon,

Goodman, and Potts, 2018). In extending this modeling framework to the visual domain, our findings provide novel evidence for the possibility that similar cognitive mechanisms may underlie pragmatic behavior across different communication modalities. This is a notion implicitly endorsed by prior work that has used non-linguistic communication modalities to investigate general constraints on communication, although these prior studies have not directly investigated contextual flexibility. (Goldin-Meadow and Feldman, 1977; Garrod, Fay, Lee, Oberlander, and MacLeod, 2007; Fay, Garrod, Roberts, and Swoboda, 2010; Theisen, Oberlander, and Kirby, 2010; Garrod, Fay, Rogers, Walker, and Swoboda, 2010; Galantucci, 2005; Verhoef, Kirby, and De Boer, 2014). Together, our findings suggest how drawings spanning a wide range of appearances can all nevertheless be effective carriers of meaning, depending on how much and what kind of information is shared between communicators. A fruitful avenue for future research would be to augment the current model with the capacity to learn from feedback and accumulate shared information over time, endowing it with the capacity to develop conventionalized ways of depicting objects that are increasingly efficient, yet still meaningful within a context (Garrod, Fay, Lee, Oberlander, and MacLeod, 2007; Hawkins, Sano, Goodman, and Fan, 2019).

Formally accounting for how shared information constrains visual communication is an important step towards a functional theory of pictorial meaning — how and why the pictures we use, including sketches, diagrams, icons, maps, and graphs look the way they do. One of the principal insights resulting from our study is that substantial progress towards this goal can be made by modeling the production of pictorial representations as fundamentally solving a social communication problem. While this general idea has roots in the philosophical literature (Goodman, 1976; Abell, 2009), our model is, to our knowledge, the first that can be directly applied to images to generate quantitative predictions. Although the trained model we present is optimized for the particular set of objects we used in our human behavioral experiment, our modeling *approach* is generic and can be applied to any new dataset. This is important because it suggests a general approach to modeling sketch understanding, which could have applications ranging from sketch-based retrieval (i.e., search engine using sketches as input; Sangkloy, Burnell, Ham, and Hays (2016)) to automatic sketch evaluation (e.g., in educational settings, how well a sketch captures the relevant properties of a target concept; Forbus, Usher, Lovett, Lockwood, and Wetzel (2008)).

There are several limitations of our model that would be valuable to address in future work. First, obtaining a visual encoder that could produce accurate predictions of perceptual correspondence between sketch-object pairs required substantial supervision. While heavy supervision is not uncommon when developing neural network models of sketch representation (Sangkloy, Burnell, Ham, and Hays, 2016; Yu, Yang, Liu, Song, Xiang, and Hospedales, 2017; Song, Yu, Song, Xiang, and Hospedales, 2017), future work should investigate architectures that require weaker supervision to estimate image-level correspondences between sketches and natural photographs. Moreover, future work should develop higher-capacity models that can scale up to a wider range of visual referents than the limited set of objects tested in our study, while avoiding a corresponding increase in model complexity and supervision. One promising approach may be to exploit the hierarchical and compositional structure of natural objects (i.e., parts, subparts, and their relations), as they are expressed in both natural images and sketches of objects (Mukherjee, Hawkins, and Fan, 2019).

Second, our model produces a decision over which *type* of sketch to produce in context, rather than producing a *particular* sketch. This is of course different from the action selection problem human participants face — they must decide not only what stroke to make, but where to place them, how many, and in what order. While there have been recent and promising advances in modeling sketch production as a sequence of such actions (Lake, Salakhutdinov, and Tenenbaum, 2015; Ha and Eck, 2017; Ganin, Kulkarni, Babuschkin, Eslami, and Vinyals, 2018), these approaches have not yet been shown to successfully emulate

how people sketch real objects, much less how this behavior is modulated by communicative context. Future work should develop sketch production models that both operate on natural visual inputs and more closely approximate the action space inherent to the task.

Meeting these challenges is not only important for developing more human-like artificial intelligence, but may also shed new light on the nature of human visual abstraction, and how ongoing perception and long-term conceptual knowledge guide action selection during complex, natural behaviors. In the long term, investigating the computational basis of visual communication may shed light on the sources of cultural variation in pictorial style, and lead to enhanced interactive visualization tools for education and research.

Materials and Methods

Communication experiment: Manipulation of context in sketch-based reference game

Participants

A total of 192 unique participants, who were recruited via Amazon Mechanical Turk (AMT) and grouped into pairs, completed the experiment. They were provided a base compensation of $2.70 for participation and earned a $0.03 bonus for each correct trial. In this and subsequent behavioral experiments, participants provided informed consent in accordance with the Stanford IRB.

Stimuli and Task

Because our goal was to understand how context influences the level of detail people use to distinguish objects from one another during visual communication, we populated our reference game with contexts possessing two key properties: (1) they contained familiar real-world objects, so that a primary source of variation would be driven by context, rather than difficulty recognizing or sketching the objects, *per se*; and (2) they systematically varied in target-distractor similarity within a session, lending greater statistical power to comparisons between context conditions. To satisfy these objectives, we obtained 32 3D mesh models of objects belonging to 4 basic-level categories (i.e., birds, chairs, cars, dogs), containing eight objects each. Each object was rendered in color on a gray background at three-quarter perspective, 10° viewing angle (i.e., slightly above), and fixed distance. Independently in each experimental session, objects were allocated to eight sets of four objects: Four of these sets contained objects from the same category (“close”); the other four of these sets contained objects from different categories (“far” condition). The assignment of objects to set and condition was randomized across pairs. Each set of four objects was presented four times each, such that each object in the quartet served as the target exactly once.

Sketchers drew using black ink on digital canvas (pen width = 5 pixels; 300 x 300 pixels) embedded in a web browser window using Paper.js (http://paperjs.org/). Participants drew using the mouse cursor, and were not able to delete previous strokes. Each stroke of which was rendered on the viewer’s screen immediately upon the completion of each stroke. There were no restrictions on how long participants could take to make their drawings. After clicking a submit button, the viewer guessed the identity of the drawn object by clicking one of the four objects in the array. Otherwise, the viewer had no other means of

communicating with the sketcher. Both participants received immediate task-related feedback: the sketcher learned which object the viewer had clicked, and the viewer learned the identity of the target. Both participants earned bonus points for each correct response.

Statistics

We primarily employed non-parametric analysis techniques (i.e., bootstrap resampling) to estimate the effects of experimental manipulations (Efron and Tibshirani, 1994). We favored this approach owing to its emphasis on estimation of effect sizes, by contrast with the dichotomous inferences yielded by traditional null-hypothesis significance tests (Cumming, 2014). Nevertheless, we found that traditional parametric statistical inference tests (i.e., *t*-tests) gave similar results, suggesting that our findings were robust to the particular choice of statistical analysis technique.

Recognition experiment: Measuring perceptual similarity between sketches and objects

Participants

A total of 112 participants were recruited via Amazon Mechanical Turk (AMT). They were provided a base compensation of $1.00 for their participation, and earned an additional $0.01 bonus for each correct response.

Task

On each trial, participants were presented with a randomly selected sketch collected in the communication experiment, surrounded by a grid containing the 32 objects from that experiment. Their goal was to select the object in the grid that best matched the sketch. Participants received task feedback in the form of a bonus earned for each correct trial. Participants were instructed to prioritize accuracy over speed. We applied a conservative outlier removal procedure based on response latency, whereby trials that were either too fast to have supported careful consideration of the sketch and menu of objects ($RT < 1000ms$), or too slow and suggestive of an attentional lapse ($RT > 30s$), were filtered from the dataset. The removal of these outlier trials (8.01%) did not have a substantial impact on the pattern of recognition behavior. In order to mitigate the possibility that participants could adjust their matching strategy according to any particular sketcher's style, each session was populated with 64 sketches sampled randomly from different reference games. To obtain robust estimates of sketch-object perceptual correspondences, each sketch was presented approximately 10 times across different sessions.

Computational modeling

Sketch data preprocessing

To train and evaluate our sketcher model, we first filter the sketch dataset to retain only sketches that were correctly identified by the viewer during the communication task (6.2% incorrect) and were compliant with task instructions by not including 'drawn'

text annotations (4.4% non-compliant). This filtered sketch dataset was then split into training, validation, and test sets in a 80%, 10%, and 10% ratio, and this split was performed in a 5-fold crossvalidated manner. Splits were based on context, defined as the set containing a specific target object and three distractor objects, such that no context appeared both in the training and test splits of any cross-validation fold. We ensured that: (1) the number of sketches from each category (i.e. car) and (2) the proportion of sketches from close and far trials were equated across splits. This was done to control for biases in model performance due to imbalances in the training or test set.

Deriving empirical estimates of perceptual correspondence between sketches and objects

In the recognition experiment, most sketches were not matched exclusively to a single object, but to several. We treated these sketches as thus displaying some degree of correspondence to the several objects it was matched to at least once. For a single sketch, we estimate the perceptual correspondence between that sketch and any object as the proportion of recognition task trials on which it was matched to that object. For sketches in each of 64 object-context categories, we estimate the "aggregated" sketch-object correspondence to be the proportion of recognition task trials on which any sketch from this category was matched to that object. Because our goal was to understand how well each model could produce informative sketches according to the context condition, and not necessarily to reproduce exactly the same sketch as a particular participant had on a specific trial, we use this aggregate correspondence measure in all modeling experiments. As a result, sketch-object correspondence scores lie in the range $[0,1]$, and sum to 1 for sketches in the same object-context category. Because all sketches from the same object-context category share the same correspondence to each object, there are a total of 32 sketch categories x 32 objects x 2 contexts = 2048 empirical perceptual correspondence scores.

Deriving empirical estimates of sketch costs

We reasoned that the amount of time taken producing each sketch would be a natural proxy for the cost incurred by workers on Amazon Mechanical Turk, who increase their total compensation by completing tasks in a timely manner. However, as there were no absolute constraints on the amount of time that could be spent on each trial, there was considerable variability across different participants in terms of how much time they spent producing their sketches. To control for this variability across participants and to ensure robust estimates, we first removed outliers (draw times exceeding 5 s.d. from the mean), then z-score normalized drawing times across all remaining trials within a participant, and finally averaged these normalized draw times across sketches within the same object-context category as above, yielding 32 objects x 2 contexts = 64 empirical cost estimates in total.

Visual encoder architecture

The visual encoder is a function that accepts a pair of images as input (both 224 x 224 RGB; see Fig. 4A): a sketch, s, and an object rendering, o, and returns a scalar value reflecting the degree of perceptual correspondence between the sketch and object, $\text{sim}(s,o)$, which lies in the range $[0,1]$, where $\text{sim}(s,o) = 0$ reflects minimal correspondence and $\text{sim}(s,o) = 1$ reflects maximal correspondence.

The encoder consists of two components: a base visual encoder and an adaptor network. We employed VGG-19 (Simonyan and Zisserman, 2014) as our base visual encoder architecture. We augmented VGG-19 with a shallow fully-connected *adaptor*

network that is trained to predict the perceptual correspondence between individual sketch-object pairs. Only the parameters of this adaptor network are trained; the parameters of the base visual encoder remain frozen. We compared three adaptor networks that intercept VGG-19 image representations at different layers: the first max pooling layer (early), the tenth convolutional layer (mid), and the first fully connected layer (high). To facilitate comparison between adaptor networks, we ensured that each of the three adaptors contain a comparable number of trainable parameters (number of learnable parameters for high: 1048839; mid: 1049115; low: 1048833) with identical training hyperparameters (i.e., learning rate, batch size, etc.). To discriminate which layer provides the best starting feature basis for predicting sketch-object correspondence, these adaptor networks were also deliberately constrained to be shallow, i.e., consisting only of two linear layers with an intervening point-wise nonlinearity.

High. When applying the high-level visual encoder, a sketch and object were first passed through VGG and a feature vector in $\mathbb{R}^{4096}$ for each image is extracted from the one of the highest layers (i.e., the first fully-connected layer, also known as *fc6*). These two vectors were then concatenated to form a single vector in $\mathbb{R}^{8192}$, to be passed into the high adaptor network. The high adaptor is composed of one linear layer that maps from $\mathbb{R}^{8192} \rightarrow \mathbb{R}^{128}$, followed by a "Swish" nonlinearity (Ramachandran, Zoph, and Le, 2018) and dropout, then a second linear layer mapping from $\mathbb{R}^{128} \rightarrow \mathbb{R}^{1}$. Swish is a recently discovered nonlinearity that outperforms the common rectified linear nonlinearity (ReLU) in deep models on several benchmarks (Ramachandran, Zoph, and Le, 2018). Dropout was applied to mitigate overfitting and improve generalization (Hinton, Srivastava, Krizhevsky, Sutskever, and Salakhutdinov, 2012; Gal and Ghahramani, 2015).

Mid. When applying the mid-level visual encoder, sketch and object representations are intercepted from an intermediate layer (i.e., the 10th convolutional layer, *conv_4_2*). Features in this layer are of dimensionality 512 x 28 x 28. Each of the sketch and object feature tensors were then "flattened" to a one dimensional vector in $\mathbb{R}^{512}$ using a weighted linear combination over the spatial dimensions $\sum_{i=1}^{28}\sum_{j=1}^{28} w_{ij} * x_{ij}$, where x_{ij} indexes a spatial location in the image representation at this layer (i.e., 'soft attention' over the spatial dimension, (Xu, Ba, Kiros, Cho, Courville, Salakhudinov, Zemel, and Bengio, 2015)). These weights $\{w_{ij} | 1 \leq i, j \leq 28\}$ are learned jointly with the parameters of the rest of the mid adaptor, but learned independently between sketch and object image modalities (Xu, Ba, Kiros, Cho, Courville, Salakhudinov, Zemel, and Bengio, 2015).

The two feature vectors in $\mathbb{R}^{512}$ are then concatenated to form a single vector in $\mathbb{R}^{1024}$. Following the architecture of the high adaptor, the mid adaptor consists of a linear layer that maps from $\mathbb{R}^{1024} \rightarrow \mathbb{R}^{1021}$, followed by a Swish nonlinearity, dropout, then a linear layer from $\mathbb{R}^{1021} \rightarrow \mathbb{R}^{1}$.

Low. When applying the low-level visual encoder, sketch and object representations are intercepted from the first max pooling layer (i.e., *pool1*). Features in this layer are of dimensionality 64 by 112 by 112. As above, a weighted sum of model activations over the spatial dimension was applied first (112 x 112), yielding a sketch and object vector, both in $\mathbb{R}^{64}$, which were then concatenated to form a single vector in $\mathbb{R}^{128}$. This was followed by a linear layer that maps from $\mathbb{R}^{128} \rightarrow \mathbb{R}^{7875}$, then a Swish nonlinearity, dropout, and a final linear layer that maps from $\mathbb{R}^{7875} \rightarrow \mathbb{R}^{1}$.

The penultimate hidden layer sizes in the mid (i.e., 1021 units) and low adaptors (i.e., 7875 units) were chosen to ensure that the total number of learnable parameters matched the high adaptor as closely as possible.

Visual encoder training

We trained each adaptor (i.e., high, mid, low) to predict, for each sketch, a 32-dimensional vector that captures the *pattern* of perceptual correspondences between that sketch and all 32 objects. The rationale for having the adaptor generate a 32-dimensional vector, rather than only the correspondence to the target, is to explicitly encourage it to match the *pattern* of correct responses and errors in human sketch recognition behavior, rather than to achieve maximal accuracy on the task.

Each encoder accepts a sketch-object pair as input and returns a real number as output, reflecting their perceptual correspondence. We iterate over all objects in the stimulus set I to generate the predicted 32-vector for each sketch, and then apply softmax normalization, yielding a vector that sums to 1. We define the loss function, $\mathcal{L}$, to be the cross entropy loss between the predicted distribution, q and the empirically estimated perceptual correspondence vector, p (which also sums to 1):

$$\mathcal{L} = \sum_{x \in I} p(x) \log q(x) \tag{5}$$

We use the Adam optimization algorithm (Kingma and Ba, 2014) (learning rate = 1e-4) over minibatches of size 10 for 100 epochs, where an epoch is a full pass through the training set.[1] After training each adaptor for 100 epochs, we select the model found during training with the best performance on the validation set.

Generating encoder-based estimates of perceptual correspondence between sketches and objects

To generate sketch-object correspondence scores for sketches in each test split, we first pass each sketch-object pair into a visual encoder, yielding a single image-level correspondence score lying in the range $(-\infty, +\infty)$. To map these raw image-level scores to the appropriate range for a correspondence score ($[0,1]$), we first z-score them ($f(x) = \frac{x-\bar{x}}{s}$), then apply the logistic function ($f(x) = \frac{1}{1+e^{-x}}$). These normalized image-level correspondence scores are then averaged across all sketches belonging to the same object-context category, yielding 32 objects x 32 sketches x 2 contexts = 2048 model-based perceptual correspondence scores for each visual encoder variant (i.e., high, mid, low).

Model comparison

In order to test the contribution of each component of our sketcher model, we conducted a series of lesion experiments and formal model comparisons. To quantify the evidence for one model over another, we computed Bayes Factors: the ratio of likelihoods for each model, integrating over all their respective parameters under the prior:

$$BF = \frac{\int P(D|M_1, \theta_1) P(\theta_1)}{\int P(D|M_2, \theta_2) P(\theta_2)}$$

Unlike classical likelihood ratio tests, which use the maximum likelihood, the Bayes Factor naturally penalizes models for their complexity (Wagenmakers, Marsman, Jamil, Ly, Verhagen, Love, Selker, Gronau, Šmíra, Epskamp et al., 2018; Jefferys and

[1] As a property of the input domain, the gradients with respect to adaptor parameters are very small (1.51e-4 $\pm$ 2.61e-4), inevitably resulting in poor learning (we can reproduce this effect from several initializations). We find that naively increasing the learning rate led to unstable optimization, but that multiplying the loss by a large constant C leads to a much smoother learning trajectories and good test generalization. Critically, increasing the learning rate and multiplying the loss by a constant are not equivalent for second moment gradient methods. In practice, $C = 1e4$.

Berger, 1992). We placed uninformative uniform priors over all five parameters required to specify our models: a discrete choice over alternative approaches to computing perceptual correspondance:

$$m \sim \text{Unif}\{\text{“human recog”}, \text{“high”}, \text{“mid”}, \text{“low”}\}$$

and over the continuous latent parameters,

$$w_i, w_c, \alpha \sim \text{Unif}(0, 500),$$

$$w_d \sim \text{Unif}(0, 1).$$

Note that w_i, w_c, and α were allowed to take any nonnegative real value (i.e., were not restricted to fall between 0 and 1). In practice, an upper bound of 500 for the uniform prior was found to be sufficiently large to support robust inference. By contrast, w_d, which balances the contributions of absolute perceptual correspondence and relative diagnosticity in context, was constrained to fall between 0 and 1. To compute the likelihood function $P(D|M, \theta)$ for a speaker model M under parameters θ, we perform exact inference for our sketcher model using (nested) enumeration and sum over all test set datapoints within a crossvalidation fold.

Specifically, we compute the exact likelihood at every point on a discrete grid of parameters. This is of particular interest for nested model comparisons, e.g. comparing our full model to a context-insensitive variant. Rather than computing the full marginalized likelihood for both models, we can use the Savage-Dickey method (Wagenmakers, Lodewyckx, Kuriyal, and Grasman, 2010) to simply compare the posterior probability against the prior at the nested point of interest (e.g. $w_c = 0$) for the full model.

To evaluate the contribution of pragmatic inference, we begin by comparing the pragmatic sketcher model using empirically estimated perceptual correspondences to nested “cost-insensitive” ($w_c = 0$) and “context-insensitive” ($w_d = 0$) variants. To evaluate the contribution of visual abstraction, we then proceed to compare the three visual encoder variants that adapt features from different layers of VGG-19, marginalizing over all other parameters. Finally, we perform the same context and cost lesion experiments on the full model that employed the best-performing visual encoder (i.e., “high”).

Evaluating model predictions

We implemented our models and conducted inference in the probabilistic programming language WebPPL (Goodman and Stuhlmüller, 2014). We use MCMC to draw 1000 samples from the joint posterior with a lag of 0, discarding 3000 burn-in samples. We constructed posterior predictive distributions by computing each measure of interest (i.e., target rank, context congruity, sketch cost) over the test data set, for every MCMC sample. To estimate standard errors on predictions across models, we employed the following procedure to account for three nested sources of variation: variation across trials within a test split, variation across the parameter posterior within a test split, and variation across test splits. Specifically, for each model variant and for each test split we bootstrap resampled trials with replacement from the test dataset 1000 times to estimate the mean and standard error on each measure of interest, marginalizing over MCMC samples from the parameter posterior. We applied inverse-variance weighting to aggregate these estimates of the mean and standard error across test splits, such that test splits with lower variance contribute more than do splits with higher variance, yielding an overall estimate of the mean and standard

error for each measure of interest, for each model variant. We estimated the half-widths of the 95% confidence interval for each measure of interest under the assumption of normality for the sampling distribution of the mean.

Statistics

In our behavioral experiments, we employ non-parametric analysis techniques (i.e., bootstrap resampling) to construct 95% confidence intervals and compute p-values for key parameters and comparisons of interest. All p-values reported for comparisons between conditions are two-sided, found by determining the proportion of 10,000 bootstrap iterations that fell below zero, multiplied by two. In our computational modeling experiments, we employ Bayesian data analysis to infer full posterior distributions over latent parameters from data, and perform formal model comparison by computing Bayes Factors using marginal likelihoods. We additionally ensure robustness of all modeling results using five-fold crossvalidation.

Code and data availability

All code and data used to produce the results in this article are publicly available in a Github repository at: `https://github.com/judithfan/visual_communication_in_context`. The code used to train the visual encoder module is available at: `https://github.com/judithfan/visual-modules-for-sketch-communication-public`.

Acknowledgements

Thanks to Dan Yamins and the Stanford CoCo Lab for helpful comments and discussion. RXDH was supported by the Stanford Graduate Fellowship and the National Science Foundation Graduate Research Fellowship under Grant No. DGE-114747.

Author contributions statement

J.E.F and R.X.D.H. designed and conducted human experiments, J.E.F, R.X.D.H, and M.W. analyzed data and performed computational modeling. J.E.F, R.X.D.H, M.W., and N.D.G. formulated models, interpreted results, and wrote the paper.

Additional information

The authors declare no competing interests.

References

Abell C (2009) Canny resemblance. Philosophical Review 118(2):183–223

Allen JP (2000) Middle Egyptian: An introduction to the language and culture of hieroglyphs. Cambridge University Press

Aubert M, Brumm A, Ramli M, Sutikna T, Saptomo EW, Hakim B, Morwood MJ, van den Bergh GD, Kinsley L, Dosseto A (2014) Pleistocene cave art from Sulawesi, Indonesia. Nature 514(7521):223–227

Bergen L, Levy R, Goodman N (2016) Pragmatic reasoning through semantic inference. Semantics and Pragmatics 9

Boltz WG (1994) The origin and early development of the Chinese writing system, vol 78. Eisenbrauns

Cohn-Gordon R, Goodman ND, Potts C (2018) Pragmatically informative image captioning with character-level inference. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pp 439–443

Cumming G (2014) The new statistics: Why and how. Psychological science 25(1):7–29

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009., IEEE, pp 248–255

Donald M (1991) Origins of the modern mind: Three stages in the evolution of culture and cognition. Harvard University Press

Efron B, Tibshirani RJ (1994) An introduction to the bootstrap. CRC press

Fan JE, Yamins DLK, Turk-Browne NB (2018) Common object representations for visual production and recognition. Cognitive Science 0(0), DOI 10.1111/cogs.12676, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12676, https://onlinelibrary.wiley.com/doi/pdf/10.1111/cogs.12676

Fay N, Garrod S, Roberts L, Swoboda N (2010) The interactive evolution of human communication systems. Cognitive Science 34(3):351–386

Forbus KD, Usher JM, Lovett AM, Lockwood K, Wetzel J (2008) Cogsketch: Open-domain sketch understanding for cognitive science research and for education. SBM 8:159–166

Frank MC, Goodman ND (2012) Predicting pragmatic reasoning in language games. Science 336(6084):998–998

Franke M, Jäger G (2016) Probabilistic pragmatics, or why bayes rule is probably important for pragmatics. Zeitschrift für sprachwissenschaft 35(1):3–44

Gal Y, Ghahramani Z (2015) Dropout as a bayesian approximation: Insights and applications. In: Deep Learning Workshop, ICML, vol 1, p 2

Galantucci B (2005) An experimental study of the emergence of human communication systems. Cognitive Science 29(5):737–767

Ganin Y, Kulkarni T, Babuschkin I, Eslami S, Vinyals O (2018) Synthesizing programs for images using reinforced adversarial learning. arXiv preprint arXiv:180401118

Garrod S, Fay N, Lee J, Oberlander J, MacLeod T (2007) Foundations of representation: where might graphical symbol systems come from? Cognitive science 31(6):961–987

Garrod S, Fay N, Rogers S, Walker B, Swoboda N (2010) Can iterated learning explain the emergence of graphical symbols? Interaction Studies 11(1):33–50

Gibson JJ (1979) The ecological approach to visual perception: classic edition. Psychology Press

Goldin-Meadow S, Feldman H (1977) The development of language-like communication without a language model. Science 197(4301):401–403

Gombrich E (1969) Art and Illusion: A Study in the Psychology of Pictorial Representation. (Bollingen series, 35. The A. W. Mellon lectures in the fine arts, 5), Princeton University Press

Gombrich E (1989) The story of art. Phaidon Press, Ltd.

Goodman N (1976) Languages of art: An approach to a theory of symbols. Hackett publishing

Goodman N, Frank M (2016) Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Sciences 20(11):818–829

Goodman N, Stuhlmüller A (2013) Knowledge and implicature: Modeling language understanding as social cognition. Topics in cognitive science 5(1):173–184

Goodman N, Stuhlmüller A (2014) The design and implementation of probabilistic programming languages

Grice HP, Cole P, Morgan JL (1975) Syntax and semantics

Güçlü U, van Gerven MA (2015) Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience 35(27):10005–10014

Ha D, Eck D (2017) A neural representation of sketch drawings. arXiv preprint arXiv:170403477

Hawkins R, Sano M, Goodman N, Fan J (2019) Disentangling contributions of visual information and interaction history in the formation of graphical conventions

Hinton GE, Srivastava N, Krizhevsky A, Sutskever I, Salakhutdinov RR (2012) Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:12070580

Hochberg J, Brooks V (1962) Pictorial recognition as an unlearned ability: A study of one child's performance. the american Journal of Psychology pp 624–628

Hoffmann D, Standish C, García-Diez M, Pettitt P, Milton J, Zilhão J, Alcolea-González J, Cantalejo-Duarte P, Collado H, De Balbín R, et al. (2018) U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. Science 359(6378):912–915

Jefferys WH, Berger JO (1992) Ockham's razor and bayesian analysis. American Scientist 80(1):64–72

Kao J, Bergen L, Goodman N (2014) Formalizing the pragmatics of metaphor understanding. In: Proceedings of the annual meeting of the Cognitive Science Society, vol 36

Kennedy JM, Ross AS (1975) Outline picture perception by the songe of papua. Perception 4(4):391–406

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980

Kubilius J, Bracci S, de Beeck HPO (2016) Deep neural networks as a computational model for human shape sensitivity. PLoS computational biology 12(4):e1004896

Lake BM, Salakhutdinov R, Tenenbaum JB (2015) Human-level concept learning through probabilistic program induction. Science 350(6266):1332–1338

Lewis D (1969) Convention: A philosophical study. Harvard University Press

Malach R, Levy I, Hasson U (2002) The topography of high-order human object areas. Trends in cognitive sciences 6(4):176–184

Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychological review 85(3):207

Monroe W, Hawkins RX, Goodman ND, Potts C (2017) Colors in context: A pragmatic neural model for grounded language understanding. arXiv preprint arXiv:170310186

Mukherjee K, Hawkins R, Fan J (2019) Conveying semantic part information in drawings

Nosofsky RM (1988) Exemplar-based accounts of relations between classification, recognition, and typicality. Journal of Experimental Psychology: learning, memory, and cognition 14(4):700

Nosofsky RM (2011) The generalized context model: An exemplar model of classification. Formal approaches in categorization pp 18–39

Peterson JC, Abbott JT, Griffiths TL (2018) Evaluating (and improving) the correspondence between deep neural networks and human representations. Cognitive science 42(8):2648–2669

Ramachandran P, Zoph B, Le QV (2018) Searching for activation functions

Rolls ET (2001) Functions of the primate temporal lobe cortical visual areas in invariant visual object and face recognition. In: Vision: The Approach of Biophysics and Neurosciences, World Scientific, pp 366–395

Sangkloy P, Burnell N, Ham C, Hays J (2016) The sketchy database: learning to retrieve badly drawn bunnies. ACM Transactions on Graphics (TOG) 35(4):119

Sayim B (2011) What line drawings reveal about the visual brain pp 1–4

Shepard RN (1958) Stimulus and response generalization: tests of a model relating generalization to distance in psychological space. Journal of Experimental Psychology 55(6):509

Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556

Song J, Yu Q, Song YZ, Xiang T, Hospedales TM (2017) Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In: ICCV, pp 5552–5561

Tanaka M (2007) Recognition of pictorial representations by chimpanzees (pan troglodytes). Animal cognition 10(2):169–179

Theisen CA, Oberlander J, Kirby S (2010) Systematicity and arbitrariness in novel communication systems. Interaction Studies 11(1):14–32

Tomasello M (2009) The cultural origins of human cognition. Harvard university press

Verhoef T, Kirby S, De Boer B (2014) Emergence of combinatorial structure and economy through iterated learning with continuous acoustic signals. Journal of Phonetics 43:57–68

Wagenmakers EJ, Lodewyckx T, Kuriyal H, Grasman R (2010) Bayesian hypothesis testing for psychologists: A tutorial on the savage–dickey method. Cognitive psychology 60(3):158–189

Wagenmakers EJ, Marsman M, Jamil T, Ly A, Verhagen J, Love J, Selker R, Gronau QF, Šmíra M, Epskamp S, et al. (2018) Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. Psychonomic bulletin & review 25(1):35–57

Wilson D, Sperber D (1986) Relevance: Communication and cognition. Mass.

Wittgenstein L (1953) Philosophical investigations. Macmillan

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057

Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ (2014) Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences 111(23):8619–8624

Yu Q, Yang Y, Liu F, Song YZ, Xiang T, Hospedales TM (2017) Sketch-a-net: A deep neural network that beats humans. International journal of computer vision 122(3):411–425

Zipf GK (1936) The psycho-biology of language: An introduction to dynamic philology. Routledge