Aesthetic Preferences for Residential Architecture: Finding Ground Truth with Machine Learning Approaches*

Thies Lindenthal

htl24@cam.ac.uk

Carolin E. Schmidt

cs2100@cam.ac.uk

Wayne X. Wan

xw357@cam.ac.uk

April 2, 2021

Abstract

ML-enabled classifiers are regularly criticized for being 'black boxes': While their predictive power is undisputed, it is difficult to understand why the model arrived at a particular classification. The same can be said for humans classifying photos according to their aesthetic appeal. They can quickly say whether they like a photo or not – but giving justifications for such a choice is often challenging. Also, human classifiers exhibit inconsistencies and biases, adding to the black box nature of their classifications.

This paper first collects binary classifications of house pictures from a large group of participants and then trains personalized ML classifiers for each participant. Predictions from these automated yet personal classification machines shed light on biases and inconsistencies in the participants' assessment of residential real estate's visual appeal.

Keywords: machine learning, computer vision, residential real estate, aesthetic preference

*This publication is the result of a project sponsored within the scope of the SEEK research programme which was carried out in cooperation between the University of Cambridge and ZEW – Leibniz-Zentrum für Europäische Wirtschaftsforschung GmbH Mannheim, Germany. We thank Peter Buchmann for setting up the infrastructure and providing excellent technical support.

"It is harder to be unhappy when you are eating Craig's ice cream." Kurt Vonnegut (1973)

1 Introduction

In March 2021, Cambridge City Council's planning committee rejected a project that would have created modern living quarters for 113 students[1]. During the planning process, 307 objecting comments had been filed, many of which offered outspoken feedback on the suggested design, including "quite possibly[...]the ugliest building in Cambridge", "another hideous[...]structure", "eyesore", or "looking like a prison" (Greater Cambridge Shared Planning, 2021). The building's design was not the only reason why so many residents and planners objected, but aesthetic concerns offered one more opportunity to block the project.

In situations like this, planning committees need to assess whether the design is indeed as outrageous as suggested by the comments and whether the objectors' statements are truly reflecting the residents' aesthetic preferences. The perspective of having 113 students as neighbors could have darkened the view on the architectural merits. A "That's ugly!" might simply mean "Not in my backyard!".

In this paper, we develop ML-enabled classifiers to predict personalized responses to residential architecture. By engineering personalized prediction machines that can tell whether somebody will find a house ugly or appealing, we can remove the design evaluation from other considerations. We can prevent shifting standards or after-the-fact justifications for a biased aesthetical judgment.

Also, this work offers new insights into the heterogeneity of personal tastes: The diversity in aesthetic preferences when it comes to housing might be larger than developers may believe – Using an array of personalized ML-enabled classifiers allows them to test designs before making a well-informed decision. Potentially, this could lead to fewer home

[1] `https://www.architectsjournal.co.uk/news/all-design-plans-for-cambridge-digs-on-stilts-rejected`

builders playing safe and opting for bland cookie-cutter homes.

Human aesthetic ratings are often inconsistent and are also evolving in time. From an ML-modeling perspective, surveyed ratings of building designs, for instance, are not hard ground truth data. At best, ratings are noisy proxies for a ground truth that eludes direct observation. Contrasting an imperfect yet time-consistent ML classifier with a more dynamic 'human classifier' provides insights into the black box of aesthetic judgments, or rather the ensemble of black boxes that jointly form an opinion each time we are looking at houses.

To emphasize the inconsistency of aesthetical evaluations, we collect ratings from a large number of participants in a way that makes consistency difficult for most people. We ask participants simply whether they like or dislike a sample of houses shown in photos. The simple question is not that easy since most participants cannot break the problem down into easy formulas. Many iterations make the task tedious and tiring influencing the mood of the classifier. Also, people learn throughout the data collection and constantly update their explicit and implicit criteria and benchmarks. Our findings are relevant for a growing body of behavioral and experimental housing research that uses images in combination with user-generated ratings on e.g. the perceived safeties of streets Naik et al. (2016) or the attractivity of places. How big is the conceptual problem of potentially biased training data really?

Finally, an automated classifier for 'soft' characteristics such as the exterior or interior design of homes could automate parts of the search process and reduce costs for buyers. Also, it improves the selection of comparables for appraisals, as shown in Lindenthal (2020).

2 Literature Review

2.1 Inconsistency in Housing Preference and Decision Making

Past literature has widely documented the heterogeneity in housing preferences, especially for the correlation between housing preference and socio-demographical factors. Eichholtz and Lindenthal (2014) link the demographics at the household level with the housing demand. Gebru et al. (2017) find that the external of residential buildings in a neighborhood can predict the socio-demographic factors of residents, such as income, race, education and voting patterns. Naik et al. (2016) also find that, in the 19 major American cities, the perception of safety of streetscapes is positively correlated with the household income in that street.

A large strand of literature has discussed the formation of and changes in heterogeneous housing preferences. Familiarity bias—the psychological phenomenon where people opt for the more familiar options rather than the optimal ones—explains why home buyers prefer properties that they feel more familiar with and tend to overestimate the values of these properties (Agarwal, 2007; Gonzalez-Navarro and Quintana-Domeque, 2009; Seiler et al., 2013). While some other papers discussed the potential changes in personal house preferences due to big events in life like migration and social integration (Büchel et al., 2020; Fan et al., 2020), most of these studies consider the formation and changes in housing preference as a gradual and slow process.

Fewer real estate studies, however, have investigated the potential inconsistency in housing preferences in an instant searching process, although such inconsistency in humans' decision-making process over a relatively short period has been well documented by the literature in many other fields like psychology and behavioral economics (Milkman et al., 2009; Sugden, 2021). An important strand of real estate literature discusses the reference points in mental accounting of home values using the theoretical prospect theory (Anenberg, 2011; Bokhari and Geltner, 2011; Genesove and Mayer, 2001), which implies that the personal home valuations might change given the updates in market in-

formation. Nevertheless, there still exists a knowledge gap in the short-term consistency in the more latent perceptions of real estate, such as the preferences for home aesthetics (Coburn et al., 2019).

2.2 ML in Real Estate and Urban Economics

There is a rapidly growing literature in real estate and urban economics using ML techniques and novel unstructured data such as texts and images (Aubry et al., 2019; Shen and Ross, 2020). One important application of ML and computer vision is to understand human perceptions of housing quality and urban environment, which is costly and challenging with traditional methods like surveys (Koch et al., 2019).

It has been documented in the literature that home aesthetics carry impacts on the market housing price. Coulson and McMillen (2008) disentangle the vintage effects on housing prices from other temporal variables like building age and time of sale. Francke and van de Minne (2017) further separate the effects of vintage and external obsolescence on housing price from the effects of physical deterioration and functional obsolescence. Using over 60,000 transactions in the Netherlands, Buitelaar and Schilder (2017) find a significant price premium of 15% for houses built in a neo-traditional architectural style.

Apart from the direct effect of building vintage on housing prices, the exteriors of the building also introduce externalities that spill over to the market price of surrounding buildings. Ahlfeldt and Mastro (2012) estimate that homeowners' willingness to pay increases by 5% to 8.5% for residential buildings adjacent to iconic architectural landmarks. For the city of Rotterdam, homogeneous ensembles in neighborhoods are found to exert a price premium of approximately 3.5% (Lindenthal, 2020).

The majority of these studies use traditional approaches, such as human assessments by experts or surveys, to measure the building vintage (Freybote et al., 2016). However, human assessment is normally costly and time-consuming, and it is also threatened by limited sample sizes and large bias from unobserved factors. Other studies use indirect measures for the building styles, such as the zoning of conserved buildings or the introduc-

tion of redevelopment projects, to achieve cleaner institutional settings for the evaluation (Ahlfeldt et al., 2017). Unfortunately, few of these approaches can scale up well.

Emerging literature aims to address these challenges by applying deep learning techniques to classify human perceptions towards housing aesthetics. These studies affirm the impact of building appearance on housing prices, although the magnitude of the effect varies across market sectors and cycles (Glaeser et al., 2018; Lindenthal and Johnson, 2021; Johnson et al., 2020). The uniqueness of building vintages relative to the surrounding homes also impacts the reference point of home sellers, which leads to a more pronounced rounding of prices (Schmidt and Lindenthal, 2020).

3 Methodology

3.1 Image Classifier

We train personalized image classifiers for the aesthetics of residential real estate based on deep convolutional neural networks (CNN) and transfer learning techniques (Glaeser et al., 2018; Lindenthal and Johnson, 2021). Specifically, we first transform each image into a 2,048-dimensional feature vector using the Inception computer vision model pretrained by Szegedy et al. (2016). Then, we include additional layers in the model to map the feature vectors to each participant's ratings, which gives us a personalized classifier of housing aesthetics.

Unlike training image classifiers with large labeled databases, one main technical challenge in our task is to deal with imbalanced and small samples in classes (Buda et al., 2018). For example, people may dislike many images but like only a few ones or vice versa. If we include imbalanced samples in our training procedure, the model may over-fit the class (i.e., "liked" or "disliked") with more training samples. Common methods to deal with the imbalanced data include under-sampling (e.g., using fewer images from the larger class), over-sampling (e.g., using data augmentation to increase the sample size of the smaller class), and class-weight modification in models (Cui et al., 2019). In our models,

we use the under-sampling method for two reasons. First, applying data augmentation, such as rotation, distortion or flipping, to the smaller class could potentially impact the aesthetics of houses and influence the accuracy of the ratings. Second, applying class-weight adjustments may introduce biases when we manually tune the parameters of each individual model.

More specifically, for each participant in the experiment, we train ML models to predict whether they like an image using the following steps. We define the bottleneck-class sample size of a participant i as the smaller one in the numbers of images they like and dislike and denote it by $bottleneck_i$. Firstly, we randomly select 100 images that participant i likes and another 100 that the same participant i dislikes as the out-of-sample testing data. Secondly, we use all remaining $bottleneck_i - 100$ images in the smaller class, and another $bottleneck_i - 100$ images randomly selected in the larger class as the in-sample training data. In other words, we sampled a balanced testing data with 200 images and a balancing training data with $2 \times (bottleneck_i - 100)$ images. Thirdly, we train an ML model (with the same model setting and parameters) and generate predictions on the out-of-sample data with the model. We repeat this process 10 rounds. In the 11th round, we use the images for which we have not obtained predictions in the previous 10 rounds as the new out-of-sample testing data, and the same under-sampling method is applied to draw training data from the remaining images. Lastly, we use the majority vote to determine the classification of an image. If an image is both classified to be "liked" and "disliked", we select the class with the higher average classification score.

3.2 Object Detection

To analyze the consistency in participants' rankings, we also apply ML techniques to construct two variables that describe the features of the houses and their surrounding environments. The first variable, *Trees*, is defined as the fraction of the image size taken up by trees. The second variable, *Density*, represents the housing density and is approximated with the log number of houses in an image. These two variables are

selected because trees and housing density are well-documented to impact home preferences (Dehring and Dunse, 2006; Wachter and Wong, 2008). Following the methodology in Wan and Lindenthal (2021), we automatically detect trees and houses in an image using the Inception object detection model. We also obtain a rectangular mask that is drawn tightly around each identified object, with which we can calculate the area of the identified object.

4 Data

We select the first 3,000 images of homes from Flickr after searching for the keyword "house".[2] Then, we manually remove images of derelict houses, toy houses, artwork, etc. Nevertheless, we do not remove homes atypical in style (e.g., English Mansions). The geographic coverage is global, but North America and Europe are over-represented. There are 2,139 images left after the sample screening.

The advantage of using images from a photography platform is that the images are of better quality than those retrieved from e.g. Google Street View. Also, the buildings are more interesting since photographers do not take pictures of a representative stock. For our task, this is desirable, as we hope for a diverse and divisive sample of photos that will lead to strong variation in ratings.

We designed an app that lets participants like or dislike images on their mobile phones[3] and Appendix Table B1 shows the user interface of the app. Participants log into the app anonymously and are first asked to fill in a short survey about their background (e.g., gender, age, ethnicity, education, etc.) After that, the images are shown to each participant in the same order sequentially.[4] The participants can swipe an image to the right if they like it, or to the left if they dislike it. They may also choose to terminate the experiment early if they do not want to rate the entire sample.

[2]These images are offered by the authors under a creative commons license.

[3]`https://4walls.cremll.com`

[4]We randomized the images to ensure that the key image features we study, including tree and density, are not correlated with the order of appearance in the app (Appendix Table A2).

We collected responses from 133 participants in total. To permit training a reliable ML model of each participant based on our methodology, we include only samples by participants who have liked or disliked at least 200 images each. This yields a sample of 36 participants and 43,150 ratings for our main regression analysis.

Table 1 reports the summary statistics of our main sample. The definition of variables is reported in Appendix Table A1. Panel A of Table 1 presents the summary of ratings. On average, 36.7% of the images are liked by the participants. Participants spend around 2.5 seconds rating an image on average. Panel B reports the summary at the participant level. Each person likes 39.1% of the images. All participants have at least rated the first 488 images. Our sample has relatively even distributions in terms of gender and age group, but white people with higher education levels are over-sampled.

— Insert Table 1 about here —

5 Results

5.1 Baseline Results

The first set of descriptive results from the experiment reveal a wide distribution of tastes across the participants. If all had similar aesthetic preferences, the distribution of the share of likes per image should be bimodal: Either the majority likes an image or not. This is clearly not the case, as Figure 1 plots the share of participants that liked an image. (a) is based on the subset of images that were ranked at least by 70 participants while (b) is derived from a larger sample of images that have been ranked by at least 10 participants. The black lines represent estimated densities. Both distributions suggest substantial differences in tastes; participants that rank more images like fewer of them (Figure 1b). In Appendix B, we also show examples of houses that most participants liked (Figure B2), disliked (Figure B3), or have mixed opinions on (Figure B4).

— Insert Figure 1 and Figure 2 about here —

Figure 2 adds a time dimension to the picture. The horizontal axis represents the order in which the images appeared in the app. The vertical axis denotes the share of participants that liked an image. For the first half of the images presented in the app, the rankings were on average higher and more varied than for the second half. Possibly, participants refine their criteria, learn about other images in the sample, and become more critical. In addition, the monotonous nature of the task might lead to tiredness or inattention.

How consistent are the participants' rankings? A few randomly selected images were presented to participants twice. Figure 3 plots the share of consistent rankings of the repeated images conditional on the number of other images shown before the image re-emerges (horizontal axis). The vertical axis denotes the share of consistent rankings of the repeat images. Overall, participants rank images consistently, with almost perfect consistency for quick repetitions. However, stated preferences are more likely to change when many other images are shown before an image is presented for ranking again.

Figure 4 pictures the relationship between the prediction accuracy of the personalized ML models and the training data sample size. Training data are balanced and feature the same number of likes and dislikes, which implies that the maximum training data depends not only on the overall number of images ranked by users but also on their respective share of likes.[5] The vertical axis is the F_1-*score*, the harmonic mean of *Precision* and *Recall*, of the ML model. Overall, all models arrive at an out-of-sample predictive performance above the value of 0.5 which a random predictor (e.g. tossing a coin) would achieve. The variation in the predictive power of the personalized ML classifiers varies strongly across respondents, as visualized by the vertical dispersion in Figure 4. Not surprisingly, larger training data samples lead to higher F_1-*scores*. The distributions of *Precision, Recall,* and F_1-*scores* summarized in Table 2 highlight the heterogeneity of the fit between human and machine classifications once more.

— Insert Figure 3, Figure 4, and Table 2 about here —

[5]See Table 1B for a distribution of % Likes per user.

5.2 Preference (Dis-)Similarities and Rating Persistence

Next, we investigate how different tastes between pairs of participants are and, more importantly, whether these differences are preserved in the predictions from the personalized ML classifiers. If all ML classifiers, for instance, responded to some image details that are not meaningful for human classifiers, then the ML classifications should be more correlated between participants than the training data. Figure 5 plots estimated density functions for the distribution of ranking similarities across participants. For any two users, a similarity score is calculated as the number of image rankings they agreed on (either liked or disliked) divided by the number of images they have both rated. The red line shows the distribution of similarity scores in ratings by participants while the blue line represents the distribution of similarity scores in predicted ratings (ML).

Again, strong differences in tastes are evident as the average similarity score is estimated to be 0.598. Interestingly, the distribution of the similarity scores based on ML predictions does not differ much, notwithstanding a fatter tail to the right.

— Insert Figure 5 and Figure 6 about here —

Moving on from pairwise similarities to clustering analysis of aesthetic tastes we test whether predicted tastes can be captured by fewer clusters than the participants' direct responses.

The horizontal axis in Figure 6 denotes the number of clusters (K) and the vertical axis is the total sum of squared errors within clusters. The red line represents the ratings and the blue line represents the ML-predicted ratings. With the blue line below the red line, we conclude that groups of predicted aesthetic preferences have more in common than groups of directly revealed preferences.

The ML-enabled classifiers will arrive at persistent predictions by construction. The training of models is separated from the prediction. Human minds will constantly update the internal model when making new predictions and therefore might rate images differently over time. Mood, boredom, blood sugar levels, or other factors might influence a

participant's classification. In the next section, we will use the consistent ML ratings to better understand the dynamics in the participants' ratings.

Table 3 presents estimated marginal effects at means from a logit regression in which participants' ratings (*Liked by Participant*) are partially explained by the ML rating (*Liked by ML*) and additional covariates.

The coefficients on the ML classifications are positive and statistically significant in all 5 model specifications. The magnitudes of the marginal effects are also sizeable, given the mean of 0.37 (Table 1). User fixed effects account for differences in the propensity to like an image within participants.[6] Overall, the machines can capture the aesthetic preferences of 'their' users well.

The coefficients on the share of the image area taken up by trees in Model (2) offer a first glimpse of the time-varying nature of participants' classifications. If all information from the images is reflected in the ML classifications then adding the *Trees* variable will not change the coefficients on *Liked by ML*. We find a positive and significant relationship between trees and users liking an image. In a sense, greenery can improve the perception of buildings that are, subjectively, not architectural masterpieces.

In addition, the negative coefficient on the interaction term *Liked by ML* $\times$ *Trees* indicates that the ML classifications are less helpful when trees are present. This could be explained either by the ML classifiers not picking up the presence of trees or by participants changing their minds about trees or focusing more on the actual home, consciously or unconsciously, while rating images. The latter is more likely, as Table 4 will show. At the beginning of the experiment, users respond more positively to trees but start to ignore them once they have rated more pictures. The ML classifier tries to learn from these contradictory inputs – and offers an imperfect compromise as a prediction. As a result, the predictive power of the ML classifier will decrease for images containing trees.

In contrast, the coefficient on our measure of development density (the number of

[6]We do not further analyze the coefficients on the demographic control variables and leave that for a future version of this paper when we have a larger and more representative sample of participants.

buildings detected in an image), is significant and negative (Model 3) while the interaction term with *Liked by ML* is statistically insignificant. This means that the ML classifier might not be able to fully incorporate the density information (or the surrounding buildings make the house itself look less attractive). However, the main insight is, that the participants do not *update* their (negative!) attitudes towards density while rating and thus do not present a moving target for the ML classifier while training.

Not all images are equally easy to rate. We capture the *Rate Time* between the timestamps when an image is sent to a participant and when the response is received by the app. We assume that the time needed to rate an image is not randomly distributed but partially depends on the house shown in the images. For some images, participants might need more time to make up their minds, or they might enjoy looking at the scenery. We do not know which one it is – but we observe that the predictions by the ML classifiers are less meaningful for images with longer rating times (negative coefficient for interaction term), which suggests that some images are simply more difficult to rate, for both the machines and humans.

Reassuringly, combining the previously independently discussed variables into the more comprehensive Model (5) does not change the coefficient estimates qualitatively.

— Insert Table 3 about here —

Table 4 offers more insights into the dynamic black box between our ears. When comparing the responses to images appearing early in the app to images that are presented later we find that people tend to like later images less. Apparently, people are careful to reject in the beginning, as they need time to find their bearings and to learn. Later the ratings become more critical.

Model (2) confirms the changes in attitudes towards trees: Initially, more trees lead to more likes and the effect is twice as strong for images from the first half than for images appearing later. Potentially, respondents learn to abstract from surroundings and focus on the building more? Model (3), however, shows that density, indeed, is disliked

throughout. This preference appears to be robust in time with no updating in the second half of the sample.

Again, we find that images that participants inspect for longer are liked more often (Model 4). This complexity effect is stronger for images appearing later. The coefficient on *log(rate time)* is not statistically significant for early images but becomes significant later on.

— Insert Table 4 about here —

6 Conclusion

ML-enabled classifiers are regularly criticized for being 'black boxes': While their predictive power is undisputed, it is difficult to understand why the model arrived at a particular classification. The same can be said for humans classifying photos according to their aesthetic appeal. They can quickly say whether they like a photo or not – but giving justifications for such a choice is often challenging. Also, human classifiers exhibit inconsistencies and biases, adding to the black-box nature of their classifications.

This paper first collects binary classifications of house pictures from a large group of participants and then trains personalized ML classifiers for each participant. The heterogeneity of personal tastes is preserved in the ML predictions: The automatic classifiers are useful 'digital twins' that could be used to evaluate designs, assess building applications, search for suitable homes, or to find comparables. The predictive power is far from perfect but already useful with F_1-scores ranging between 0.6 and 0.7.

Predictions from these automated yet personal classification machines shed light on biases and inconsistencies in the participants' assessment of residential real estate's visual appeal. In the coming months, we will increase the number of participants, train additional personal classifiers, and research whether preferences and biases can be linked to demographic characteristics – the current sample is still too small and unbalanced in that respect.

In addition, we intend to generate two sets of new images of houses based on generative adversarial networks (GAN). The first set will be trained on images the human classifiers 'liked' whereas the second will be based on images the personalized ML classifier predicted as 'liked' by the participant. Finally, participants will rate the generated images from both sets. We can test whether imperfect but consistent classifications might represent the unobservable aesthetical 'ground truth' better than ratings from a better but time-varying classifier.

References

S. Agarwal. The impact of homeowners' housing wealth misestimation on consumption and saving decisions. *Real Estate Economics*, 35(2):135–154, 2007.

G. Ahlfeldt and A. Mastro. Valuing iconic design: Frank lloyd wright architecture in oak park, illinois. *Housing Studies*, 27(8):1079–1099, 2012.

G. M. Ahlfeldt, K. Moeller, S. Waights, and N. Wendland. Game of zones: The political economy of conservation areas. *The Economic Journal*, 127(605):F421–F445, 2017.

E. Anenberg. Loss aversion, equity constraints and seller behavior in the real estate market. *Regional Science and Urban Economics*, 41(1):67–76, 2011.

M. Aubry, R. Kräussl, G. Manso, and C. Spaenjers. Machine learning, human experts, and the valuation of real assets. *CFS Working Paper Series*, 2019.

S. Bokhari and D. Geltner. Loss aversion and anchoring in commercial real estate pricing: Empirical evidence and price index implications. *Real Estate Economics*, 39(4):635–670, 2011.

K. Büchel, M. V. Ehrlich, D. Puga, and E. Viladecans-Marsal. Calling from the outside: The role of networks in residential mobility. *Journal of Urban Economics*, 119:103277, 2020.

M. Buda, A. Maki, and M. A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

E. Buitelaar and F. Schilder. The economics of style: Measuring the price effect of neo-traditional architecture in housing. *Real Estate Economics*, 45(1):7–27, 2017.

A. Coburn, O. Kardan, H. Kotabe, J. Steinberg, M. C. Hout, A. Robbins, J. MacDonald, G. Hayn-Leichsenring, and M. G. Berman. Psychological responses to natural patterns in architecture. *Journal of Environmental Psychology*, 62:133–145, 2019.

N. E. Coulson and D. P. McMillen. Estimating time, age and vintage effects in housing prices. *Journal of Housing Economics*, 17(2):138–151, 2008.

Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

C. Dehring and N. Dunse. Housing density and the effect of proximity to public open space in aberdeen, scotland. *Real Estate Economics*, 34(4):553–566, 2006.

P. Eichholtz and T. Lindenthal. Demographics, human capital, and the demand for housing. *Journal of Housing Economics*, 26:19–32, 2014.

Y. Fan, H. P. Teo, Y. Tu, and W. X. Wan. The road to integration: Post-migration experience and migrant housing behavior in singapore. *Available at SSRN*, 2020.

M. K. Francke and A. M. van de Minne. Land, structure and depreciation. *Real Estate Economics*, 45(2):415–451, 2017.

J. Freybote, L. Simon, and L. Beitelspacher. Understanding the contribution of curb appeal to retail real estate values. *Journal of Property Research*, 33(2):147–161, 2016.

T. Gebru, J. Krause, Y. Wang, D. Chen, J. Deng, E. L. Aiden, and L. Fei-Fei. Using deep learning and google street view to estimate the demographic makeup of neighborhoods across the united states. *Proceedings of the National Academy of Sciences*, 114(50): 13108–13113, 2017.

D. Genesove and C. Mayer. Loss aversion and seller behavior: Evidence from the housing market. *The Quarterly Journal of Economics*, 116(4):1233–1260, 2001.

E. L. Glaeser, M. S. Kincaid, and N. Naik. Computer vision and real estate: Do looks matter and do incentives determine looks. *Working paper*, 2018.

M. Gonzalez-Navarro and C. Quintana-Domeque. The reliability of self-reported home values in a developing country context. *Journal of Housing Economics*, 18(4):311–324, 2009.

Greater Cambridge Shared Planning, 2021. URL `https://applications.greatercambridgeplanning.org/online-applications/applicationDetails.do?activeTab=neighbourComments`.

E. B. Johnson, A. Tidwell, S. V. Villupuram, et al. Valuing curb appeal. *Journal of Real Estate Finance and Economics*, 60(1):111–133, 2020.

D. Koch, M. Despotovic, S. Leiber, M. Sakeena, M. Döller, and M. Zeppelzauer. Real estate image analysis: A literature review. *Journal of Real Estate Literature*, 27(2): 269–300, 2019.

T. Lindenthal. Beauty in the eye of the home-owner: Aesthetic zoning and residential property values. *Real Estate Economics*, 48(2):530–555, 2020.

T. Lindenthal and E. B. Johnson. Machine Learning, Building Vintage and Property Values. *Journal of Real Estate Finance and Economics*, 2021.

K. L. Milkman, T. Rogers, and M. H. Bazerman. Highbrow films gather dust: Time-inconsistent preferences and online dvd rentals. *Management Science*, 55(6):1047–1059, 2009.

N. Naik, R. Raskar, and C. A. Hidalgo. Cities are physical too: Using computer vision to measure the quality and impact of urban appearance. *American Economic Review*, 106(5):128–32, 2016.

C. Schmidt and T. Lindenthal. The odd one out: Asset uniqueness and price precision. *Working paper*, 2020.

M. J. Seiler, V. L. Seiler, D. M. Harrison, and M. A. Lane. Familiarity bias and perceived future home price movements. *Journal of Behavioral Finance*, 14(1):9–24, 2013.

L. Shen and S. Ross. Information value of property description: A machine learning approach. *Journal of Urban Economics*, page 103299, 2020.

R. Sugden. Hume's experimental psychology and the idea of erroneous preferences. *Journal of Economic Behavior & Organization*, 183:836–848, 2021.

C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

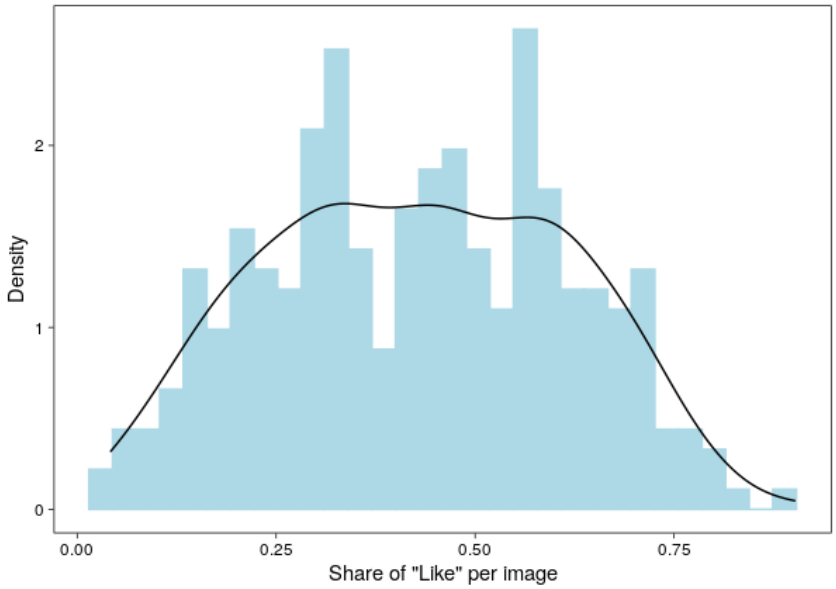
K. Vonnegut. *Breakfast of Champions*. Delacorte Press, 1973.

S. M. Wachter and G. Wong. What is a tree worth? green-city strategies, signaling and housing prices. *Real Estate Economics*, 36(2):213–239, 2008.

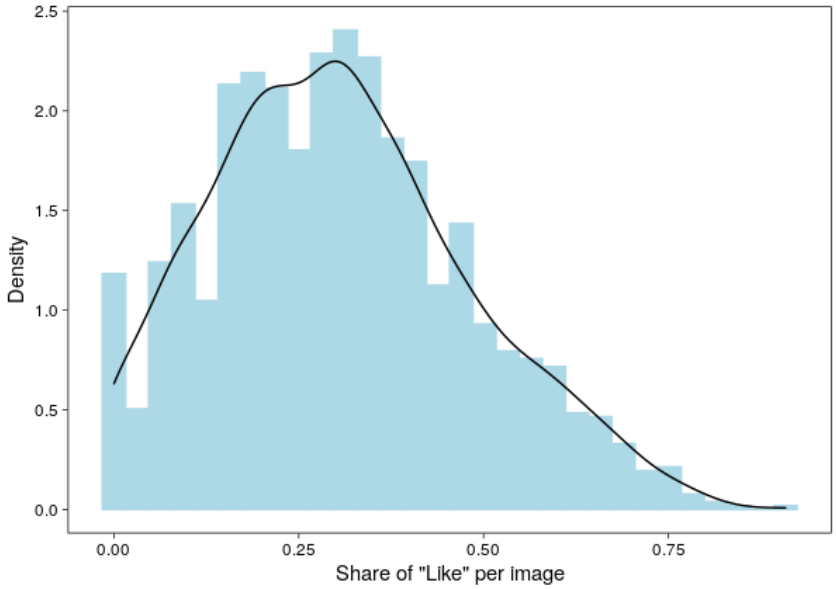
W. X. Wan and T. Lindenthal. Towards accountability in machine learning applications: A system-testing approach. *Available at SSRN: https://ssrn.com/abstract=3758451*, 2021.

Tables and Figures

Figure 1: Share of "Likes" per Image



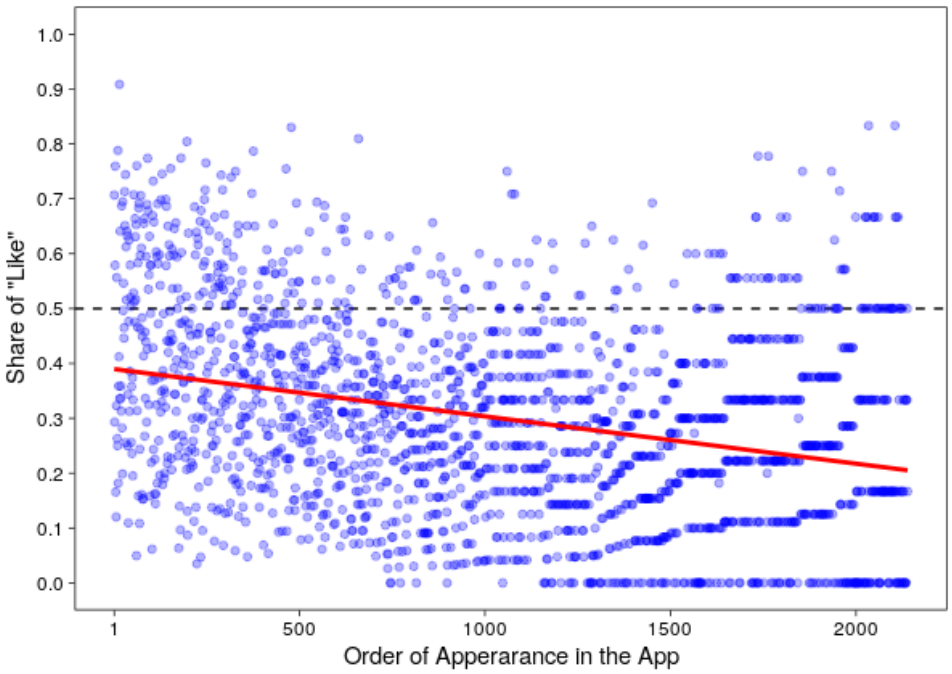
(a) Images ranked by at least 70 participants



(b) Images ranked by at least 10 participants

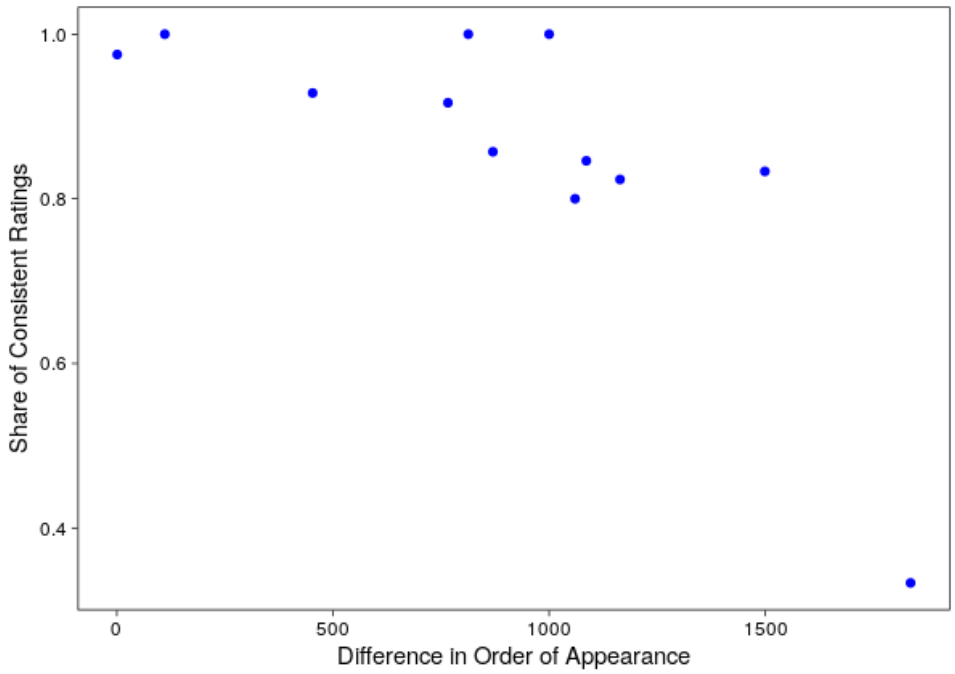
Notes: This figure plots the share of participants that liked an image. (a) is based on the subset of images that were ranked at least by 70 participants while (b) is derived from a larger sample of images that have been ranked by at least 10 participants. The black lines represent estimated densities. Both distributions suggest substantial differences in tastes: If all participants had identical tastes, a bimodal distribution would emerge. This is clearly not the case. Participants that rank more images like fewer of them (b).

Figure 2: Share of "Likes" per Image, in Order of Image Appearance



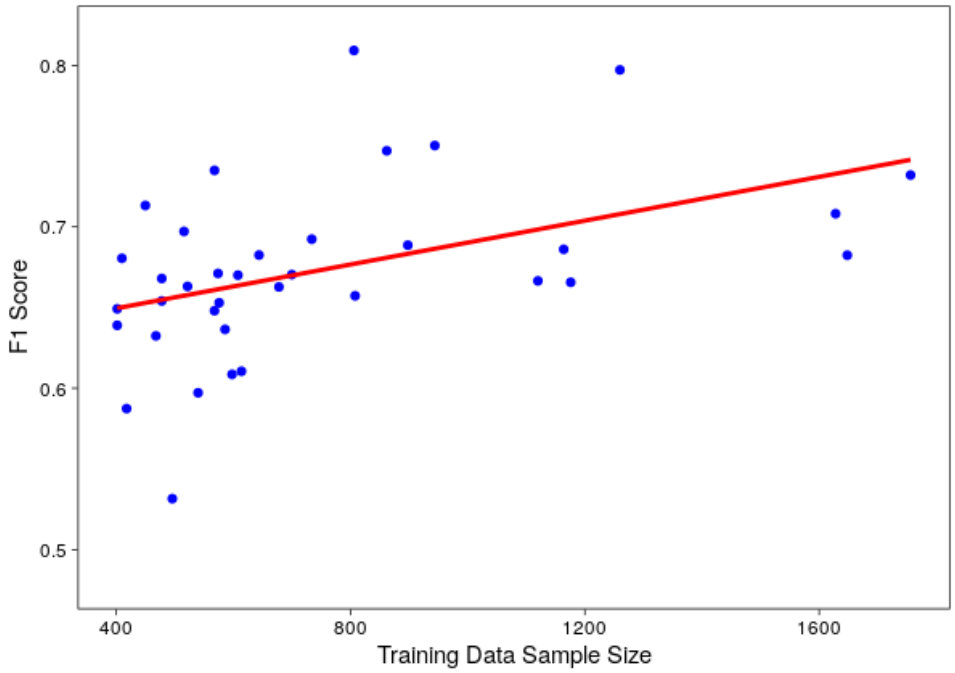
Notes: The horizontal axis represents the order in which the images appeared in the app. The vertical axis denotes the share of participants that liked an image. For the first half of the images presented in the app, the rankings were on average higher and more varied than for the second half. Possibly, participants refine their criteria, learn about other images in the sample, and become more critical. The red line depicts a fitted linear trend.

Figure 3: Consistency of Rating for Repeated Images



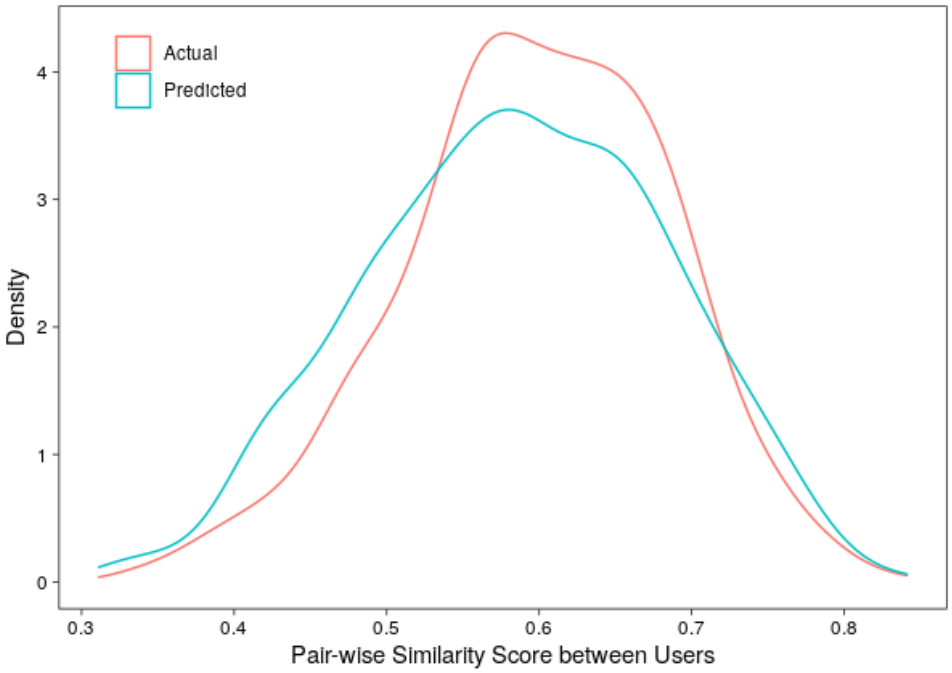
Notes: A few randomly selected images were presented to participants twice. This figure plots the share of consistent rankings of the repeated images conditional on the number of other images shown before the image re-emerges (horizontal axis). The vertical axis denotes the share of consistent rankings of the repeat images. Overall, participants rank images consistently. However, stated preferences are more likely to change when many other images are shown before the repeated ranking.

Figure 4: ML Model Accuracy and Training Sample Size



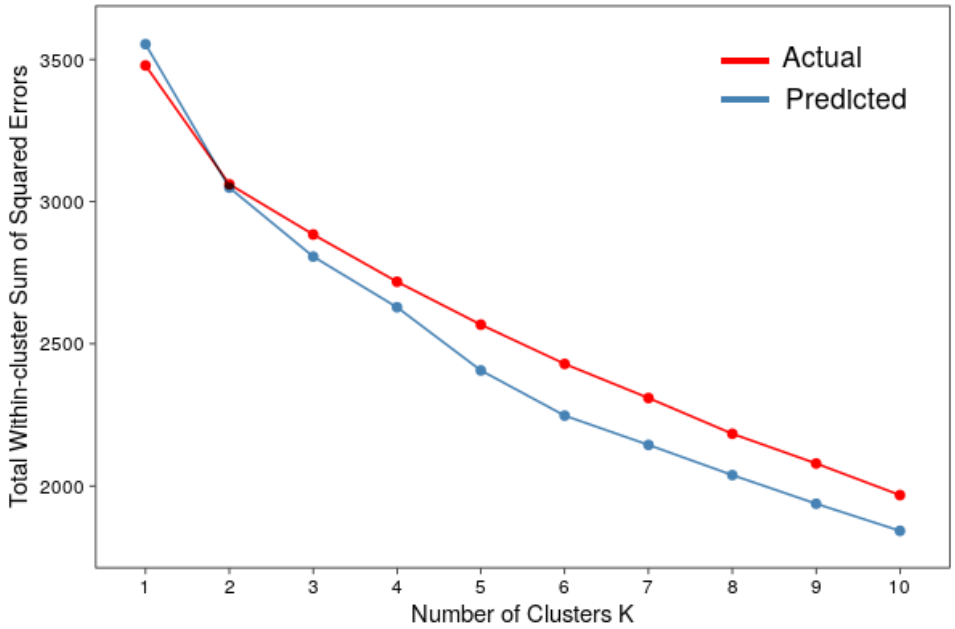
Notes: This figure shows the relationship between the prediction accuracy of the ML model and the training data sample size. Training data are balanced and feature the same number of likes and dislikes, which implies that the maximum training data depends not only on the overall number of images ranked by users but also on their respective share of likes (see Table 1B for a distribution of % Likes per user). The vertical axis is the F_1-*score* of the ML model. Overall, all models arrive at an out-of-sample predictive performance above 0.5 (also see Table 2). The predictive power of the personalized ML classifiers varies strongly across respondents, as visualized by the vertical dispersion in the figure. Not surprisingly, larger training data samples lead to higher F_1-*scores*. Fitted line in red.

Figure 5: Density Distributions of Human and ML Classifications



Notes: This figure plots estimated density functions for the distribution of ranking similarities across participants. For any two users, a similarity score is calculated as the number of image rankings they agreed on (either liked or disliked) divided by the number of images they have both rated. The red line shows the distribution of similarity scores in ratings by participants while the blue line represents the distribution of similarity scores in predicted ratings (ML).

Figure 6: Cluster Analysis of Aesthetic Tastes: Participants' Ratings vs. ML Predictions



Notes: This figure presents the clustering analysis results of aesthetic tastes in ratings and ML-predicted ratings, using the K-means method. The horizontal axis denotes the number of clusters (K) and the vertical axis is the total sum of squared errors within clusters. The red line represents the ratings and the blue line represents the ML-predicted ratings. Predicted user preferences apparently are easier to capture in fewer clusters than the participants' direct rankings (blue line below red line).

Table 1: Summary Statistics

A. Summary by Image								
	(1) N	(2) mean	(3) sd	(4) min	(5) p25	(6) p50	(7) p75	(8) max
Liked by Participant	43,150	0.367	0.482	0	0	0	1	1
Liked by ML	43,150	0.450	0.498	0	0	0	1	1
Trees	43,150	0.276	0.235	0.000	0.074	0.226	0.428	0.977
Density	43,150	1.113	0.406	0.000	0.693	1.099	1.386	2.890
Rate Time	43,150	0.938	0.368	0.000	0.693	0.693	1.099	2.398
B. Summary by Participant								
	(1) N	(2) mean	(3) sd	(4) min	(5) p25	(6) p50	(7) p75	(8) max
Share of Liked by Participant	36	0.394	0.169	0.132	0.259	0.383	0.510	0.736
Share of Liked by ML	36	0.460	0.079	0.273	0.391	0.469	0.519	0.581
Number of Images Rated	36	1,199	521	488	754	1,088	1,485	2,139
Gender	36	0.472	0.506	0	0	0	1	1
Ethnicity	36	2.278	0.741	1	2	2	2	4
Age Group	36	4.583	2.247	1	3	4	6	9
Education Level	36	3.222	0.959	1	3	3	4	4

Notes: The table presents the summary statistics of the key variables in our analysis. Panel A reports the summary by images and Panel B reports the summary by participants. The definition of the variables is presented in Appendix Table A1.

Table 2: Prediction Accuracy of Personalized ML Classifiers

	(1) N	(2) mean	(3) sd	(4) min	(5) p25	(6) p50	(7) p75	(8) max
Precision	36	0.6821	0.0560	0.6014	0.6524	0.6740	0.7112	0.8590
Recall	36	0.6677	0.0664	0.4520	0.6500	0.6675	0.7015	0.7790
F_1-score	36	0.6735	0.0548	0.5316	0.6489	0.6690	0.6936	0.8093

Notes: This table summarises the prediction accuracy of 36 ML classifiers trained on each participant's image ratings. *Recall* is the share of ratings predicted correctly while *Precision* is the share of images correctly predicted to be liked by a participant. F_1*-scores* are the harmonic means of *Precision* and *Recall*: $F_1\text{-}score = \frac{2 \times Recall \times Precision}{Recall + Precision}$

Table 3: Explanatory Power of the Personalized ML Classifiers

	(1)	(2)	(3)	(4)	(5)
	Logit Model				
	Y: Liked by Participant				
Liked by ML	0.2832***	0.3128***	0.2749***	0.3989***	0.4286***
	(0.0107)	(0.0111)	(0.0179)	(0.0256)	(0.0299)
Trees		0.1181***			0.1069***
		(0.0185)			(0.0180)
Liked by ML × Trees		-0.1087***			-0.1050***
		(0.0197)			(0.0197)
Density			-0.0322***		-0.0217***
			(0.0087)		(0.0083)
Liked by ML × Density			0.0074		-0.0024
			(0.0096)		(0.0099)
Rate Time				0.1201***	0.1190***
				(0.0217)	(0.0217)
Liked by ML × Rate Time				-0.1200***	-0.1184***
				(0.0204)	(0.0207)
Demographic Controls	Y	Y	Y	Y	Y
User Fixed Effects	Y	Y	Y	Y	Y
Observations	43,150	43,150	43,150	43,150	43,150
Pseudo-R^2	0.188	0.190	0.189	0.193	0.195

Notes: The table presents estimated marginal effects at means from a logit regression in which participants' ratings (*Liked by Participant*) are partially explained by the predicted rating *Liked by ML* and additional covariates. *Trees* is the share of the image area taken up by trees. *Density* equals the natural logarithm of the number of different buildings in the image. *Rate Time* is the natural logarithm of the time between the timestamps when an image is sent to a participant and when the response is submitted (in seconds). Demographic control variables (not shown) include gender, ethnicity, age, and education of participants. Robust standard errors are clustered by user and are reported in parentheses. *p < 0.1; **p < 0.05; ***p < 0.01.

Table 4: Consistency of Classification by Participants

	(1)	(2)	(3)	(4)	(5)
	Logit Model				
	Y: Liked by Participant				
Later Appearance	-0.0642***	-0.0442***	-0.0820***	-0.1084***	-0.0990***
	(0.0170)	(0.0159)	(0.0166)	(0.0239)	(0.0255)
Liked by ML	0.2715***	0.2713***	0.2721***	0.2700***	0.2703***
	(0.0123)	(0.0124)	(0.0123)	(0.0124)	(0.0126)
Liked by ML × Later Appearance	0.0231***	0.0225**	0.0220**	0.0225***	0.0212**
	(0.0088)	(0.0088)	(0.0089)	(0.0086)	(0.0086)
Tree		0.0931***			0.0831***
		(0.0181)			(0.0185)
Tree × Later Appearance		-0.0707***			-0.0693***
		(0.0187)			(0.0204)
Density			-0.0364***		-0.0290***
			(0.0081)		(0.0078)
Density × Later Appearance			0.0163		0.0096
			(0.0114)		(0.0123)
Rate Time				0.0376	0.0377
				(0.0295)	(0.0298)
Rate Time × Later Appearance				0.0520***	0.0519***
				(0.0179)	(0.0180)
Demographic Controls	Y	Y	Y	Y	Y
User Fixed Effects	Y	Y	Y	Y	Y
Observations	43,150	43,150	43,150	43,150	43,150
Pseudo-R^2	0.191	0.192	0.192	0.193	0.195

Notes: The table presents estimated marginal effects at means from a logit regression in which participants' ratings (*Liked by Participant*) are partially explained by the predicted rating *Liked by ML* and additional covariates. *Later Appearance* is a dummy variable denoting the second half of images rated by each user. *Trees* is the share of the image area taken up by trees. *Density* equals the natural logarithm of the number of different buildings in the image. *Rate Time* is the natural logarithm of the time between the timestamps when an image is sent to a participant and when the response is submitted (in seconds). Demographic control variables (not shown) include gender, ethnicity, age, and education of participants. Robust standard errors are clustered by user and are reported in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$

Appendix A: Supplementary Tables

Table A1: Definition of Variables

Variable Name	**Definition**
Liked by Participant	A dummy variable equal to 1 if the image is ranked as "liked" by a participant; otherwise, it equals 0.
Liked by ML	A dummy variable equal to 1 if the ML model predicts that a participant will like the image; otherwise, it equals 0.
Trees	The share of the image area taken up by trees.
Density	The number of houses detected in an image, in logarithmic form.
Rate Time	The log of 1 + the time between the timestamps when an image is sent to a participant and when the response is submitted (in seconds). Rating times over 10 seconds are winsorized at 10 seconds.
Share of Liked by Participant	The share of images rated as "liked" by a participant among all images rated by them.
Share of Liked by ML	The share of images predicted as "liked" by a participant's ML model among all images rated by them.
Number of Images Rated	The total number of images rated by a participant in the app.
Gender[a]	The gender of the participant, encoded as: 1 = Female; 2 = Male
Ethnicity	The ethnicity of the participant, encoded as: 1 = Others; 2 = Other white background (except British); 3 = Asian; 4 = White British
Age Group	The age group of the participant, encoded as: 1 = 20–24; 2 = 25–29; 3 = 30–34; 4 = 35–39; 5 = 40–44; 6 = 45–49; 7 = 50–54; 8 = 55–60; 9 = over 60
Education Level	The highest education level that the participant has obtained, encoded as: 1 = High school or lower; 2 = Bachelor; 3 = Master; 4 = PhD

Notes: This table presents the definitions of key variables in this study.

[a]While having more options in the initial survey, all participants identified unambiguously as female or male.

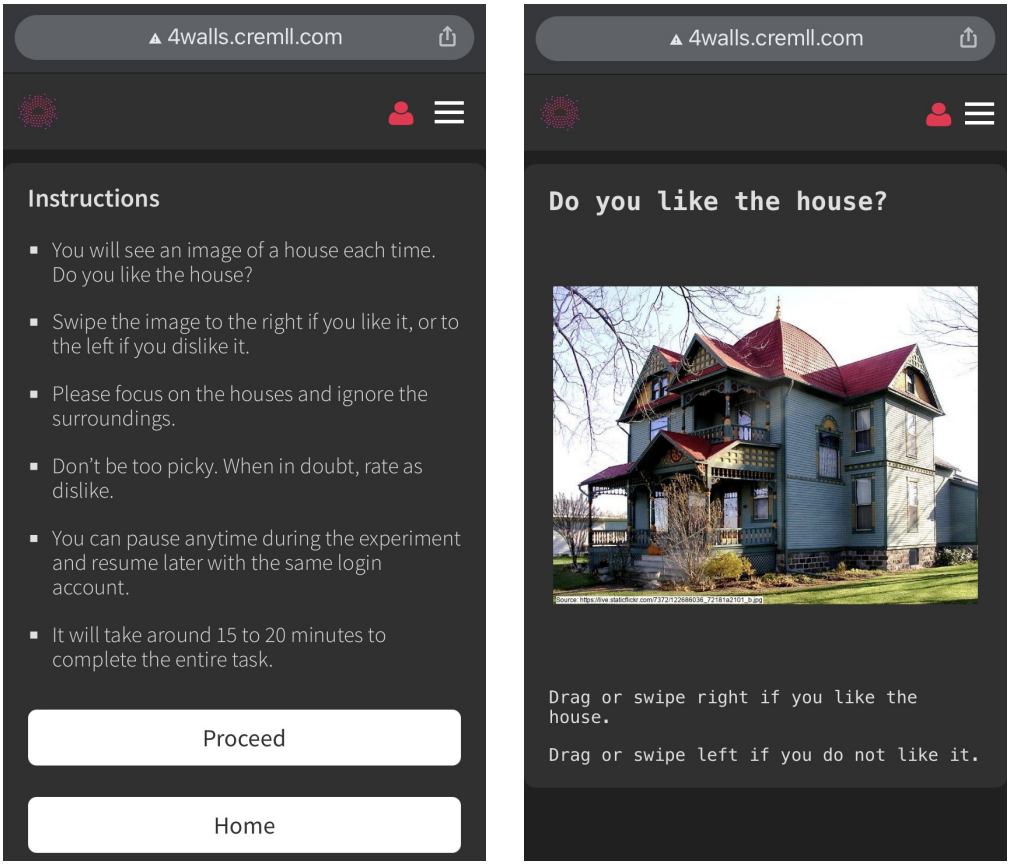
Table A2: Image Features and the Order of Appearance

	(1) Order of Appearance
Trees	0.4587
	(0.5803)
	[1.05]
Density	-0.5280
	(0.3397)
	[1.05]
Observations	2,139
R-squared	0.002

Notes: This table presents the regression results of image features (*Trees* and *Density*) on the order that an image appears in the app. Robust standard errors are reported in parentheses and variance inflation factors (VIF) in brackets. *p<0.1; **p<0.05; ***p<0.01.

Appendix B: Supplementary Figures

Figure B1: User Interface of the App



(a) The page of instructions

(b) The page to rate images

Notes: This figure shows the user interfaces of the app that we used in the experiment.

Figure B2: Examples of Images that Most Participants Liked

Notes: This figure shows some examples of images that most (over 80%) of the participants liked.

Figure B3: Examples of Images that Most Participants Disliked

Notes: This figure shows some examples of images that none of the participants liked.

Figure B4: Examples of Images that Have Most Mixed Opinions

Notes: This figure shows some examples of images that have the most mixed opinions (the share of "liked" is around 50%).