AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

© 2023 American Psychological Association
ISSN: 0096-3445

2023, Vol. 152, No. 10, 2713–2734
https://doi.org/10.1037/xge0001413

Generative Visual Common Sense: Testing Analysis-by-Synthesis on Mondrian-Style Image

Ning Tang[1, 2], Siyi Gong[3, 4], Jifan Zhou[1], Mowei Shen[1], and Tao Gao[2, 3, 4]
[1] Department of Psychology and Behavioral Sciences, Zhejiang University
[2] Department of Statistics, UCLA
[3] Department of Communication, UCLA
[4] Department of Psychology, UCLA

The well-known Mondrian-style images, aside from being aesthetically amusing, also reflect the core principles of human vision in their viewing experience. First, when we see a Mondrian-style image consisting only of a grid and primary colors, we may automatically interpret its causal history such that it was generated by recursively partitioning a blank scene. Second, the image we observe is open to many possible ways of partitioning, and their probabilities of dominating the interpretation can be captured by a probabilistic distribution. Moreover, the causal interpretation of a Mondrian-style image can emerge almost spontaneously, not being tailored to any specific task. Using Mondrian-style images as a case study, we demonstrate the generative nature of human vision by showing that a Bayesian model based upon an image-generation task can support a wide range of visual tasks with little retraining. Our model, learned from human-synthesized Mondrian-style images, could predict human performance in the perceptual complexity ranking, capture the transmission stability when images were iteratively passed among participants, and pass a visual Turing test. Our results collectively show that human vision is causal such that we interpret an image from the angle of how it was generated. The success of generalization with little retraining suggests that generative vision constitutes a type of common sense that supports a wide range of tasks of different natures.

Public Significance Statement

By using Mondrian-style images as a case study, this study demonstrated that modeling how humans draw images can well explain how humans perceive images across a variety of tasks. This study suggests that a deep understanding of how an image is generated can serve as a source of visual common sense.

Keywords: generative vision, analysis-by-synthesis, common sense, Bayesian model, Turing test

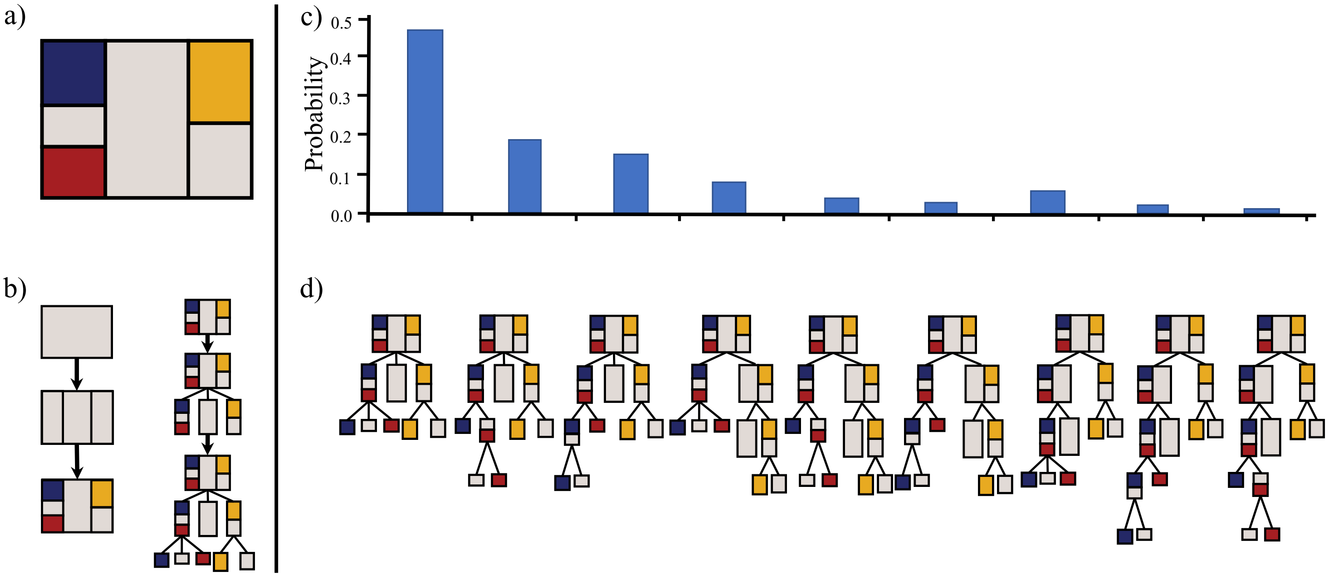
Contemporary visual art can be very abstract, to the extent that the art vocabulary may be reduced to simple geometric shapes. This characteristic is well illustrated by artist Piet Mondrian, whose influential Neo-Plasticism style is characterized by only horizontal and vertical lines as well as primary colors. Despite its visual simplicity, the stacking of rectangles in his signature painting can endure viewers' sustained attention. This is in part because the structure of this Mondrian-style image is open to many potential explanations competing for the visual interpretation. Such abstract artistic style nevertheless reflects several core principles of human vision. First, vision is "causal," automatically interpreting the "history" of an image—how an image was sequentially generated (e.g., Chen & Scholl, 2016; Freyd, 1987; Leyton, 1989). Specific to the Mondrian-style image in Figure 1a, one particular structural interpretation represents one unique way to recursively "parse" a blank scene into smaller pieces until the observed image is eventually produced, as presented in Figure 1b. Second, vision is "probabilistic" (e.g., Bennett et al., 1989; Rock, 1983) instead of a logically deductive process. That is, one cannot deductively conclude that a particular generative process constitutes the history of an image while other processes do not. Specific to the Mondrian-style image in Figure 1a, it may be the product of any one of the generative processes presented in Figure 1d. Thus, to measure the plausibility of all possible processes

This article was published Online First May 18, 2023.
Ning Tang https://orcid.org/0000-0001-5715-6338
Tao Gao https://orcid.org/0000-0003-2523-7867
Individuals can access a full description of the data and code through the Open Science Framework website (https://osf.io/u62xb/) and access the supplemental materials through (https://sites.google.com/view/generativevision-commonsense). We have no conflicts of interest to disclose. Some of the data in this work were previously presented at the annual convention of the Association for Psychological Science in Boston, Massachusetts, in May 2017 and the annual meeting of the Vision Sciences Society in St. Pete Beach, Florida, in May 2018.

We thank Michael Lee and other four anonymous reviewers for their insightful comments and helpful suggestions. This work was performed with the support of UCLA startup funding to Tao Gao and grants from National Natural Science Foundation of China (31871096 to Jifan Zhou, 32071044 to Mowei Shen).

Correspondence concerning this article should be addressed to Tao Gao, Department of Communication, UCLA, Los Angeles, CA 90095, United States, or Jifan Zhou, Department of Psychology and Behavioral Sciences, Zhejiang University, Xixi Campus, Hang Zhou 310007, People's Republic of China, or Mowei Shen, Department of Psychology and Behavioral Sciences, Zhejiang University, Xixi Campus, Hang Zhou 310007, People's Republic of China. Email: taogao@ucla.edu or jifanzhou@zju.edu.cn or mwshen@zju.edu.cn

Figure 1
A Mondrian-Style Image and Its Possible Generative Processes



Note. (a) Illustration of a Mondrian-style image. (b) Illustration of one specific process to create a Mondrian-style image by recursively partitioning an image into smaller pieces. (c) The probabilities of all generative processes estimated by the human prior hierarchical Bayesian model that will be later reported in this study. (d) All possible generative processes for a single Mondrian-style image. See the online article for the color version of this figure.

quantitatively, probability becomes a necessary tool. For example, the probabilities of generative processes based on the model we reported in this study are shown in Figure 1c. Third, the causal interpretation of vision serves as a common sense that emerges rather automatically without any specification of a task. Imagine yourself in an art gallery viewing a Mondrian-style image—are you purposefully engaging in a particular task? The answer is most likely no. Nevertheless, we believe that spontaneous causal vision can support a wide range of tasks (e.g., Lake et al., 2015), a notion referred to as the "big tasks" nature of human vision (Y. Zhu et al., 2020). The main focus of this study is to show that the causal understanding of image generation in vision can support a wide range of visual tasks.

Here, we explore the generative nature of human vision on Mondrian-style images as a case study. We adopt a visual commonsense perspective, applying a model to many different visual tasks which it has never encountered before. In fact, the model we use here is not trained on how humans *analyze* images, but on how humans *synthesize* images. Our results offer direct support for the analysis-by-synthesis perspective of human vision (Yuille & Kersten, 2006). In the following sections, we will introduce the theoretical backgrounds of each of the core principles in human vision and how they can guide psychophysical experiments, using Mondrian-style images as a running example.

Vision as Inverse Graphics

"What I cannot create, I do not understand."—Richard Feynman

All images are generated by certain causal processes, deployed spatially and temporally. Once finished, the dynamic generation process makes its exit, leaving behind only a static image. For humans, vision is not just about precisely registering every pixel of an image as it is, but includes a crucial process of recovering the causal history of an image by "searching" for a plausible generation process. This mechanism is referred to as the analysis-by-synthesis process (Grenander, 1976; Yuille & Kersten, 2006; S.-C. Zhu & Mumford, 2007), also known as inverse graphics (e.g., Kulkarni et al., 2015; Yildirim et al., 2015). Analogous to computer graphics, a generative process is a graphic program that procedurally creates a scene consisting of geometrical entities and then uses it to render a 2D image. On the contrary, vision inverts the process, reconstructing the geometrical entities and their relationships as graphic programs from a 2D image.

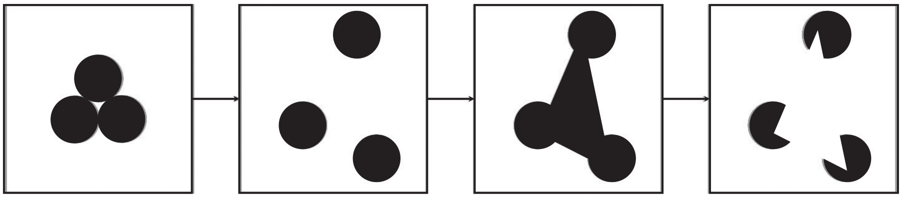
Here we illustrate the notion of inverse graphics on an iconic illusion, the Kanizsa Triangle. We often see a white triangle in the Kanizsa Triangle because it is the easiest way to construct an image. For example, we can create a Kanizsa Triangle by (a) creating three black circles, (b) placing them apart from each other, (c) putting a triangle on top of them, and (d) painting the triangle white (Figure 2). Generating a Kanizsa Triangle this way is much more efficient than meticulously arranging three pac-man shapes to make the edges of their mouths align to those of each other. Technically, for Figure 2, it may be unjustifiable to call it an illusion, since a white triangle was indeed present in the image created out of photoshop.

Historically, the inverse graphics perspective is deeply rooted in constructivism, following the idea that the rich world we "see" is often not directly accessible from stimuli but is nonetheless constructed or interpreted by a vision based on those stimuli (e.g., Bruner, 1973; Gregory, 1970; Palmer, 1999; Rock, 1983; von Helmholtz, 1867/1925). This constructivism idea has been used to explain amodal completion (Kanizsa, 1976), color constancy (Brainard & Freeman, 1997), size constancy (Geisler & Kersten, 2002), and many other visual phenomena (e.g., Adelson, 2000; Palmer, 1999; Rock, 1983).

Recently, an increasing body of studies have been focused on how vision understands images by recovering their causal history,

Figure 2

Illustration of an Efficient Way of Creating a Kanizsa Triangle by Photoshop



especially on the perception of shapes. The idea is that "shape is time"—static shapes contain temporal information that recapitulates the causal histories of objects (Leyton, 1989, 1992). For example, the logo of Apple Inc. is frequently viewed as a whole apple that was bitten, and a dented can is often regarded as an undamaged can that was subsequently dented. An increasing number of studies have provided empirical evidence supporting the perceptual causal history. A study showed that causal history could bias people's judgment on the geometric properties of static shapes: When a shape appeared like an object that has been bitten, observers tended to compensate for the "missing part" and perceive a symmetry axis biased toward that of the "complete" shape of that object, instead of strictly following the shape's true skeleton representations (Spröte et al., 2016). Further, humans could recognize transformations subjected to objects: When an object was deformed in its shape, observers were able to identify the type of deformation; and when the object was partially deformed, observers could perceive a clear distinction between its intact portions and deformed portions (Schmidt et al., 2019). Moreover, observers even perceived a motion illusion as if they chronologically relived the causal history: When a change in the contours of a shape suggested a historical intrusion, observers perceived a gradual development of the intrusion that was, in fact, made abrupt (Chen & Scholl, 2016).

More broadly, human performance in recognizing images has been shown to be connected with how those images are generated. For example, in studies on visual working memory, it has been shown that the capacity of working memory can be explained by a model that recovers the latent hierarchical structure of images (Brady & Tenenbaum, 2013; Suchow et al., 2014). Moreover, recent studies on visual production, although not particularly focusing on the analysis-by-synthesis perspective, have shown that generation and recognition of images recruit the same visual representation in processing objects: Adults exhibited enhanced recognition of objects after learning how to draw them (Fan et al., 2018), and children exhibited parallel developmental changes in drawing and recognition of visual concepts (Long et al., 2021).

Formulating Inverse Graphics as Bayesian Inference

Image generation is a one-to-one process from a graphic procedure (cause) to an image (effect). On the contrary, inverse graphics is a one-to-many process, from which an image can be explained by a variety of graphic procedures. As an example, the same Mondrian-style image, comprising a specific set of rectangular pieces, can be rendered by multiple ways of partitioning a canvas. Such a one-to-many inverse process is ill-posed for logical deduction (e.g., Bennett et al., 1989; Marr, 1982; Rock, 1983; von Helmholtz, 1867/1925), yet it can be naturally represented by a probabilistic distribution. Given the image, the task of vision is to find out the posterior distribution of those explanations following Bayes' rule (Knill & Richards, 1996; Yuille & Kersten, 2006; S.-C. Zhu & Mumford, 2007). This insight is neatly captured by the following equation, though solving it can be computationally challenging.

$$
\begin{aligned}
P(\text{graphics}|\text{image}) &= \frac{P(\text{graphics},\ \text{image})}{P(\text{image})} \\
&= \frac{P(\text{image}|\text{graphics}) \times P(\text{graphics})}{P(\text{image})} \qquad (1)
\end{aligned}
$$

In Equation 1, the prior P(graphics) formulates an observer's knowledge about the scene preceding any observation. The likelihood P(image|graphics), serving as the probability of rendering an image given a graphic procedure, represents the image generation process.

The Bayesian approach has been adopted by the field of cognitive science in studying human vision. Early Bayesian cognitive modeling studies mainly focus on 3D graphical concepts of simple objects (Kersten et al., 2004; Knill & Richards, 1996), such as shaping (e.g., Freeman, 1996), lighting (Mamassian et al., 2002), viewpoint (e.g., Nakayama & Shimojo, 1992), surface geometry (e.g., Vetter & Troje, 1997), material reflectivity (e.g., Adelson, 2000), and object pose (e.g., Richards et al., 1996).

To date, cognitive modeling of human vision has been embracing a richer range of generative processes. For example, in a study of shape representation, a shape is assumed to be "grown" from a skeleton through a stochastic generative process that recursively develops "ribs" from the skeleton to support the contour of the shape. This representation enables possibilities for a great variety of shapes (Feldman & Singh, 2006). In the study of perceptual grouping, with the view that the configuration of image elements is generated by a mixture of distinct objects, various perceptual grouping problems could be solved within the framework of Bayesian hierarchical clustering (Froyen et al., 2015). More recently, the generative process has been more broadly defined as a probabilistic program (Gelman et al., 2015; Goodman et al., 2008; Lake et al., 2017), which includes statements, grammar, and recursions similar to other programs. This idea has been reflected in applications such as CAPTCHA recognition (Mansinghka et al., 2013), face analysis, body pose estimation, object reconstruction (Kulkarni et al., 2015), and handwritten letters analysis (Lake et al., 2015).

Testing Generative Vision With "Big Tasks"

One of the biggest advantages of a causal model is that a single model can be generalized to multiple tasks, reflecting the general characteristics of human intelligence. Such capability is discussed in depth in a recent review article[1] (Y. Zhu et al., 2020) that integrates the causal perspectives from both the fields of computer vision and cognitive science. In contrast to the mainstream "big data for small tasks" methods commonly employed in building artificial

[1] Author Tao Gao contributed to this review article.

intelligence that are largely "empowered by large-scale annotated data and end-to-end training using neural networks," the article calls for a "small data, big tasks" paradigm where the application of a generative model can be generalized across multiple tasks with little training data. This is because the causality in the generative process is relatively stable compared to the transient nature of tasks that humans need to face. That is, what you do with an object does not change how it was generated. Thus, the invariant nature enables a causal model to be well adapted to novel situations with little training (Pearl, 2000).

Evidence supporting the versatility of causal models has been manifested in several cognitive science studies. In modeling intuitive physics, a physics engine model has been used for predicting both whether a tower is stable and in which direction it will fall (Battaglia et al., 2013). More related to vision, the same generative model of handwritten letters has been used to solve different tasks, including classifying characters in one-shot training, imitating a novel character, and recovering the dynamic process of character generation (Lake et al., 2015). More recently, one generative model of recursive visual concepts has been applied in both classifying new concepts and generating new examples of concepts (Lake & Piantadosi, 2020).

In the current study, we aim to demonstrate the generative nature of human vision using the "big tasks" paradigm. Previously, a computer vision model synthesizing Mondrian-style images (Roy & Teh, 2008) has been built using a generalized Poisson Process. Here, we aim to explore whether human vision explains Mondrian-style images using the reconstruction of their causal history. Importantly, while we developed a generative model for recognizing Mondrian-style images, the focus is not on the model itself but on demonstrating the utility of a deep integration of a generative model and human psychophysics experiments. Specifically, we examined it through the following perspectives of human vision. First, the tasks involved should consist of both image generation and image interpretation, two inseparable processes that lie at the central idea of "analysis-by-synthesis" (e.g., Yuille & Kersten, 2006). This entails that the model, learned from how humans generate Mondrian-style images, can be transferred into modeling how humans interpret those images.

Second, the tasks should measure the causal process both explicitly and implicitly. In the explicit case, the task requires participants to recognize the causal history of an image; whereas in the implicit case, the task itself is unrelated to the estimation of the image generation process. Importantly, the generative process should be able to explain visual performance regardless of whether it is explicit or implicit in the task. To date, the field of human vision has accumulated a diverse collection of implicit and explicit behavioral measurements of human vision, on some of which our experiment method is grounded.

Third, while we are looking for a diversity of measurements for task performance, we hope to find a unified approach for manipulating independent variables so that results across different experiments can be integrated. Since our model is probability-driven, we use the model's estimation of the probability of generating an image, otherwise known as the P(image), as the major independent variable of the study. Specifically, we measure how task performance varies as a function of P(image) (A detailed review of the theoretical implications of P(image) can be found in the Discussion). This approach is different from traditional psychophysics in which physical features are extracted for manipulation. Although estimating the probability of an image is necessarily affected by features, P(image) integrates and absorbs a variety of visual features, offering an opportunity for experimenters to utilize a single unified index to make predictions without the need to identify each and every one of them.

Following the three perspectives under the "big task" paradigm, we split the four human experiments in our study into two major parts. Experiment 1 focused on image generation with the goal of eliciting the subjective P(image) from participants' minds. A "just cut it" task was conducted where human participants were asked to create Mondrian-style images by splitting a blank paper into pieces through recursively applying horizontal and vertical divisions. Then, a model of P(image) for Mondrian-style images was learned from samples of human-generated data.

In the rest of the study, we conducted three visual tasks to demonstrate that the very same image-generation model could be applied to solve different image-interpretation tasks. Experiment 2 tested whether the model could predict humans' subjective complexity of images, adopting the pairwise comparison paradigm (Thurstone, 1927). In Experiment 3, we used a "telephone game" (Bartlett, 1932; Griffiths & Kalish, 2007; Uddenberg & Scholl, 2018) where players formed a chain to pass an image from the head to the tail, testing whether the transmission stabilities of images exhibited by humans could be predicted by the model. In Experiment 4, we set up a visual Turing test (Lake et al., 2015) to investigate whether the model could generate responses indistinguishable from those of a human.

In the following section, we will introduce how we construct the causal model of Mondrian-style image and conduct specific experiments to test it under the "big task" paradigm.

A Generative Model on Mondrian-Style Images

In our study, we built a model on how to generate Mondrian-style images, which can be represented as a "parse tree," a tree structure originally introduced for recursively parsing a sentence into words as the terminal nodes of the tree (Chomsky, 1965). In our case, it is the Mondrian-style image that is recursively parsed into rectangles. For computational tractability, here we set a constraint to the number of terminal nodes in a tree structure to 6. By doing this, we create an image space containing all structured images that can be generated by recursively splitting a scene with five horizontal or vertical cuts.

As illustrated in the causal graph (Figure 3), the exact method for modeling the distribution of images follows a three-step image-generation process: First, a parse tree is created; second, the scene is split into different pieces given the tree; and third, every pixel in the image is filled given the hierarchical partitions. By multiplying the distributions of each procedure in the graph, we obtain the joint distribution of all variables. The probability of an image is computed by marginalizing the latent variables on the joint distribution, following the mathematical formulations (Equations 2 and 3):

$$\begin{aligned} P(\text{image, partition, tree}) = P(\text{tree}) \times P(\text{partition}|\text{tree}) \\ \times P(\text{image}|\text{partition}) \end{aligned} \tag{2}$$

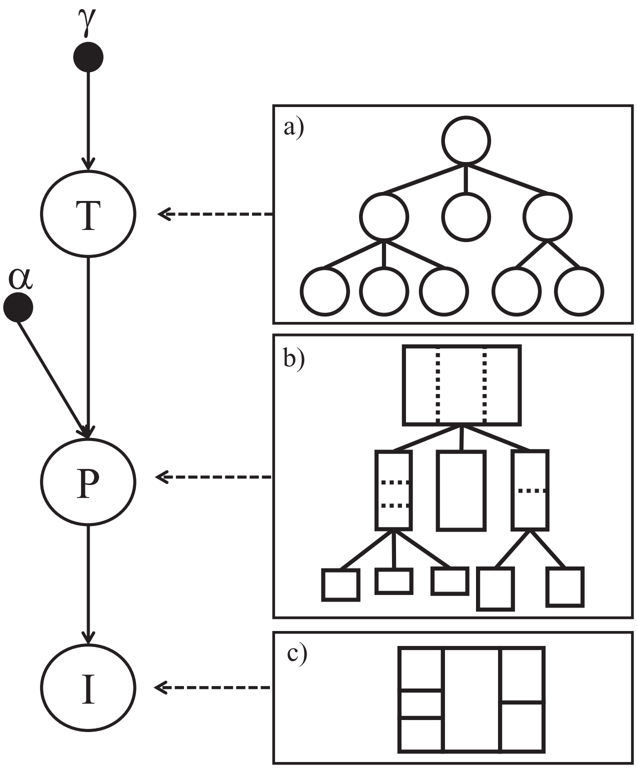
$$P(\text{image}) = \sum_{\text{partition, tree}} P(\text{image, partition, tree}) \tag{3}$$

Probability of Parse Tree P(tree)

A limited number of nodes can nevertheless create a large space of tree structure. To assign a probability to each tree in a systematic way, we use

Figure 3

Illustrations of a Hierarchical Bayesian Model for Synthesizing Mondrian-Style Images



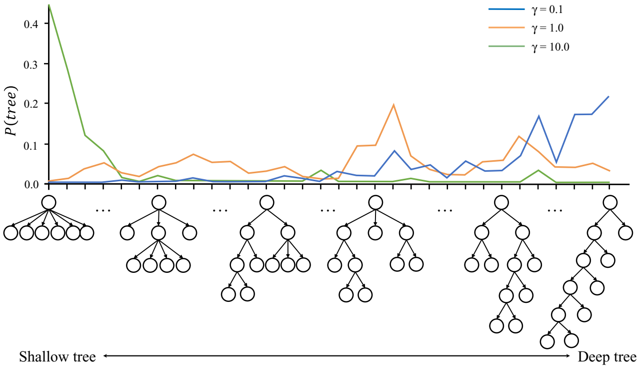
Note. Images in the boxes on the right illustrate variables (see next section for a detailed explanation) in the Bayesian network. (a) A parse tree (T) with the branching factor (γ) as the hyperparameter. (b) A hierarchical partition (P) with the evenness factor (α) as the hyperparameter. (c) The synthesized Mondrian-style image (I) given the hierarchical partition (P).

the nested Chinese Restaurant Process (nCRP) to formulate the tree prior. The nCRP recursively assigns a primitive unit to different branches of a tree, with a certain probability of creating a new branch (Blei et al., 2010). This provides nCRP the potential to create trees that allow for an unfixed number of branching and assign a probability to each of them. As the number of terminal nodes in a tree is fixed to 6, a total of 32 trees can be generated. The only free parameter of the nCRP controls how likely it will create a new branch. We refer to it as the "branching factor," following the convention of the traditional Artificial Intelligence research on tree search. A higher branching factor is more likely to generate wide and shallow trees, whereas a lower branching factor is more likely to generate narrow and deep trees. Thus, for all possible trees with limited branches in our experiments, we create a categorical distribution over the limited number of trees (Figure 4), with the parameter of branching factor controlling how probability mass is allocated among these trees. Details of the nCRP computation of trees are documented in Appendix A, and a more comprehensive tutorial on the nCRP can be found in another study (Blei et al., 2010).

Probability of Partition Given Parse Tree $P(\text{partition}|\text{tree})$

Figure 4

Illustration of the Effect of Branching Factor on P(*tree*)

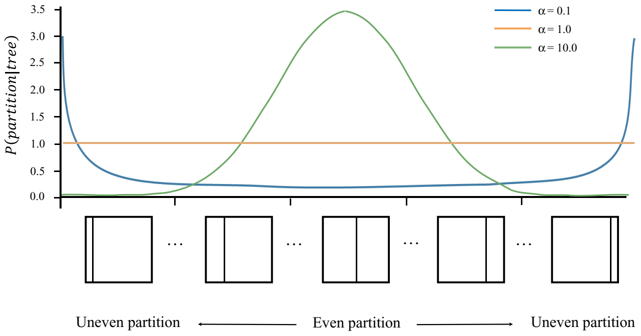


Note. Wide and shallow trees are more likely to be generated with a higher branching factor, whereas narrow and deep trees are more likely to be generated with a lower branching factor. See the online article for the color version of this figure.

After the number of branches under each node of a tree is determined, the model can then partition a blank scene into the components of a Mondrian-style image. To model the uncertainty between a vertical or horizontal cut, we assign equal probability to each orientation. Yet another uncertainty arises regarding the sizes of the cuts: What percentage of a part should each subpart occupy? We model this by introducing an "evenness factor" parameter that controls how likely a part will be split in an even fashion (Figure 5). The higher the evenness factor, the more likely it is to have subparts of similar sizes. It is modeled by a Dirichlet distribution, which can be intuitively understood as a dice factory that produces even dice or biased dice, depending on the evenness factor (see Appendix A for further details). The evenness factor ranges from 0 to 10 for practical reasons, in which case the probability of even partitions increases with a greater value. An evenness factor of 10 indicates a partition closely resembling an even split, whereas an evenness factor of 0 indicates a highly uneven partition. An evenness factor of 1 indicates that all partitions, even and uneven, have identical probabilities.

Figure 5

Illustration of the Effect of the Evenness Factor on P(*partition*|*tree*) *With Two Partitions*



Note. A high evenness factor is likely to result in a partition closely resembling an even split, whereas a low evenness factor is likely to result in a highly uneven partition. See the online article for the color version of this figure.

In our study, the branching factor and evenness factor were learned from human-generated images in Experiment 1. They serve as human prior distributions that are used to estimate a *P*(image) for our central model, which we call the human prior hierarchical Bayesian model (HP).

Baseline Models

Our experiments mainly rely on *P*(image) to predict human performance. Hence, success in prediction is largely contingent upon whether *P*(image) accurately captures the subjective estimation of images in the human mind. To demonstrate that a causal model learned from human-generated data (HP) is essential in predicting humans' visual interpretation, we also incorporate two baseline models in the current study: (a) a "noninformative prior model" (NP) that shares the same hierarchical Bayesian structure but lacks human priors, instead of sampling the branching factor and evenness factor from a noninformative uniform prior (details of the priors can be found in Appendix B); and (b) a "feature multivariate Gaussian model" (MG) that learns the summary statistics of features from human-generated images, without inferring the causal structure of those images.

Transparency and Openness

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study, following Journal Article Reporting Standards (JARS, Kazak, 2018). All data, analysis code, and research materials are available at https://osf.io/u62xb/ (Tang et al., 2021). Models were implemented using Python, version 2.7 (van Rossum, 1995) and the package NetworkX, version 2.1.1 (Hagberg et al., 2008). Data were analyzed using Python, version 2.7 (van Rossum, 1995). The design and analysis of this study have not been preregistered.

All procedures of the experiments in this study were approved by the ethics committee of the Department of Psychology and Behavioral Sciences at Zhejiang University (XL201709021—"Testing analysis-by-synthesis on Mondrian-style image").

Experiments

Experiment 1: Just Cut It

To elicit the subjective *P*(image) from participants' minds, we first set up an image generation task to collect human synthesized images for training the models. In this image-generation task, participants were requested to generate images by cutting a rectangle into six pieces with horizontal and vertical cuts. They could apply their cuts freely with a single constraint: only split one existing rectangle at a time.

Method

Participants. Sixty (fifty for the training dataset, ten for the testing dataset) undergraduate and graduate students at Zhejiang University participated in this experiment in exchange for payment.

Procedure. Participants were only instructed for the legal actions to split an image. The word "Mondrian" was never mentioned, nor was there any example image showing how to apply cuts. Instead, participants split the image in whichever way they felt intuitive. Figures are generated in two environments. For the training dataset, twenty-five participants each drew their cuts on blank paper with a pen, and another 25 participants drew their cuts on the computer screen using a mouse (Figure 6). For the testing dataset, ten participants drew their cuts on the computer screen. Each participant was asked to generate two figures.

Results

Data Preprocessing. A demo showing how an image was generated on the computer step by step can be found at https://sites.google.com/view/generativevision-commonsense. Sample images generated in the two environments are shown in Appendix B. The images drawn on the paper were pre-processed with a computer vision line detection algorithm (Bradski, 2000) to match the format of images drawn on the computer screen (Figure 7). Two groups of images were merged, as the summary statistics on them showed no significant difference (see Appendix B for further details).

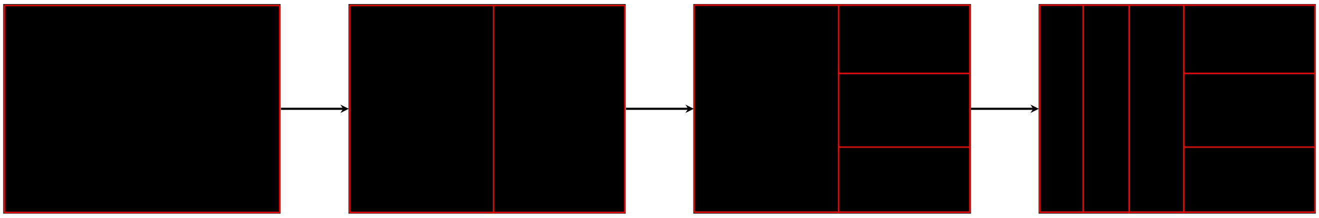
Model Training. Human-generated images were used to train the following two models, using Maximize A Posterior Estimation method (Bishop, 2006, Section 2.3).

For the HP model, the parameters of branching factor γ and evenness factor α were learned from human data (Equation 4) and treated as human priors in the following experiments.

$$P(\gamma, \alpha|\text{images}) \propto \prod_{\text{images}} P_{\gamma,\alpha}(\text{image}) \times P(\gamma, \alpha) \tag{4}$$

The branching factor and the evenness factor were estimated by the grid search method (Gelman et al., 2013, Chapter 5). Their posterior estimation is illustrated as a heat map (Figure 8), with red indicating a higher posterior and blue indicating a low posterior. The

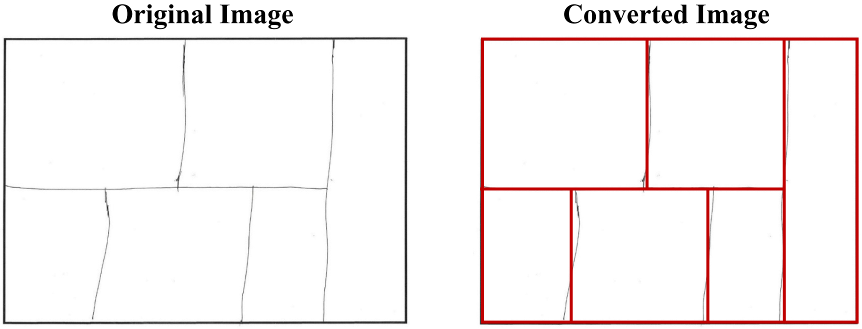
Figure 6
Sequence of Applying Cuts on a Computer Screen Using a Mouse



Note. See the online article for the color version of this figure.

Figure 7

Illustration of Converting an Image on Paper Into the Format on a Computer Screen



Note. See the online article for the color version of this figure.

results show that the posterior was maximized when $\gamma = 0.9$ and $\alpha = 3.5$. Intuitively, the results reflect that human-generated structured images were “balanced” in the following ways. In terms of the structure of the tree, humans preferred to produce trees with 2–3 layers—neither too deep/narrow nor too shallow/wide. In terms of the ratio of each split, humans preferred even partitions within each layer, but the preference was not exceedingly strict.

For the MG model, we select the logarithms of aspect ratios and the sizes of all six primitive rectangles as the features to describe images. In this sense, an image now is represented as a vector of the logarithms of aspect ratios and sizes ($x_{1\text{ratio}}$, $x_{2\text{ratio}}, \ldots, x_{6\text{ratio}}, x_{1\text{size}}, x_{2\text{size}}, \ldots, x_{6\text{size}}$). By the definition of the Gaussian linear regression model, the probability of an image is a multivariate-Gaussian distribution with feature means of $\mu = (\mu_{1\text{ratio}}, \mu_{2\text{ratio}}, \ldots, \mu_{6\text{ratio}}, \mu_{1\text{size}}, \mu_{2\text{size}}, \ldots, \mu_{6\text{size}})$ and a covariance of Σ (a 12*12 matrix). The parameters of mean and covariance were learned from human data (Equation 5).

$$P(\mu, \Sigma|\text{images}) \propto \prod_{\text{images}} P_{\mu, \Sigma}(\text{image}) \times P(\mu, \Sigma) \quad (5)$$

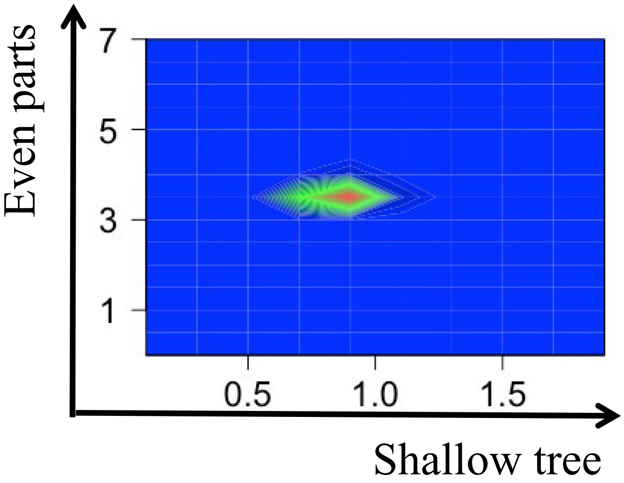
Model Evaluation. Since model training concerns optimizing a model’s estimated probability of given images, the models were evaluated by their estimation on new images’ probabilities: Models that predict new images with a larger likelihood are more desired. Here, we adopted an approach commonly used in model evaluation by computing the average log P(new images) of each model from the parameters learned from human-synthesized images. Following Bayesian model evaluation, each model was evaluated 30 times by drawing a sample of parameters from a posterior distribution over the parameter space instead of setting to the maximum a posteriori. Each sample of parameters was applied to 20 testing images from which the average log P(new images) was computed.

We began by demonstrating that the training for both the HP and MG models was effective. Two types of images were used: One set of 20 images was generated by a new group of 10 human participants, and another set was 20 nonhuman-synthesized images generated by the NP model with an uninformative prior distribution that did not capture any human bias. We predicted that the HP and MG models, trained by human-synthesized images, could explain new human-synthesized images significantly better than nonhuman-synthesized images, while the NP model would show no difference in explaining the two. The results are shown in Figure 9. As predicted, the differences between the average log probabilities of two types of new images were significant for the HP model, human-synthesized: $M = -2.35$, 95% CI $[-2.33, -2.37]$ and nonhuman-synthesized: $M = -5.77$, $[-5.81, -5.73]$; $t(58) = 155.43$, $p < .001$, $d = 40.13$, and the MG model, human-synthesized: $M = -11.80$, $[-11.88, -11.73]$ and nonhuman-synthesized: $M = -29.10$, $[-29.30, -28.90]$; $t(58) = 164.63$, $p < .001$, $d = 42.51$, but not for the NP model, human-synthesized: $M = -6.60$, $[-8.00, -5.21]$ and nonhuman-synthesized: $M = -7.16$, $[-8.07, -6.25]$; $t(58) = .68$, $p = .50$, $d = 0.18$; $\text{BF}_{01} = 3.13$. These results showed that both the HP and MG models were effectively trained for modeling human preference.

We further examined which of the trained models, the HP model or the MG model, could better capture human preference. The results in Figure 9 showed that the HP model significantly outperformed the MG model in explaining new human images, $t(58) = 251.90$, $p < .001$, $d = 65.04$. This result occurred despite

Figure 8

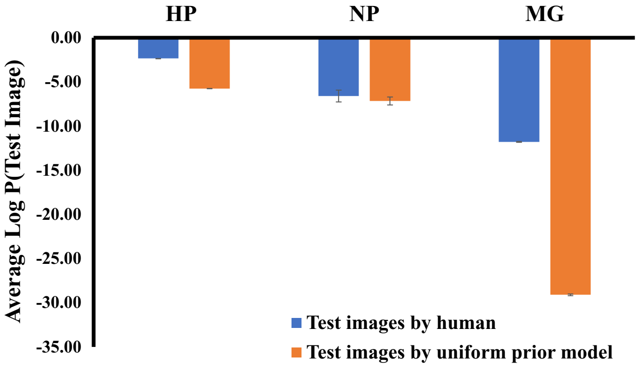
Heat Map Showing P(*images*) *in Different Combinations of Branching Factor and Evenness Factor*



Note. Dark gray away from the center indicates a low probability, medium gray near the center indicates a middle probability and light gray in the center indicates a high probablity. In the color version, blue indicates a low probability, green indicates a middle probability and red indicates a high probability. See the online article for the color version of this figure.

Figure 9

Results of Average Log Probabilities of Three Models in Explaining New Images Synthesized by Humans and the NP Model



Note. Error bars indicate standard errors. See the online article for the color version of this figure.

a greater number of free parameters in the MG model and is expected since the HP model accurately captures the causal generative process. The key focus of our study is how well these models can be generalized to various visual tasks besides generative processes per se.

Experiment 2: Perceptual Complexity

Here we test whether the model could predict the perceptual complexity of images, which can be quantitatively measured by P(image). Intuitively, images with low probability will be perceived as more complex with more information. This intuition is captured by information theory (Shannon, 1948) in which the information content of an image is defined as the negative logarithm of P(image), as shown in Equation 6.

$$\text{Information content (image)} = -\log_2 P_{\gamma,\alpha}(\text{image}) \qquad (6)$$

In this experiment, 20 images were selected and sorted through the pairwise comparison paradigm to estimate perceived subjective complexity in humans: In each trial, participants were simply presented with two images from the 20-image pool and instructed to make a choice on which one seemed to be more complex. After multiple trials of comparison, the ranking of all images was computed. The three models we used (HP, NP, and MG) predicted the ranking based on each of their computations of P(image). For each model in each trial, the probability of identifying an image as more complex was inversely proportional to its estimated P(image). To capture human's complexity ranking, the model's estimation of P(image) must be consistent with that of humans. Importantly, participants were not instructed to report anything related to the probabilistic interpretation of images.

Method

Participants. Sixteen undergraduate and graduate students at Zhejiang University participated in this experiment in exchange for payment. Our research design and predicted effects were novel. We decided in advance to use a sample size of 16 in all the following visual experiments as we expected a large effect size ($d \geq 0.8$).

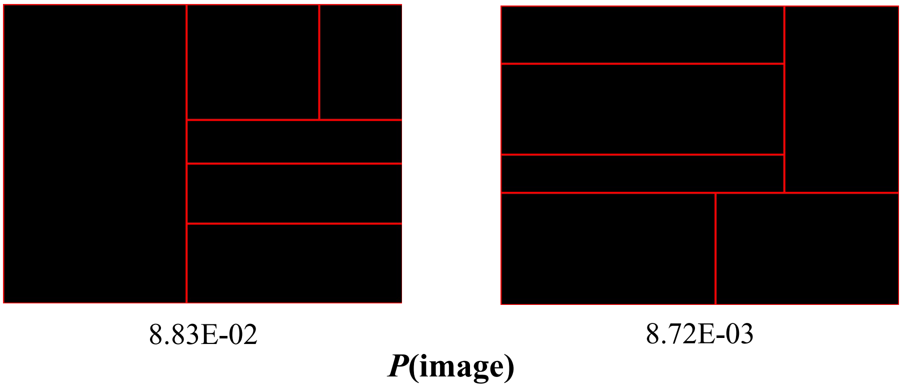
Materials. Twenty images were sampled from the generative model with great variance in their structures including tree topology and partition evenness. Details regarding the sampled image can be found in Appendix B.

Procedure. One hundred ninety possible pairwise combinations could be derived from the 20 images. Each combination was presented with a random image position (left-right or right-left), resulting in a total of 190 trials. As the model is stochastic, each run of the model generates a different result. We thus treated each run of the model as a machine participant. To match the number of human participants (16), each of the three models ran the paired-comparison 16 times.

In each trial, after a 200 ms fixation and a 300 ms blank screen, two images (20.48° × 15.36°) were presented on the screen with 25.60° between their centers horizontally. A sample of paired image is shown in Figure 10. Participants then reported which of the two images they perceived as more complex by pressing one

Figure 10

A Sample of Two Images for Pair-Comparison With Their P*(image)s*



Note. Only for illustration purpose, not shown to subjects. See the online article for the color version of this figure.

of two keys (“F” for the left image, “K” for the right). There was no restriction on reaction time. Models reported which images were more complex by estimating their *P*(image)s and then sampling responses following Equation 7.

$$P(\text{right image is more complex}) = 1 - \frac{P(\text{right image})}{P(\text{left image}) + P(\text{right image})} \quad (7)$$

Results

Human data were computed by the following procedure. First, for each participant, the percentage of each image perceived as more complex in all comparisons was computed and then converted into a *Z*-score, indicating the ranking of an image’s complexity. Then, for each image, the *Z*-scores from all participants were added up as the estimated human-perceived complexity of that image. All models’ *Z*-scores were computed exactly the same way.

The correlations between humans and each model’s ranked complexity of all 20 images are shown in Figure 11, with the *x*-axis as the model-ranking *Z*-score and the *y*-axis as the human-ranking *Z*-score. For the HP model, a strong positive correlation was found ($r = .91$, 95% CI [.79, .97], $p < .001$). For the NP model, a weaker but still positive correlation was found ($r = .62$, [.24, .83], $p = .004$). For the MG model, there was no significant correlation ($r = -.03$, [−.47, .42], $p = .90$; $\text{BF}_{01} = 3.59$). Bayesian Analysis showed that there is extreme evidence in favor of the HP model being better than the NP model in explaining the human-ranking *Z*-scores ($\text{BF}_{\text{HP, NP}} = 3.88 \times 10^4$) and strong evidence in favor of the NP model being better than the MG model in explaining the human-ranking *Z*-scores ($\text{BF}_{\text{NP, MG}} = 26.13$). Thus, we found additional evidence supporting that the causality in the Bayesian hierarchy model is essential in capturing human perceived complexity, indicating that humans recovered the underlying generative process when they interpreted the images. At the same time, we learned that priors learned from human-generated data also play an important role in encapsulating humans’ interpretation of an image.

Experiment 3: Human Iteration

Here we use a different behavioral paradigm utilizing *P*(image) to demonstrate the common-sense nature of generative vision. Specifically, we aim to show how, within a series of human communications, images with different *P*(image)s all gradually converge into the “prototype images” with high *P*(image)s.

We adopted the “human iterative message-passing” paradigm that was originally introduced in studies on false memory (Bartlett, 1932) and has been revived in recent studies as a powerful tool for eliciting human priors (Griffiths & Kalish, 2007; Kalish et al., 2007; Uddenberg & Scholl, 2018). It is essentially a “telephone game” of images: An observer sees an image shortly and then regenerates that image which will be passed to the next observer. The iterative image-passing procedure will be repeated until the end of the chain.

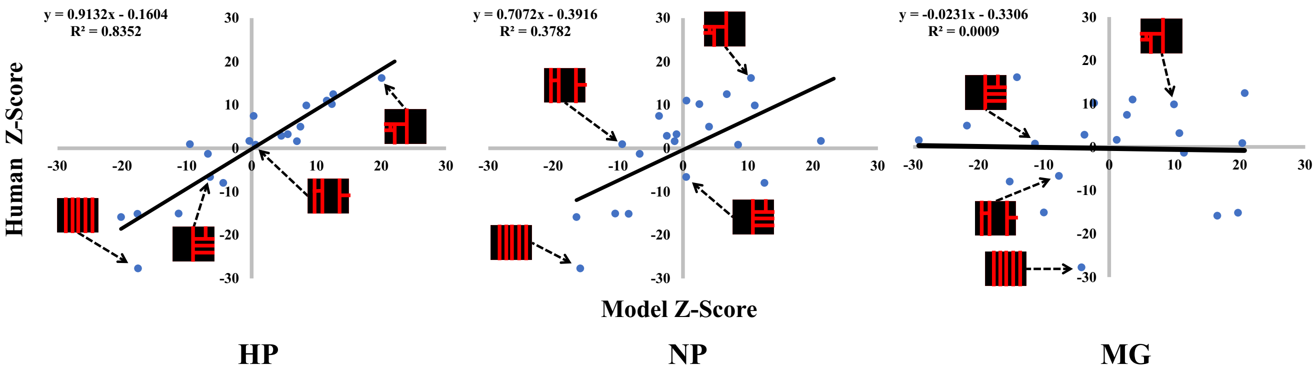
The iterative process can be considered an amplifier of human priors, as in each iteration an observer injects their subjective bias into their reproduction of the image. Due to the limited precision of human perception and memory, an image will mutate as it is regenerated along with the procedure. However, we predict that this mutation pattern is not random. With a model of *P*(image), we expect that all images will gradually mutate to the ones with high *P*(image)s in human prior. This mutation can be measured by a loss of information content (Equation 8): images at the end of the chains carry less information than those at the start.

$$\begin{aligned}\text{Information loss} &= \text{Information(initial image)}\\ &\quad - \text{Information(initial image)}\\ &= -\log_2 P_{\gamma,\alpha}(\text{initial image})\\ &\quad - (-\log_2 P_{\gamma,\alpha}(\text{end image}))\end{aligned} \quad (8)$$

Furthermore, the magnitude of mutation is not a constant value. An initial image with a high information content and a low *P*(image) should mutate more dramatically, indicating low transmission stability; inversely, an initial image with a low information content and a high *P*(image) should mutate less, indicating high transmission stability. This positive correlation between

Figure 11

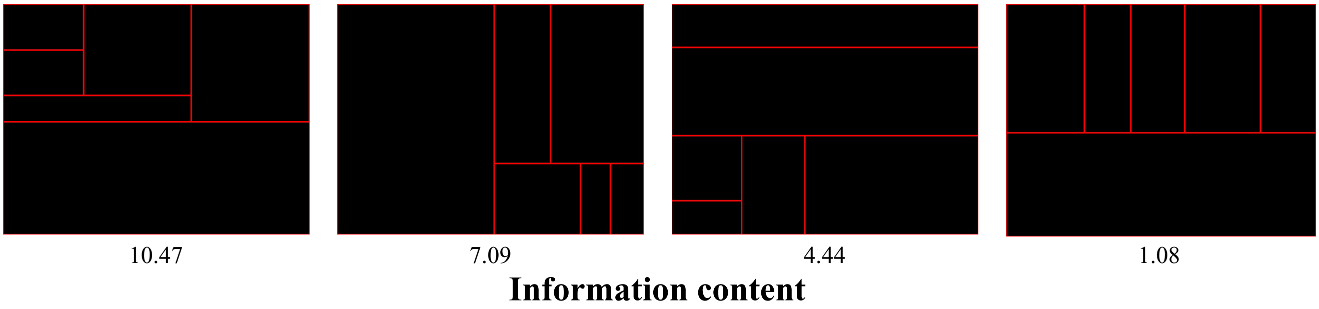
Results of the Correlations Between the Model Z*-Score and Human* Z*-Score of Subjective Complexity for All Three Models*



Note. Each point represents one image, and four representative images are illustrated here. See the online article for the color version of this figure.

Figure 12

Samples of Initial Images Given as the Start of the Chains With Their Information Contents



Note. See the online article for the color version of this figure.

initial information content and information loss is the key prediction in our experiment. However, while the P(image) plays a central role in describing the mutation process, it cannot be obtained explicitly. That is, we cannot directly ask human participants to report their subjective P(image) but instead have to estimate it by a model. Therefore, if the model estimates human P(image) accurately, we should be able to observe a positive correlation between initial information content and information loss estimated by the model. Conversely, if the estimation is less accurate, the model will be worse at explaining how images mutate in the iterative image-passing procedure, manifested by a weaker or insignificant correlation.

Method

Participants. Forty-eight undergraduate and graduate students at Zhejiang University participated in this experiment in exchange for payment. Participants were evenly split into three chains, each with 16 participants.

Materials. Twenty new images were sampled with a procedure identical to the one in Experiment 2 (Figure 12).

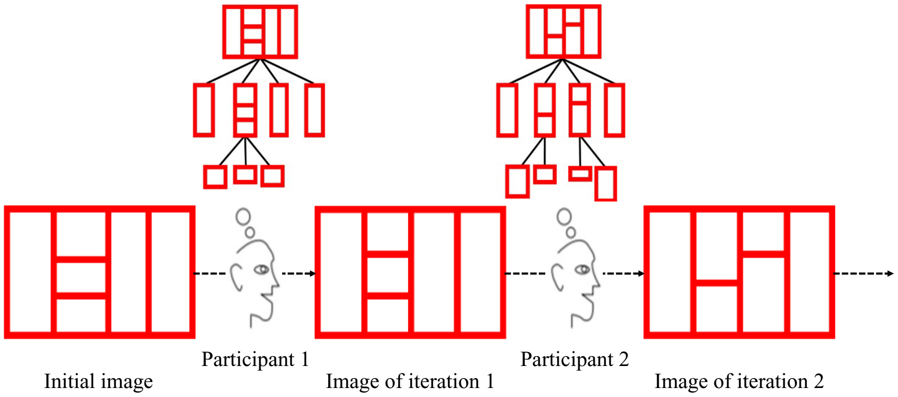
Procedure. Each chain contains 16 iterations, with one iteration for one participant (Figure 13). Each iteration contains 20 trials, with one trial for one image. Each trial started with a 200 ms fixation and a 300 ms blank screen, followed by an image (40.96° by 30.72°) presented at the center of the screen for 1,000 ms. Then the participant regenerated the image on a computer screen by the same procedure as in Experiment 1.

Results

For each model, all three chains shared the same initial information content since their start images were identical. Samples of the image passing chains with initial images of different information contents are shown in Figure 14. The correlations between initial information content and information loss in the three chains for all models are shown in Figure 15, with each dot representing an image. For the HP model, positive correlations were found in all three chains (Chain A: $r = .75$, 95% CI [.46, .90], $p < .001$; Chain B: $r = .77$, [.49, .90], $p < .001$; Chain C: $r = .81$, [.57, .92], $p < .001$). For the NP model, positive correlations were also found in all three chains, but with smaller coefficients compared to those of the HP model

Figure 13

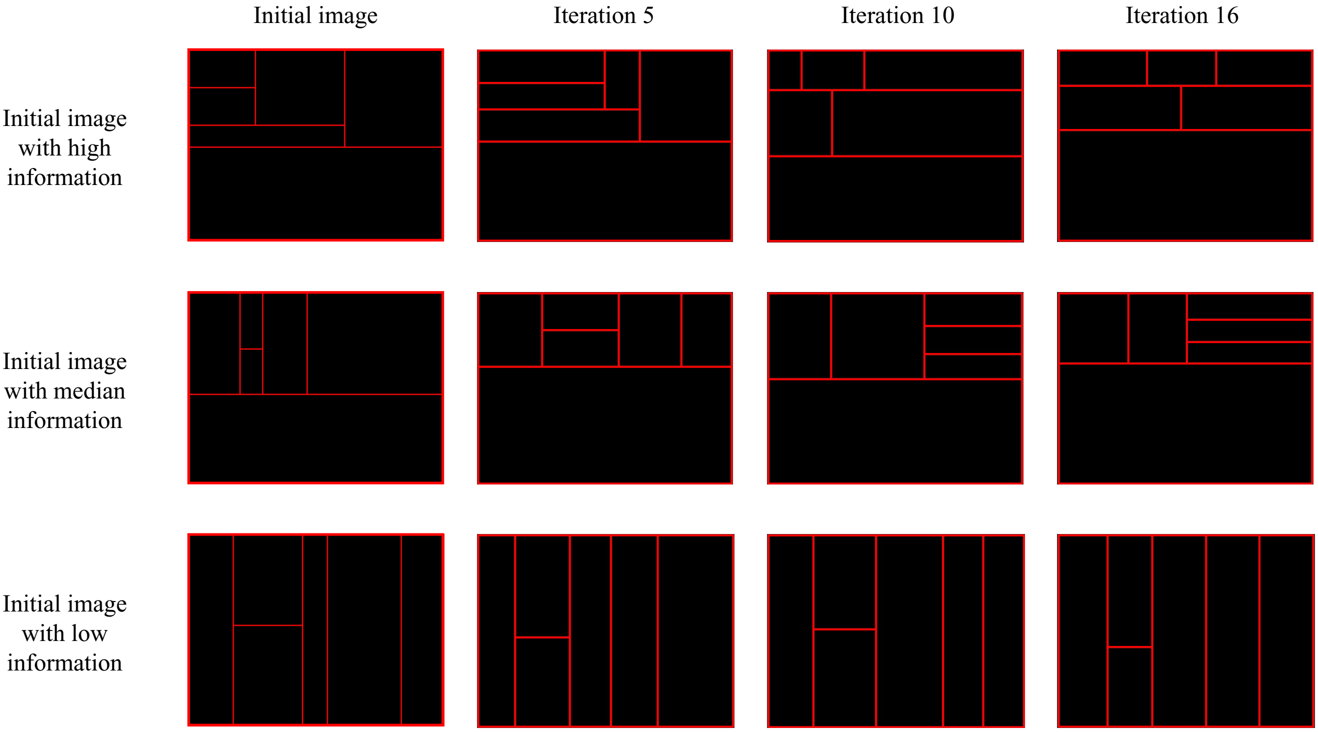
Illustration of a Chain of Image-Passing Iteration



Note. One participant views an image generated by the previous participant and regenerates that image to pass to the next participant. See the online article for the color version of this figure.

Figure 14

Samples of Image-Passing Iterations With Initial Images of Different Information Contents


Initial image
Iteration 5
Iteration 10
Iteration 16
Initial image with high information
Initial image with median information
Initial image with low information

Note. See the online article for the color version of this figure.

(Chain A: $r = .55$, [.14, .80], $p = .012$; Chain B: $r = .45$, [.003, .74], $p = .049$; Chain C: $r = .65$, [.30, .85], $p = .002$). For the MG model, however, the correlations were not significant in all three chains (Chain A: $r = .30$, [−.17, .65], $p = .21$, $BF_{01} = 1.72$; Chain B: $r = .34$, [−.12, .68], $p = .15$, $BF_{01} = 1.34$; Chain C: $r = .35$, [−.12, .68], $p = .14$, $BF_{01} = 1.28$). The results show that the MG model cannot estimate human subjective P(image) accurately, whereas both the two hierarchical Bayesian models can. This suggests that the causal generative process in Bayesian models plays a critical role in explaining how images mutate through an iterative image-passing procedure. Further, the greater strength in the correlation for the HP model indicates that human priors learned from human-generated data are also essential in capturing the message-passing process.

Experiment 4: Visual Turing Test

The Turing test is a test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of humans. It is a thought experiment that has been arguably regarded as the ultimate challenge for artificial intelligence (Turing, 1950). To date, the Turing test has been adopted as an empirical research paradigm in examining the validity of a cognitive model of the human mind. The key to this paradigm is that the response made by a model or human is not entirely open but limited within a certain space. For example, in a visual Turing test, the answers generated by humans or models were limited to a particular type of written letter (Lake et al., 2015).

Here, we applied a visual Turing test to Mondrian-style images. Although conventionally a Turing test examined judgments based on the final stimuli, here we presented participants with parse trees for making judgments since our focus was on the causal generative process itself that humans or models recovered. In preparation for the test, given a Mondrian-style image, participants were instructed to recover a parse tree that they thought would best explain the image. The same procedure was performed on a model, and all parse trees from both humans and the model were collected for the Turing test. Finally, a new group of human observers saw the dynamic process of how an image was parsed step by step, with a parse tree growing from the root. They then needed to identify whether the tree was generated by a human or a model.

For both the HP and NP models, the parse tree was sampled from the posterior distribution of tree structures conditioned on the image by integrating out latent possible partitions (Equation 9). However, the MG model failed to apply to this task as the task was causal in nature. Instead, we introduced the uniform sample model (US) as a new baseline model. It samples from all topologically correct trees, as shown in Figure 1.

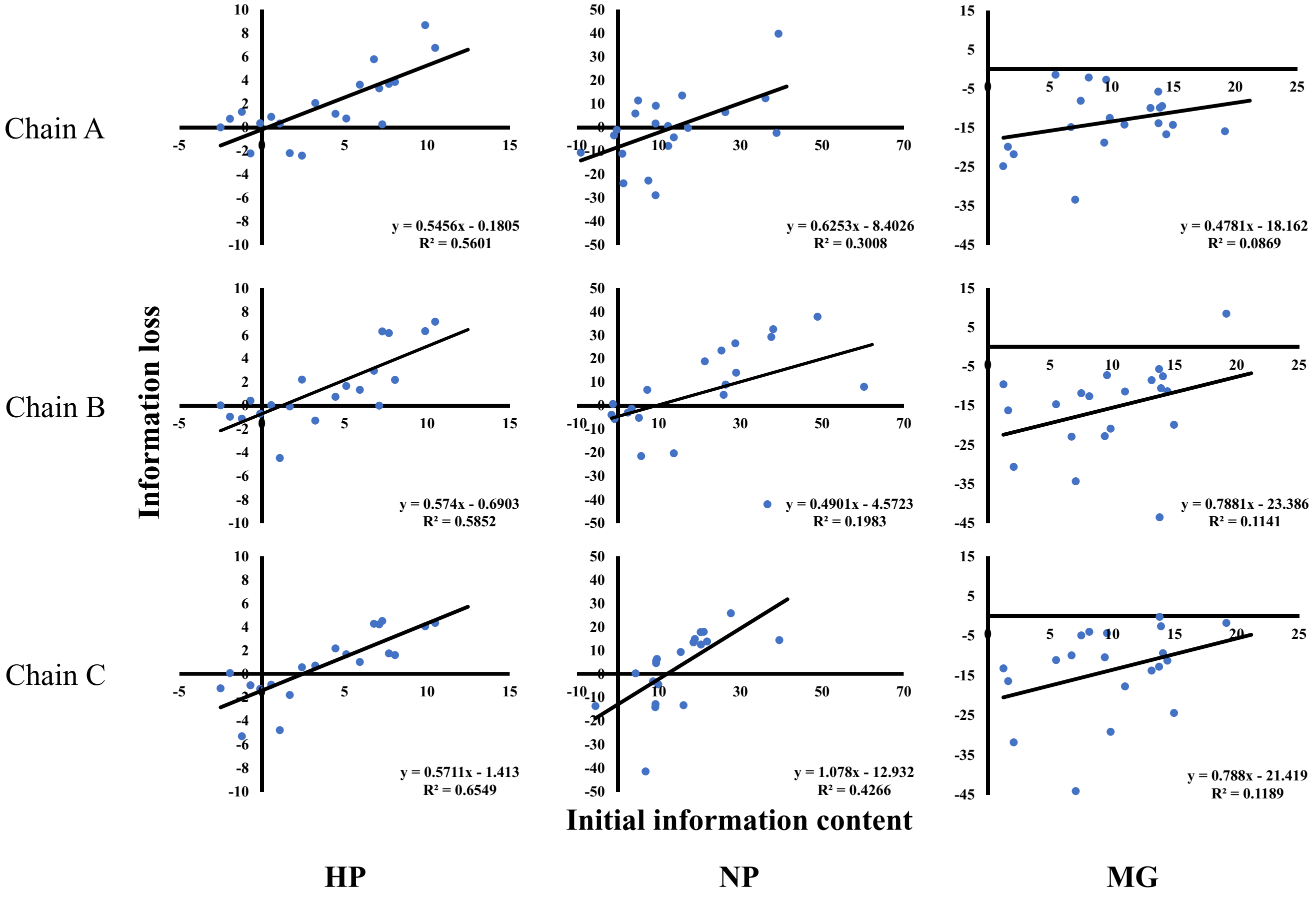
$$\begin{aligned} P_{\gamma,\alpha}(\text{tree}|\text{image}) &= \sum_{\text{partition}} P_{\gamma,\alpha}(\text{tree}|\text{partition, img}) \\ &= \sum_{\text{partition}} \frac{P_{\gamma,\alpha}(\text{image, partition, tree})}{P_{\gamma,\alpha}(\text{image, partition})} \end{aligned} \quad (9)$$

Method

Participants. Sixty-four undergraduate and graduate students at Zhejiang University participated in this experiment in exchange

Figure 15

Results of the Correlations Between the Information Loss and the Initial Information Content



Note. The *y*-axes of different models are in different scales for illustrative purposes. Each point represents an image. See the online article for the color version of this figure.

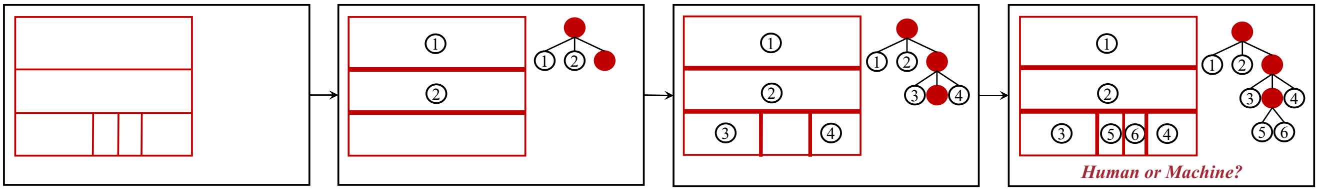
for payment or course credit. Sixteen of them were the parsers that provided parse trees which we used as human answers. The other 48 participants were observers in the human–machine identification task, evenly assigned to the three model groups. Each group only discriminated human answers from those of one model.

Materials. Sixteen new images were generated using a procedure similar to the one in previous experiments (See Appendix B for further details).

Procedure. In the parsing task (parse tree generation), for each of 20 Mondrian-style images, each participant was asked to draw a parse tree (Figure 3a) with a pen that represented the most probable generation process they believed for the observed image. These parse trees were then digitalized to serve as the stimuli of human parse tree in the identification task. To ensure that participants sufficiently understood how a parse tree could describe a step-by-step partition process of an image, participants were first instructed to

Figure 16

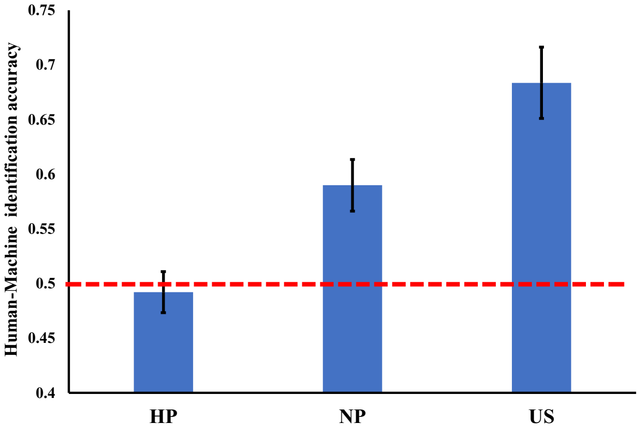
The Procedure of the Identification Task



Note. Participants were first presented with a Mondrian-style image. Then they would see a parse tree growing layer by layer on the right side of the screen, accompanied by a visualization of how partitions were applied on the Mondrian-style image on the left. See the online article for the color version of this figure.

Figure 17

Results of the Visual Turing Test Showing the Percentage of Trials Correctly Identified by Third-Party Observers on Whether a Parse Tree was Provided by Humans or One of the Three Models



Note. Error bars indicate standard errors. See the online article for the color version of this figure.

understand the correspondence between a parsing process and a parse tree. Essentially, a parse tree would grow in layers from the root node, with each node in a layer representing a subpartition from the existing partitions of the previous layers. Then, a test for participants was implemented in which a novel image and a parse tree were provided. Each participant needed to explain correctly to the experimenter how to parse the image layer by layer according to the parse tree.

For the identification task (human/machine parse tree discrimination, Figure 16), participants were given the same instruction and test about how to interpret the parse tree in terms of a parsing process. During the task, each trial started with a 200 ms fixation followed by a 300 ms blank screen. Participants were first presented with a Mondrian-style image. The image was then partitioned step by step, illustrated by two visualizations synchronously. On the right side of the screen, a parse tree grew in layers. Correspondingly, on the left side of the screen, a new layer in the parse tree could be visualized on the Mondrian-style image where the borders of newly partitioned parts were thickened. This process is self-paced—the next layer of image partition and its corresponding parse tree would only unfold after participants fully understood the partition and clicked. When the parse tree was completed, participants were asked to judge whether the partitions were made by humans or models by pressing one of the two buttons (“F” and “J” on the keyboard). The dynamic demonstration of a single trial can be found at https://sites.google.com/view/generativevision-commonsense.

Results

Human-model identification accuracy is shown in Figure 17. For the HP model, the accuracy was not significantly above the chance level, $M = .49$, 95% CI [.45, .53], $t(15) = -0.42$, $p = .68$, $d = 0.10$; $\mathrm{BF}_{01} = 3.62$, thus suggesting a pass of the visual Turing test. In contrast, for both the NP and US models, their accuracies were significantly above chance level, NP: $M = .59$, [.54, .64], $t(15) = 3.82$, $p = .002$, $d = .96$; US: $M = .68$, [.61, .75], $t(15) = 4.86$, $p < .001$, $d = 1.22$, thus suggesting they failed the visual Turing test. However, a direct t-test shows that the performance of the NP group was much lower than that of the US group, $t(30) = -2.34$, $p = .026$, $d = 0.83$, suggesting that responses from the NP model were more difficult for observers to discriminate than those from the US model. These results are consistent with Experiments 2 and 3, once again showing that the causal structure of hierarchical Bayesian captures important aspects of humans’ visual interpretation. Notably, to pass the visual Turing test, the causal structure must be integrated with the human priors in image generation, once again supporting the analysis-by-synthesis perspective of human vision.

Discussion

In the current study, we demonstrate the generative nature of human vision in the case of Mondrian-style images, using a visual common-sense perspective. The tasks used in our study constitute two core attributes of image processing: generation (synthesis) and interpretation (analysis). First, we implemented an image-generation task to collect human-synthesized Mondrian-style images. Two parameters, the branching factor and the evenness factor, were learned from those images and served as human prior distributions, which we used to estimate a P(image) for the HP model. We found the HP model based upon a generation task could explain human performance in various visual tasks: In the complexity ranking task, the HP model’s subjective rating on the complexity of Mondrian-style images was highly concordant with those of humans; in the telephone game task, it could capture the amount of information loss that measures how images mutated in an iterative image-passing procedure; lastly, in the visual Turing task, the HP model could recover the causal history of Mondrian-style images as parse trees similar to those recovered by humans, passing the Turing test as it fooled third-party humans’ judgments on the source of those parse trees.

In all experiments, we set two baseline models: the causal NP model that lacks human priors and the noncausal MG model that learns summary statistics of visual features from human-generated images. We found that (a) the HP and NP models showed greater explanatory power on human performance than the MG model that was absent of a causal process (The lower performance of the MG model is not due to overfitting. See the same pattern of results by comparing a regularized MG model with the HP and NP models in Experiments 1–3 in Appendix B), indicating that expressing the causal synthesizing process is essential for explaining human analysis on Mondrian-style images; (b) across all experiments, the HP model predicted human performance more accurately than the NP model free of human synthesized data, suggesting that human preferences in the synthesizing processes are also a necessity; and (c) the NP model failed the Turing test despite outperforming the US model in the test. From the above evidence, regarding the significance of a causal structure and human priors elicited from image generation, we demonstrate the central notion of analysis-by-synthesis in the case study of Mondrian-style images. In the following sections, we will discuss the implications of results from this case study exploring a deep integration of a generative model and human experiments. We hope these results can facilitate future studies on generative vision that apply beyond Mondrian-style images.

Implicit Measurement of Generative Vision

In the current study, we tested generative vision in both explicit and implicit tasks. The visual Turing test we adopted in Experiment 4 was the most explicit of all tasks. Participants were directly instructed to recover the generative process of observed Mondrian-style images by providing parse trees. Although the visual Turing test has only been recently adopted in cognitive science (Lake et al., 2015), explicitly asking participants questions regarding the causal process underlying an image has been an approach commonly seen in early Bayesian perceptions studies (Kersten et al., 2004).

The telephone game task in Experiment 3 takes place somewhere in-between explicit and implicit measurements. Each participant was instructed to regenerate a Mondrian-style image, which included a generative process. However, they were not instructed to come up with the most likely generative process of the observed image or follow it in their recreation of that image. That is, participants had complete freedom in their regeneration process, while a product resembling the observed image was the only requirement. Yet, we found that their products could be best explained by generative processes captured by the HP model.

The complexity ranking task in Experiment 2 is an entirely implicit measurement: The concept of a generative process was never mentioned in the task when participants were judging the complexity of different Mondrian-style images. Nevertheless, we found that the HP model built upon the generative process best explained subjective complexity, indicating that humans recover the causal history of images in a spontaneous, automatic manner. Our results are consistent with the previous study, showing that the causal history, despite never being mentioned in the tasks, was found to bias humans into perceiving a dynamic transformation on images even when such real-time changes were unobservable (Chen & Scholl, 2016; Freyd, 1987).

Successes of implicit tasks suggest that generative vision can be applied to a wide scope of tasks seemingly irrelevant to causal processes, which is consistent with the argument regarding the omnipresence of causal understanding in humans (Pearl & Mackenzie, 2018). When we see a picture in daily life, we often process tasks such as identifying objects in the scene or memorizing the whole scene. These processes, though appearing to be quite different from that of how the image was generated, often involve contributions from the generative process. Recent studies have also demonstrated that generative models have succeeded in explaining various research topics on human vision that were often conceived as not directly concerning the generative process, such as multiple object tracking (Vul et al., 2009), visual working memory (Brady & Tenenbaum, 2013) and ensemble perception (Whitney & Yamanashi Leib, 2018). These achievements highlight the gravity of causality in understanding human vision by showing that causal history underlies a wide range of seemingly noncausal tasks.

P(image) for Psychophysical Experiments

Our research largely relies on the concept of *P*(image)—the model's estimation on the probability of generating an image, which serves as a unified independent variable that was manipulated to explain human performance on multiple tasks. Our focus on *P*(image) is enlightened by theories in computer vision (S.-C. Zhu et al., 1997): Given a model, each image should be understood in the context of a space where all possible images reside. Within an image space, every image is considered as a single point rather than a collection of pixels, and all images collectively form a population of points. For example, an image with 100 grayscale pixels is equivalent to a point in an image space of 100 dimensions. From this perspective, to synthesize an entire image is to sample a single point from this high dimensional space. Within this space, meaningful images in the eyes of humans constitute only a tiny portion of all points, and thus a uniform sampling will most likely result in an image of meaningless noise. In this sense, a good model of human vision should be able to map out how humans distribute the probability mass across their image spaces, with a higher probability assigned to a more meaningful image. By sampling from this distribution, the model can better synthesize images that people have in mind.

More than a theoretical discussion, this idea also brings up significant psychophysical implications. Due to a conservation of probability mass, which must sum up to 1, assigning some images with higher probabilities means lower probabilities for other images. Thus, different *P*(image)s reflect the priority of human vision—images with high *P*(image)s are prioritized with more prompt and robust processing. This principle is nicely summarized as the No Free Lunch Theorem in machine learning, which suggests that for any model, good performances on certain sets of data are necessarily at the cost of those degraded performance on other sets of data (Wolpert & Macready, 1997).

Consistent with the prediction from the No Free Lunch Theorem, from Experiments 2 and 3, we found that images with higher *P*(image)s and thus low information content were perceived as less complex and exhibited greater stability during transmission, potentially suggesting that the visual system is better adapted in processing these images. Nevertheless, our evidence on vision's priority in processing images with high *P*(image)s is still limited, and future research is encouraged to further explore if this phenomenon is universal across different aspects of vision, such as whether images with high *P*(image)s predict faster visual search (Võ et al., 2019; Wolfe & Horowitz, 2017), more robust consolidation in working memory (Vogel et al., 2006; Xu & Chun, 2006), and more precise memory retrieval (Hyun et al., 2009; Konkle et al., 2010).

Utilizing Information Theory in Combination With a Generative Model

Information theory is the mathematical treatment of the concepts, parameters, and rules governing the transmission of messages through communication systems (Shannon, 1948). Through this case study on Mondrian-style images where information theory was heavily used in our dependent measurement of visual performance, we can see an interesting trend of how information theory plays a more important role in understanding human vision.

Directly contributed to the cognitive revolution in the 50 s, information theory has been deeply embedded into the foundation of cognitive psychology by inspiring a widely recognized conceptualization about the human mind that the mind works as an information processing system. Today, various terminologies in psychology are inspired by information theory. For example, terms such as "encoding," "decoding," "storage capacity," and "retrieval" are commonly used in cognitive psychology, and all have their counterparts in the theory.

However, it has not been a smooth journey for psychologists to use the information theory to quantify human mental processes.

After some early breakthroughs in measuring the capacity of the human mind in terms of bits (e.g., Hake & Garner, 1951; Pollack, 1952), this approach encountered a great theoretical challenge in an influential study on short-term memory capacity showing that the memory capacity of letters could not be described by a fixed number of bits but instead can be better described by "chunking" (Miller, 1956). Decades later, reflections were made on the struggle in applying information theory in empirical research (Laming, 2001; Luce, 2003). The biggest difficulty was speculated to occur in measuring the probability of stimuli—the objective probability of a stimulus being physically presented in an experiment does not necessarily reflect its subjective probability in the human mind. This challenge has been pointed out by Luce (2003, p. 185), "The probabilities are on a pure, unstructured set whose elements are functionally interchangeable...however, the stimuli of psychological experiments are to some degree structured, and so, in a fundamental way, they are not in any sense interchangeable." Without an accurate subjective probability serving as input, outcomes from information theory simply were simply incapable of capturing the human mind.

Retrospectively, we can see how this challenge failed to be resolved in the early stages of cognitive psychology for two reasons. First, it was simply hard to build a sophisticated probabilistic model of psychological structures until decades later, when probability inference as a principle of intelligence became more widely accepted (Pearl, 1988). In the current study, our HP model is a variant of the nCRP model introduced in early 2010 (Blei et al., 2010). Other successful applications of information theory also largely depended on Bayesian probabilistic models, such as explaining eye movement during visual search (Najemnik & Geisler, 2005). From this perspective, the lack of sophisticated models capable of capturing human's structured subjective probability is being addressed progressively.

Second, a difficulty was present in eliciting the subjective probabilities from humans. Traditional methods such as verbal reports are limited in measuring this probability as these responses are subjected to an ambiguity between a spontaneous response or a calculated cognitive inference of the task (Firestone & Scholl, 2016; Gao et al., 2009). Recently, eliciting representations in human minds have undergone significant development by allowing humans to freely generate stimuli, such as tapping the shape anywhere they like on a touch-sensitive tablet (Firestone & Scholl, 2014) or directly drawing objects (Fan et al., 2018), beyond measuring humans' accuracy and response time from given stimuli. In the current study, the image-generation task in Experiment 1 allowed humans to partition Mondrian-style images freely during which the human priors were elicited. The telephone game in Experiment 3, despite not allowing for free creation, still allowed for a free regeneration process.

In our case study integrating a structured probability model with elicited subjective priors, we noticed two interesting observations regarding the application of information theory to human vision. First, the structured probabilistic models assign a great variation of probabilities to different images, which can be taken advantage of in experimental design. Unlike early works (e.g., Hake & Garner, 1951; Pollack, 1952) focusing on uniform distribution by randomizing stimuli as much as possible, our study, powered by a probabilistic model, embraced varied probabilities and turned probability itself into an independent variable. By deliberately selecting samples varying dramatically in their subjective probabilities from the synthesizing process, we are able to see that, at a large scale, how human performance can be significantly influenced by the information content of images, therefore showing how information theory can explain human vision in broad strokes.

Another interesting observation is the change in the focus of information theory in understanding vision. In our study, the focus shifted from measuring the exact human channel capacity in bits to using information theory as a window to probe the structured representations in the human mind. It is only when a model accurately captures the structures of human mental representations can the information content derived from it strongly correlate with human performance.

Conclusion

In the current study, we demonstrate the generative nature of human vision using Mondrian-style images as a case study. Our study highlights the successes in applying image-generation processes to explain humans' visual interpretation, providing direct support for the analysis-by-synthesis perspective and suggesting that humans interpret an image by recovering how it was generated. At the same time, with a "big tasks" paradigm, we show that our generative model can be well generalized across a wide range of visual tasks, supporting that generative vision constitutes a type of common sense. Lastly, our study demonstrates that a structured probabilistic model combined with elicited human priors is an effective approach in addressing the long-lasting challenges in applying information theory, showing that information theory can be a powerful tool in quantifying the human mind in psychological research.

Context of the Research

We carried out this study to emphasize the perspective that vision offers more than "what" and "where" but also "why" and "how." We hypothesize that vision reflects a deep, underlying causal structure that serves as a type of common sense for humans to intelligently interpret the images, construct the hidden causal world, and support a wide range of tasks in real life. This study uses Mondrian-style images as a case study to confirm such hypothesis. Mondrian-styles images are good candidates as they possess interesting graphical structures while allowing for tractable modeling. Our study explores an integration of psychophysics and computational modeling that goes beyond traditional measurements of reaction time and accuracy. In our case, model is not just for fitting experimental data but also helps derive the important hypothesis of our experiment. We hope this study can encourage more deep integration of psychophysics and causal modeling in the future.

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 339–351). MIT Press.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Bennett, B., Hoffman, D. D., & Prakash, C. (1989). *Observer mechanics: A formal theory of perception*. Academic Press. https://doi.org/10.1016/C2013-0-10358-3

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer-Verlag. https://doi.org/10.1007/978-0-387-45528-0

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies. *Journal of the ACM*, *57*(2), 1–30. https://doi.org/10.1145/1667053.1667056

Bradski, G. (2000). The openCV library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, *25*(11), 120–123.

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*(1), 85–109. https://doi.org/10.1037/a0030779

Brainard, D. H., & Freeman, W. T. (1997). Bayesian color constancy. *Journal of the Optical Society of America A*, *14*(7), 1393–1411. https://doi.org/10.1364/JOSAA.14.001393

Bruner, J. S. (1973). *Beyond the information given: Studies in the psychology of knowing*. WW Norton.

Chen, Y.-C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*(6), 923–930. https://doi.org/10.1177/0956797616628525

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. https://doi.org/10.21236/AD0616323

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*, *42*(8), 2670–2698. https://doi.org/10.1111/cogs.12676

Feldman, J., & Singh, M. (2006). Bayesian estimation of the shape skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019. https://doi.org/10.1073/pnas.0608811103

Firestone, C., & Scholl, B. J. (2014). "Please tap the shape, anywhere you like" shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological Science*, *25*(2), 377–386. https://doi.org/10.1177/0956797613507584

Firestone, C., & Scholl, B. J. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences*, *39*. Article e229. https://doi.org/10.1017/S0140525X15000965

Freeman, W. T. (1996). Exploiting the generic viewpoint assumption. *International Journal of Computer Vision*, *20*(3), 243–261. https://doi.org/10.1007/BF00208721

Freyd, J. J. (1987). Dynamic mental representations. *Psychological Review*, *94*(4), 427–438. https://doi.org/10.1037/0033-295X.94.4.427

Froyen, V., Feldman, J., & Singh, M. (2015). Bayesian hierarchical grouping: Perceptual grouping as mixture estimation. *Psychological Review*, *122*(4), 575–597. https://doi.org/10.1037/a0039540

Gao, T., Newman, G. E., & Scholl, B. J. (2009). The psychophysics of chasing: A case study in the perception of animacy. *Cognitive Psychology*, *59*(2), 154–179. https://doi.org/10.1016/j.cogpsych.2009.03.001

Geisler, W. S., & Kersten, D. (2002). Illusions, perception and Bayes. *Nature Neuroscience*, *5*(6), 508–510. https://doi.org/10.1038/nn0602-508

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, B. D. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall. https://doi.org/10.1201/b16018

Gelman, A., Lee, D., & Guo, J. (2015). Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, *40*(5), 530–543. https://doi.org/10.3102/1076998615606113

Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., & Tenenbaum, J. B. (2008). Church: a language for generative models. In D. McAllester, & P. Myllymaki (Eds.), *Conference on Uncertainty in Artificial Intelligence* (pp. 220–229). AUAI Press.

Gregory, R. L. (1970). *The intelligent eye*. McGraw-Hill.

Grenander, U. (1976). *Pattern synthesis: Lectures in pattern theory: Volume I*. Springer-Verlag.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, *31*(3), 441–480. https://doi.org/10.1080/15326900701326576

Hagberg, A., Swart, P., & Schult, D. (2008). *Exploring network structure, dynamics, and function using NetworkX* (Report No. LA-UR-08-05495; LA-UR-08-5495). Los Alamos National Lab (LANL).

Hake, H. W., & Garner, W. R. (1951). The effect of presenting various numbers of discrete steps on scale reading accuracy. *Journal of Experimental Psychology*, *42*(5), 358–366. https://doi.org/10.1037/h0055485

Hyun, J., Woodman, G. F., Vogel, E. K., Hollingworth, A., & Luck, S. J. (2009). The comparison of visual working memory representations with perceptual inputs. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(4), 1140–1160. https://doi.org/10.1037/a0015019

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294. https://doi.org/10.3758/BF03194066

Kanizsa, G. (1976). Subjective contours. *Scientific American*, *234*(4), 48–52. https://doi.org/10.1038/scientificamerican0476-48

Kazak, A. E. (2018). Editorial: Journal article reporting standards. *American Psychologist*, *73*(1), 1–2. https://doi.org/10.1037/amp0000263

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*(1), 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Knill, D. C., & Richards, W. (1996). *Perception as Bayesian inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511984037

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, *139*(3), 558–578. https://doi.org/10.1037/a0019165

Kulkarni, T. D., Kohli, P., Tenenbaum, J. B., & Mansinghka, V. (2015). Picture: A probabilistic programming language for scene perception. IEEE Conference on Computer Vision and Pattern Recognition (pp. 4390–4399). https://doi.org/10.1109/CVPR.2015.7299068

Lake, B. M., & Piantadosi, S. T. (2020). People infer recursive visual concepts from just a few examples. *Computational Brain & Behavior*, *3*(1), 54–65. https://doi.org/10.1007/s42113-019-00053-y

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, Article e253. https://doi.org/10.1017/S0140525X16001837

Laming, D. (2001). Statistical information, uncertainty, and Bayes' theorem: Some applications in experimental psychology. European Conference on Symbolic and Quantitative Approaches to Reasoning and Uncertainty (Vol. 2143, pp. 635–646). https://doi.org/10.1007/3-540-44652-4_56

Leyton, M. (1989). Inferring causal history froms shape. *Cognitive Science*, *13*(3), 357–387. https://doi.org/10.1207/s15516709cog1303_2

Leyton, M. (1992). *Symmetry, causality, mind*. MIT Press.

Long, B., Fan, J., Chai, Z., & Frank, M. C. (2021). *Parallel developmental changes in children's drawing and recognition of visual concepts*. *PsyArXiv*. https://psyarxiv.com/5yv7x/

Luce, R. D. (2003). Whatever happened to information theory in psychology? *Review of General Psychology*, *7*(2), 183–188. https://doi.org/10.1037/1089-2680.7.2.183

Mamassian, P., Landy, M., & Maloney, L. T. (2002). Bayesian modelling of visual perception. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 13–36). MIT Press. https://doi.org/10.7551/mitpress/5583.003.0005

Mansinghka, V. K., Kulkarni, T. D., Perov, Y. N., & Tenenbaum, J. B. (2013). Approximate Bayesian image interpretation using generative probabilistic graphics programs. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K.Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 1520–1528). Curran Associates.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. W. H. Freeman.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*(7031), 387–391. https://doi.org/10.1038/nature03390

Nakayama, K., & Shimojo, S. (1992). Experiencing and perceiving visual surfaces. *Science*, *257*(5075), 1357–1363. https://doi.org/10.1126/science.1529336

Palmer, S. E. (1999). *Vision science: Photons to phenomenology*. MIT Press.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier. https://doi.org/10.1016/c2009-0-27609-4

Pearl, J. (2000). *Causality: Models, reasoning, and inference*. Cambridge University Press. https://doi.org/10.1017/cbo9780511803161

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic books.

Pitman, J. (2006). *Combinatorial stochastic processes: Ecole d'eté de probabilités de saint-flour xxxii-2002* (J. Picard, Ed.). Springer-Verlag.

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, *24*(6), 745–749. https://doi.org/10.1121/1.1906969

Richards, W., Jepson, A., & Feldman, J. (1996). Priors, preferences and categorical percepts. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian inference* (pp. 93–122). Cambridge University Press. https://doi.org/10.1017/cbo9780511984037.005

Rock, I. (1983). *The logic of perception*. MIT Press.

Roy, D. M., & Teh, Y. W. (2008). The Mondrian process. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in neural information processing systems* (pp. 1377–1384). Curran Associates.

Schmidt, F., Phillips, F., & Fleming, R. W. (2019). Visual perception of shape-transforming processes: 'Shape scission'. *Cognition*, *189*, 167–180. https://doi.org/10.1016/j.cognition.2019.04.006

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Spröte, P., Schmidt, F., & Fleming, R. W. (2016). Visual perception of shape altered by inferred causal history. *Scientific Reports*, *6*(1), 1–11. https://doi.org/10.1038/srep36245

Suchow, J. W., Fougnie, D., Brady, T. F., & Alvarez, G. A. (2014). Terms of the debate on the format and structure of visual memory. *Attention, Perception, & Psychophysics*, *76*(7), 2071–2079. https://doi.org/10.3758/s13414-014-0690-7

Tang, N., Gong, S., Zhou, J., Shen, M., & Gao, T. (2021). *Generative vision as common sense: Testing analysis-by-synthesis on Mondrian-style image* [Data and code]. https://osf.io/u62xb/

Thurstone, L. L. (1927). Psychophysical analysis. *The American Journal of Psychology*, *38*(3), 368–389. https://doi.org/10.2307/1415006

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/lix.236.433

Uddenberg, S., & Scholl, B. J. (2018). Teleface: Serial reproduction of faces reveals a whiteward bias in race memory. *Journal of Experimental Psychology: General*, *147*(10), 1466–1487. https://doi.org/10.1037/xge0000446

van Rossum, G.. (1995). *Python reference manual* (F. L. Drake, Jr. Ed.). Centre for Mathematics and Computer Science.

Vetter, T., & Troje, N. F. (1997). Separation of texture and shape in images of faces for image coding and synthesis. *Journal of the Optical Society of America A*, *14*(9), 2152–2161. https://doi.org/10.1364/josaa.14.002152

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436–1451. https://doi.org/10.1037/0096-1523.32.6.1436

von Helmholtz, H. (1925). *The perceptions of vision: Treatise on physiological optics* (J. P. C. Southall Ed.). Optical Society of America. (Original work published 1920).

Vul, E., Frank, M. C., Tenenbaum, J. B., & Alvarez, G. A. (2009). Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 1955–1963). Curran Associates.

Whitney, D., & Yamanashi Leib, A. (2018). Ensemble perception. *Annual Review of Psychology*, *69*(1), 105–130. https://doi.org/10.1146/annurev-psych-010416-044232

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour*, *1*(3), 1–8. https://doi.org/10.1038/s41562-017-0058

Wolpert, D. H., & Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, *1*(1), 67–82. https://doi.org/10.1109/4235.585893

Xu, Y., & Chun, M. M. (2006). Dissociable neural mechanisms supporting visual short-term memory for objects. *Nature*, *440*(7080), 91–95. https://doi.org/10.1038/nature04262

Yildirim, I., Kulkarni, T. D., Freiwald, W. A., & Tenenbaum, J. B. (2015). Efficient and robust analysis-by-synthesis in vision: A computational framework, behavioral tests, and modeling. In R. Dale, C. Jennings, P. P. Maglio, T. Matlock, D. C. Noelle, A. Warlaumont, & J. Yoshimi (Eds.), *Annual conference of the cognitive science society* (pp. 2751–2756). Cognitive Science Society.

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Zhu, S.-C., & Mumford, D. (2007). *A stochastic grammar of images*. Now Publishers. https://doi.org/10.1561/9781601980618

Zhu, S.-C., Wu, Y. N., & Mumford, D. (1997). Minimax entropy principle and its application to texture modeling. *Neural Computation*, *9*(8), 1627–1660. https://doi.org/10.1162/neco.1997.9.8.1627

Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., Tenenbaum, J. B., & Zhu, S.-C. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, *6*(3), 310–345. https://doi.org/10.1016/j.eng.2020.01.011

(*Appendices follow*)

Appendix A

Models

Human Prior Hierarchical Bayesian Model (HP)

Probability of Tree Controlled by Evenness Factor γ—P_γ(tree)

In a Mondrian-style image, there are six primitive rectangles that would not be further divided, whose configuration reflects the topology of the image's parse tree. Conceptually, the process of applying partitions to an image is equivalent to building a tree by assigning six items to all its nodes. This item assignment is a recursive process: For each node, the items in it will be assigned to its child nodes, and the same process continues until every child node contains only one item. This concept can be implemented by a stochastic process known as the nCRP (Blei et al., 2010). As the name suggests, "nest" shows that it is recursive, whereas the "Chinese Restaurant Process" is the one describing how items are assigned to nodes at one iteration (CRP, Pitman, 2006).

The CRP originates from its description of how customers (items) in a Chinese restaurant (parent node) are assigned to tables (child nodes). It assigns N items one at a time, from item 1 to item N (Figure A1). Each item is either assigned to an existing child node or a newly created child node. The probabilities of two possible assignments follow the equation below (Equation 10).

$$P(C_n = j | C_{1:n-1}) = \begin{cases} \dfrac{M_j}{n-1+\gamma}, & j \le J, \text{ to an existing child node} \\ \dfrac{\gamma}{n-1+\gamma}, & j = J+1, \text{ to a new child node} \end{cases} \tag{10}$$

Here, n is the current item being assigned, and C_n represents the child node to which the item is assigned, whereas $C_{1:n-1}$ represents the child nodes to which all previous $n-1$ items are assigned; j is the index of a child node, and J represents the total number of existing child nodes; M_j represents the total number of items already sitting at node j; γ is the branching factor controlling the process' tendency of creating a new child node: A larger γ indicates a higher possibility that an item will be assigned to a newly created child node.

After all the items are assigned, an assignment plan for this unique composition of a tree is formed that can be computed into a probability (Equation 11). Different assignment plans will have distinct probabilities.

$$P(B = \{b_1, b_2 \ldots b_K\}) = \frac{\Gamma(\gamma)\gamma^{\sum_{k=1}^{K} |b_k|}}{\Gamma(\gamma + N)} \prod_{k=1}^{K} \Gamma(|b_k|) \tag{11}$$

Here, B is an assignment plan for N items; K is the total number of child nodes; b_k is a set of items in the order they are assigned to the node k; and Γ is the gamma function. For example, for Figure A1, $B = \{b_1, b_2\}$, $b_1 = \{``1", ``3"\}$, $b_2 = \{``2"\}$).

The probability of a tree can then be computed by multiplying the probabilities of all the assignment plans that can possibly constitute the tree (Equation 12).

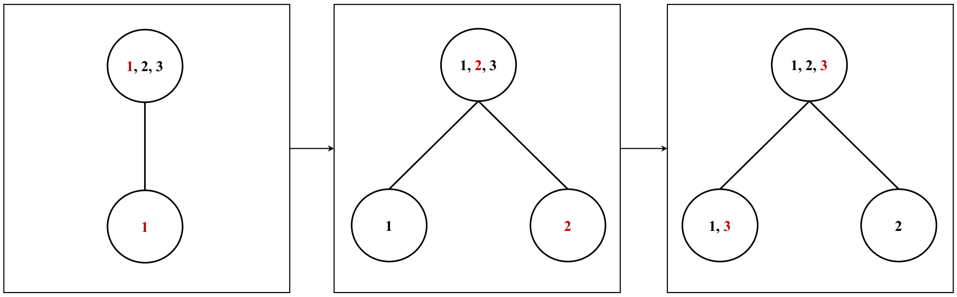
$$P(\text{tree}) = \prod_{t=1}^{T} P(B_t) \tag{12}$$

Here B_t is an assignment plan. T is the total number of item assignments in the tree.

While the index of items in CRP is for identifying distinct customers, the abstract primitive rectangles of Mondrian-style images in our study are indistinguishable from each other. Thus, instead of concerning the particular order in which distinct entities are assigned, we are interested in the general topology of the tree that depends on the total amounts of items (primitive rectangles) assigned to each child node. That is, we only discriminated assignment plans that show differences in their quantity distributions (Figure A2). For example, a three-item assignment can only give rise to two distinct quantity distributions (2-1 and 1-1-1), despite that the 2-1 distribution can be achieved by three unique assignment plans.

Figure A1

Illustration of Assigning Three Items Into Two Child Nodes

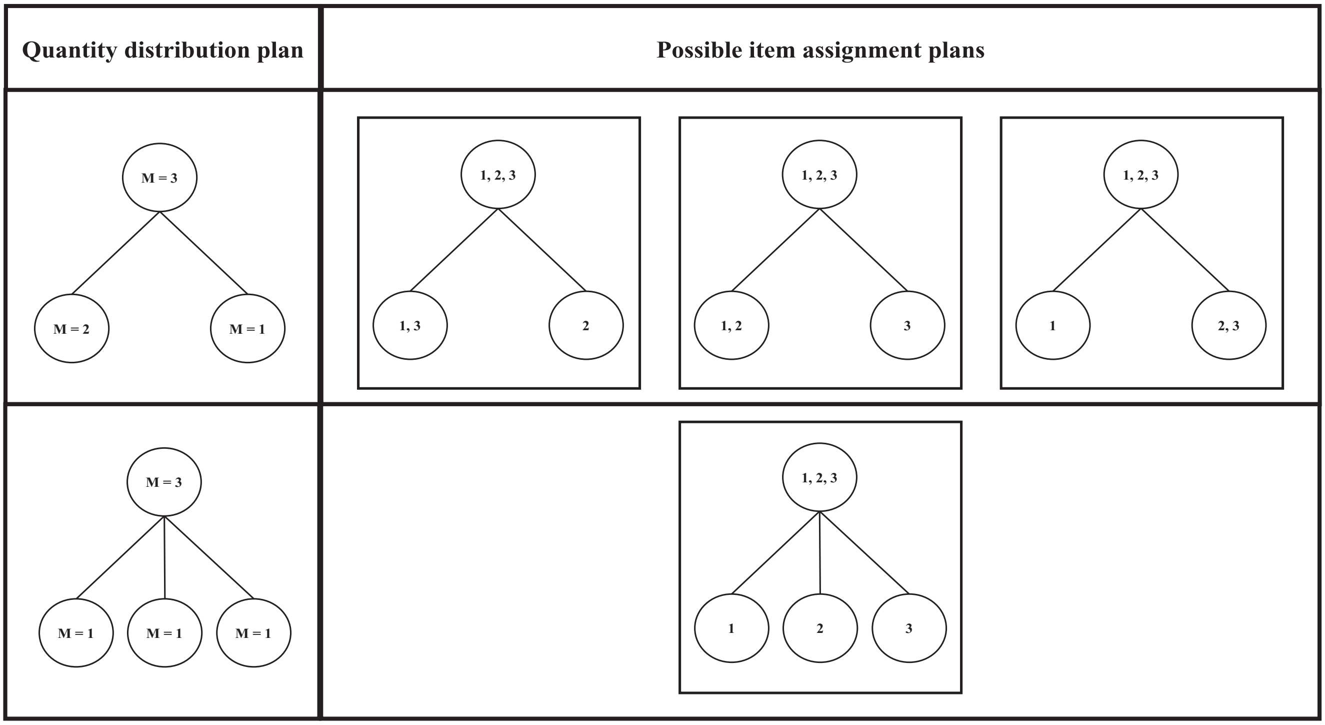


Note. See the online article for the color version of this figure.

(*Appendices continue*)

Figure A2

Illustration of the Possible Three-Item Quantity Distributions and Their Corresponding Assignment Plans



Correspondingly, we can obtain the probability of each quantity distribution by summing the probabilities of all assignment plans that can possibly lead to this distribution (Equation 13).

$$P(A = \{|b_1|, |b_2| \ldots |b_K|\}) = \prod_{k=1}^{K} C^{|b_k|-1}_{N-\sum_{1}^{k}|b_{k-1}|-1} \quad P(B = \{b_1, b_2 \ldots b_K\}) \tag{13}$$

Here, A is the quantity distribution of N items; K is the total number of child nodes; $|b_k|$ is the total number of items assigned in the child node k; C is the combination calculator.

Further, since a Mondrian-style image is sensitive to the spatial organization of its parts and subparts, it is important to reflect the spatial relationship in the topology of a tree—the placement of a child node matters within the layer it resides. Specifically, here we use nodes on the left to represent regions on the left/top of a Mondrian-style image, whereas nodes on the right to represent regions on the right/bottom. Through this convention, we can distinguish parsing proposals that constitute different spatial organizations with the same quantity distribution of items (Figure A3).

The probability of a quantity distribution should be equally shared by its distinct parsing proposals (Equation 14).

$$P(P_r^A) = \frac{P(A)}{R} \tag{14}$$

Here, P_r^A is a parsing proposal with a quantity distribution A; R is the total number of the parsing proposals with the same quantity distribution.

Moreover, the nCRP allows for one-to-one assignment, where items in a parent node can all be assigned to a single child node. However, this assignment plan is not allowed in our parse tree representation since simply transferring items from a parent node to a child node in a different layer will not result in any change in a Mondrian-style image (Figure A4). Thus, the probabilities of other valid parsing proposals are renormalized by excluding one-to-one assignment.

The probability of a tree can then be computed by multiplying the probabilities of all parses after the three modifications are made in the tree (Equation 15), where T is the total number of parsing in the tree.

$$P(\text{tree}) = \prod_{t=1}^{T} P(P_t) \tag{15}$$

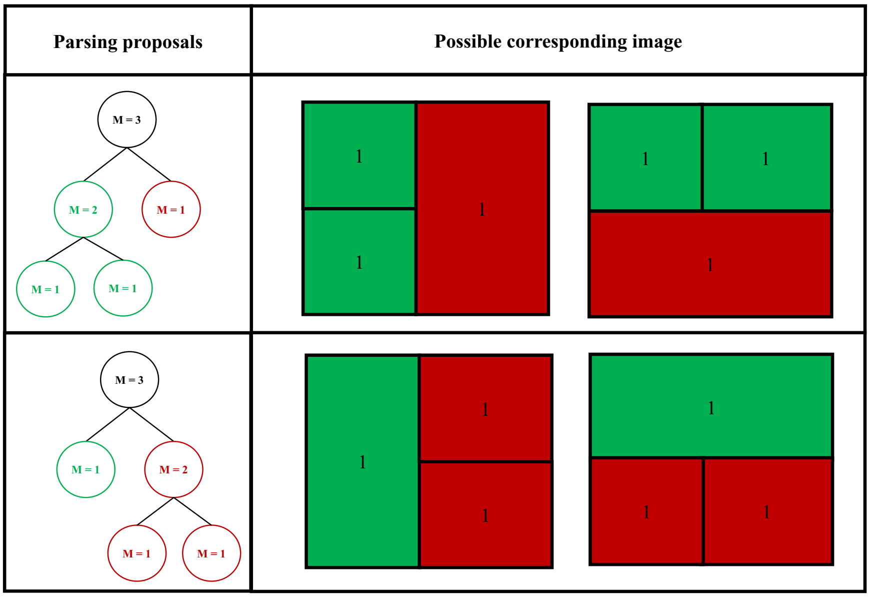
Probability of Partition Given Tree Controlled by Evenness Factor α—P_α(partition|tree)

With a tree structure determining the number of subparts into which each part should be parsed, considerations need to be made on the orientation of cuts and percentage of subparts.

(*Appendices continue*)

Figure A3

Illustration of the Possible Parsing Proposals for a Three-Item Quantity Distribution and the Corresponding Visualization on Images

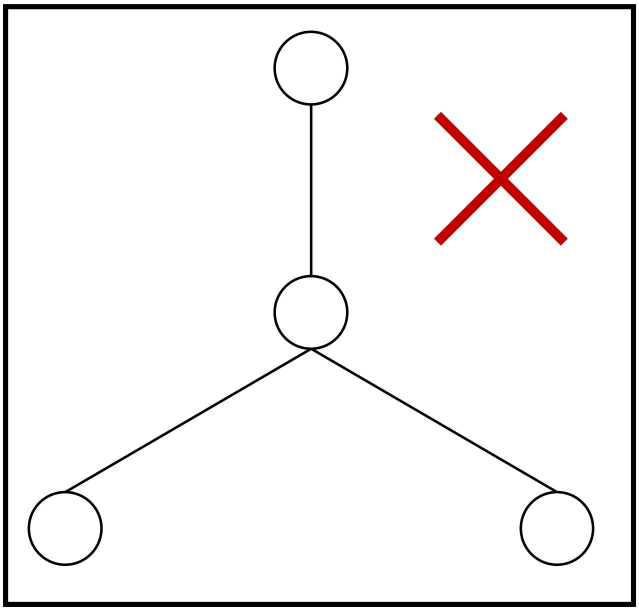


Note. See the online article for the color version of this figure.

For each nonterminal node in the tree, the model would partition its corresponding part of the image into subparts represented by its child nodes horizontally or vertically. Here, we assigned an equal probability of 50% to each orientation. Suppose a node in a tree has K child nodes. Then, there should be K variables indicating the percentages of K subparts occupying a part in an image. This ratio can be modeled by the systematic Dirichlet distribution with a parameter α, the concentration parameter that is referred to as the evenness factor in our study to better match the subjective consequence. The probability of observing a set of occupied percentages (z_1, z_2…z_K) is shown in Equation 16, where Q_k is the occupied percentages of the k subparts and Γ is the gamma function.

Figure A4

An Illustration of an Illegal Tree in Our Model



Note. In this illegal tree, all items in a parent node are assigned to a single child node. See the online article for the color version of this figure.

$$P(Q_k = \{z_1, z_2 \ldots z_K\}) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{i=1}^{k-1} z_i^{\alpha-1} \tag{16}$$

Here we define the partition as a particular set of cuts applied to an image. The probability of observing a partition given a tree, computed by multiplying the probabilities of cut orientations and the probabilities of percentages in all parsing proposals along the tree, is shown in Equation 17, where T is the total amount of parses in the tree.

$$P(\text{partition}|\text{tree}) = \frac{1}{2^T} \prod_{t=1}^{T} P(Q_t) \tag{17}$$

Noninformative Prior Hierarchical Bayesian Model (NP)

Noninformative Prior of Branching Factor

The NP model shares the same hierarchical Bayesian structure as the HP model, but with a noninformative uniform prior instead of

(*Appendices continue*)

human prior. According to the definition of CRP (Pitman, 2006), the branching factor γ has the constraint of $\gamma > 0$. Although a uniform prior on (0, +∞) would be an intuitive way to construct a noninformative prior, it is in fact "informative" for the branching factor. This is because $\gamma > 1$ implies that the probability of assigning an item to a new child node is bigger than that of an existing child node, and thus a uniform distribution of γ on (0, +∞) would result in a model preferring wide and shallow trees with more branches, instead of showing no preference in every decision on branching. To resolve this, we constructed a "noninformative" prior with equal preferences for all types of trees by using the reparameterization trick (Gelman et al., 2013, Chapter 5), in which the branching factor γ is reparametrized in terms of θ, where $\theta = \log_{10}(\gamma)$. When computing the P(tree) for the NP model, we then uniformly sampled θ and applied $\gamma = 10^{\theta}$ to obtain P_{γ}(tree).

Noninformative Prior of Evenness Factor

For a similar reason, we also reparametrized the evenness factor α in terms of β, where $\beta = \log_{10}(\alpha)$. We sampled β and applied $\alpha = 10^{\beta}$ to obtain P_{α}(partition|tree).

Appendix B

Experiments

Experiment 1

Samples of Synthesized Images by Human

Figure B1.

Data Merge

We analyzed the systematic difference between the training images from both paper and computer environments by comparing the logarithms of aspect ratios and sizes of the primitive rectangles in these two types of images. The feature values of all primitive rectangles are shown in Figure B2. Most points representing the values are concentrated in one area of the scatter. The statistical analysis (Figure B3) showed that no significant difference was found between the two environments on logarithms of aspect ratio, Paper: $M = -0.06$, 95% CI [−0.20, 0.08]; PC: $M = -0.17$, [−0.30, −0.03]; $t(598) = -1.10$, $p = .27$, $d = 0.09$; $\mathrm{BF}_{01} = 6.11$, or size, Paper: $M = 0.167$, [0.16, 0.17]; PC: $M = 0.167$, [0.16, 0.17]; $t(598) = -0.003$, $p = .998$, $d < 0.001$; $\mathrm{BF}_{01} = 10.99$.

Experiments 2–4

Stimuli

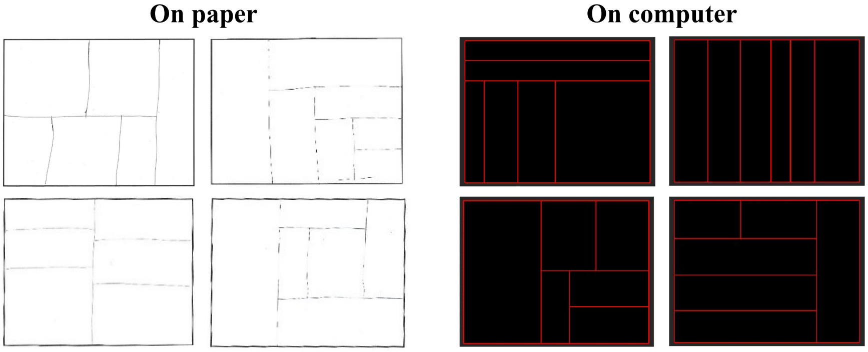
In each experiment, 10,000 new images were sampled from the HP model. These images were then sorted by their information contents computed by the HP model. One image is selected from every 500 images, resulting in 20 images in total. In Experiment 4, we selected 16 images from 20 images since the posterior distributions of the parse trees were concentrated on one hypothesis for the other four images.

Model Comparison With the Regularized "Feature Multivariate Gaussian Model"

We regularized the multivariate Gaussian model (LMG) by adding a weighted diagonal matrix to the covariance matrix (Gelman et al., 2013, Chapter 14). The weight λ was selected by a 10-folder cross-validation method. We further compared this model with the hierarchical models (HP and NP) in Experiments 1–3.

Figure B1

Samples of Images Generated on Paper and on the Computer Screen

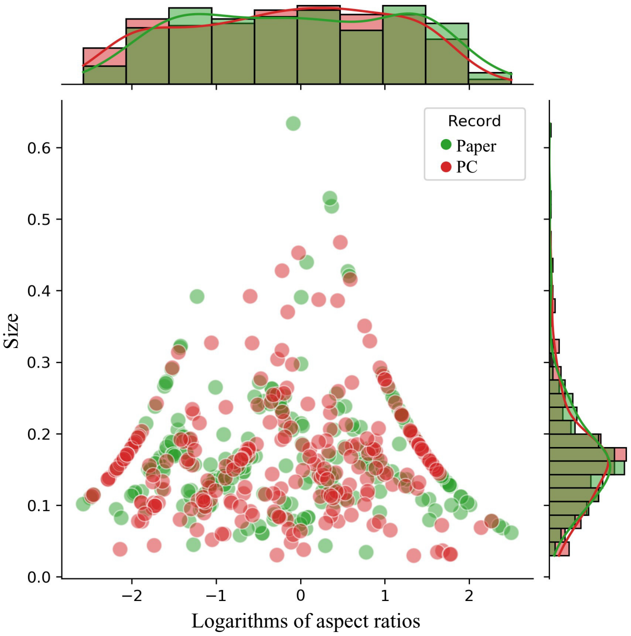


Note. See the online article for the color version of this figure.

(*Appendices continue*)

Figure B2

Scatterplot and Histogram of Data Points Representing the Logarithms of Aspect Ratios and Sizes of the Primitive Rectangles in Images



Note. Different colors represent the environments in which the images were drawn. See the online article for the color version of this figure.

The most important finding is that the experimental results from the regularized model did not change our conclusion in any qualitative way.

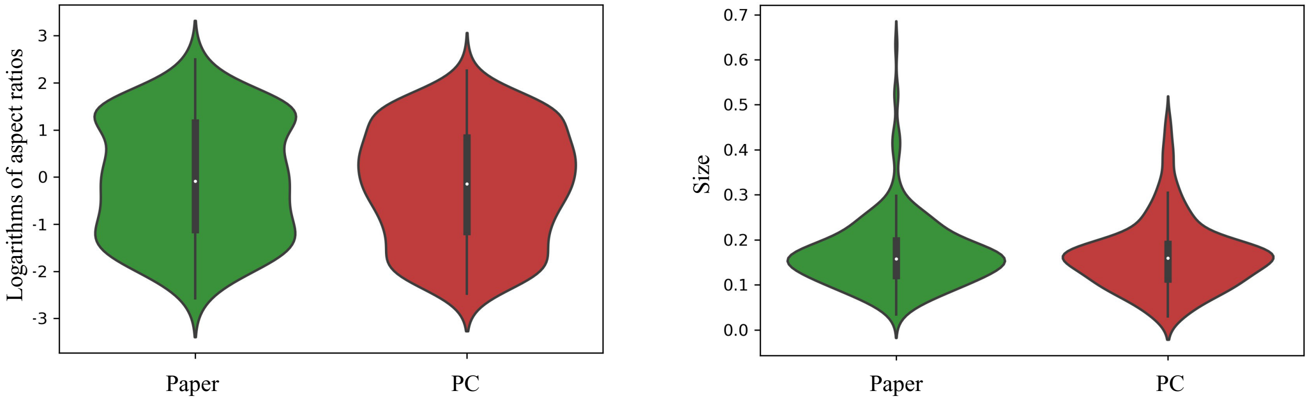
In Experiment 1, the image generation task, we first found a similar training effect as the nonregularized MG model: After being trained by human-synthesized images, the regularized MG model could explain new human-synthesized images significantly better than nonhuman-synthesized images, human-synthesized: $M = -6.57$, 95% CI [−6.61, −6.52]; nonhuman-synthesized: $M = -14.07$, [−14.16, −13.99]; $t(58) = 157.97$, $p < .001$, $d = 40.79$. Second, we also further examined which of the trained models, the HP model or the regularized MG model, could better capture human preference. The results showed that the HP model significantly outperformed the regularized MG model in explaining new human images, $t(58) = 173.22$, $p < .001$, $d = 44.73$.

In Experiment 2, the complexity ranking task, like the original nonregularized MG model, there was still no significant correlation between the ranking *Z*-scores of humans and those of the regularized MG model ($r = -.18$, 95% CI [−0.57, 0.29], $p = .48$; $BF_{01} = 2.79$). Bayesian Analysis also showed that there is strong evidence in favor of the NP model being better than the regularized MG model in explaining the human-ranking *Z*-scores ($BF_{NP,\ LMG} = 21.43$). This additional evidence still supports that the causality in the Bayesian hierarchy model is essential in capturing human perceived complexity compared to the regularized MG model.

In Experiment 3, the human iteration task, the correlations between initial information content and information loss were still not significant in all three chains for the regularized MG model (Chain A: $r = .30$, 95% CI [−0.16, 0.66], $p = .19$, $BF_{01} = 1.64$; Chain B: $r = .24$, [−0.23, 0.62], $p = .31$, $BF_{01} = 2.24$; Chain C: $r = .38$, [−0.08, 0.70], $p = .10$, $BF_{01} = 1.04$). The results showed that the regularized MG model could not estimate human subjective *P*(image) accurately, the same as the nonregularized MG model, whereas both the hierarchical Bayesian models can. This suggests that the causal generative process in Bayesian models plays a critical role in explaining how images mutate through an iterative image-passing procedure.

Figure B3

Violin Plots of the Logarithms of Aspect Ratios and Sizes of the Primitive Rectangles in Images



Note. Different colors represent the environments in which the images were drawn. See the online article for the color version of this figure.

Received March 23, 2022
Revision received August 30, 2022
Accepted February 21, 2023 ■