Exploring Biases Between Human and Machine Generated Designs

Christian E. Lopez
Industrial and Manufacturing Engineering
The Pennsylvania State University,
State College, PA 16802
e-mail: cql5441@psu.edu
Mem. ASME

Scarlett R. Miller
Engineering Design and
Industrial Engineering
The Pennsylvania State University,
State College, PA 16802
e-mail: shm13@psu.edu
Mem. ASME

Conrad S. Tucker[1]
Engineering Design and
Industrial Engineering
The Pennsylvania State University,
State College, PA 16802
e-mail: ctucker4@psu.edu
Mem. ASME

ABSTRACT

The objective of this work is to explore the possible biases that individuals may have towards the perceived functionality of machine generated designs, compared to human created designs. Towards this end, 1,187 participants were recruited via Amazon Mechanical Turk to analyze the perceived functional characteristics of both human created 2D sketches as well as sketches generated by a deep learning generative model. In addition, a computer simulation was used to test the capability of the sketched ideas to perform their intended function and explore the validity of participants' responses. The results reveal

[1] Corresponding author. 213 N Hammond Building, University Park, PA 16802, USA

that both participants and computer simulation evaluations were in agreement, indicating that sketches generated via the deep generative design model were more likely to perform their intended function, compared to human created sketches used to train the model. The results also reveal that participants were subject to biases while evaluating the sketches, and their age and domain knowledge were positively correlated with their perceived functionality of sketches. The results provide evidence that supports the capabilities of deep learning generative design tools to generate functional ideas and their potential to assist designers in creative tasks such as ideation.

1. INTRODUCTION

Recent advancements in generative design, topology optimization, and deep learning algorithms, are enabling designers to integrate computational tools into the design process at an increased pace [1]. Researchers argue that as these computational tools become more efficient at creating novel and functional ideas, they will foster designers' creativity. Hence, both machines and designers will co-create solutions that surpass each of their independently created ideas [2].

Deep learning algorithms are being implemented to automatically generate new design ideas [3,4]. Though an idea needs to be new and novel to be considered creative, it also has to meet its intended functionality and be useful [5]. During the latter stages of the design process, designers create CAD models and implement advanced computational methods to test the functionality of their design ideas. However, during the early stages of the design process, rough 2D sketches are typically the primary communication source of ideas [6]. During these stages, designers use their experience and domain knowledge to ensure that their new ideas are relevant to the design

problem at hand. Similarly, in the literature, crowdsourcing methods have been implemented to assess the ability of generative computational tools to produce new design ideas [3,4]. Recently, researchers have started exploring the functionality of 2D sketched ideas generated by computational tools using human raters [7]. If computational tools are to co-create new products and solutions alongside designers, their ability to produce not only novel, but also functional ideas, needs to be further explored. In this work, the term "computer generated" is used as an encompassing term to represent various deep learning-related methods of automated design generation.

The ability to generate creative ideas is an insufficient condition for innovation because decision-makers need to not only generate, but also select creative ideas for innovation to occur [8]. However, studies have shown that gender effects can influence the idea selection process [9,10]. Similarly, the educational level, experience, and domain knowledge of individuals have been related to individuals' risk attitudes and decision-making processes [11,12], while age has been related to technology adoption, acceptance, and perceived usability [13–15]. Hence, as designers integrate computational tools to assist in the design process, their possible bias towards computer generated and human created ideas, as well as the potential confounding effects of their demographic characteristics and domain knowledge, need to be explored. In light of this, the authors of this work present a crowdsourcing method to explore the perceived functional characteristics of 2D design sketches created by humans and 2D design sketches generated by a deep learning generative model, as well as the effects of participants' demographic characteristics and domain knowledge on

their perceived functionality of design sketches. Moreover, computer simulation is used to test the capability of the sketches to perform their intended function and test the validity of participants' responses.

2. LITERATURE REVIEW

2.1 Generative design

Generative design methods have captured the interest of both the design research and industry communities [16,17]. In Chandrasegaran et al. [1], the authors present a review of some of the challenges and future direction for computational support tools used in the product design process. Recently, designers have started to integrate deep learning models into their generative design methods. Deep learning models are a class of hierarchical statistical models composed of multiple interconnected layers of nonlinear functions [18]. Designers have gained a particular interest in Recurrent Neural Networks (RNNs) [19,20] and Generative Adversarial Networks (GANs) [3,4,21]. RNNs are deep learning models that contain multiple interconnected hidden layers. The hidden layers in an RNN are able to use information from their previous state via a recurrent weight layer, which allows them to have a recollection of their previous states [22]. GANs are deep learning generative models composed of a generator and a discriminator network. For example, the generator can be trained to generate new images that the discriminator classifies as "real" images (i.e., drawn from the same distribution as the training dataset). In contrast, the discriminator is trained to detect the generator's output images as being "fake" (i.e., classify images produced by the

generator as being drawn from a distribution other than the training data) [23]. This iterative game between the generator and discriminator results in GANs being capable of generating designs that are different from the training dataset (i.e., unique at a pixel level), while still maintaining some degree of similarity (see [22,23] for additional details).

Deep generative methods have been used to help in the representation of the design space. For example, Burnap et al. [3], train a deep generative model with a dataset of automotive designs able to generate new design ideas that morphed different body types and brands of vehicles. Dosovitskiy et al. [24] train a deep generative model to generate new 2D images of chairs. Kazi et al. [6] implement deep generative models into their *DreamSketch* tool. The *DreamSketch* tool takes as input, a rough 2D sketch, and generates multiple augmented solutions in 3D. Recently, Chen et al. [20] present a modification of Ha and Eck's *Sketch-RNN* model [19] capable of recognizing and generating 2D sketches from multiple classes. As highlighted by the authors, this model has the potential to help with creative tasks [20]. Deep generative methods have also been implemented to increase the veracity of big-data pipelines by generating new images [4]. However, an inherent challenge of these generative methods is that their objective to create new design ideas that still maintain a degree of similarity with the training data used are conflicting and challenging to evaluate. While studies have implemented pixel-level Euclidean distance and structured similarity indices to evaluate these methods, in many cases, these scores do not correlate to visual quality scores given by human raters [25].

2.2 Crowdsourcing and generative design validation

As a result of the current limitations in the evaluation metrics of generative models, researchers are starting to integrate crowdsourcing methods to evaluate their models. For example, Burnap et al. [3] use a crowdsourcing method to recruit 69 participants and assess the ability of their deep generative model to generate realistic designs. Their results show that their model was able to generate realistic designs while exploring the design space. Chen et al. [20] conduct a Turing test to compare the capability of 61 human raters and four deep learning models to distinguish between human and computer generated sketches. Their results reveal that some of the deep learning models outperformed the human raters in accurately distinguishing between human and computer generated sketches. Dering and Tucker [4] use 252 human raters to evaluate the capability of their method to generate new 2D sketches that were recognized to belong to a specific class. Their results indicate that human raters were able to accurately recognize the sketches of certain classes. These studies have analyzed the accuracy of human raters in classifying new images and sketches into specific classes, and not necessarily evaluating the functionality of sketches themselves.

Research indicates that crowdsourcing methods might constitute a promising paradigm for the product design process [26]. Table 1 shows a summary of existing literature related to deep generative design tools and the implementation of crowdsourcing methods used to evaluate them. Most of the current works focus on evaluating the capability of deep generative models to create new sketches that can be classified as belonging to a specific category. Though an idea needs to be new and novel

in order to be considered creative, it also has to meet its intended functionality and be useful [5].

During the latter stages of the design process, designers create CAD models and implement advanced numerical methods to evaluate the functionality of their design ideas. However, these methods are time-consuming and complex to implement, which limits their scalability [27]. Because of these limitations, researchers have started to explore how deep learning algorithms can be implemented to predict the ability of a 3D artifact to perform a function [28]. Nonetheless, during the early stages of the design process, detailed 3D models are not widely available, compared to rough 2D sketches. Sketches are typically the primary communication source of ideas, especially in the early phases of the design process [6,29]. Sketches can be categorized in terms of their intended purpose, design progression, and physical elements [30–32]. Based on their physical elements, Rodger et al. [31] present categories ranging from simple monochrome line drawings that do not include shading or annotations (Level 1), to high fidelity realistic sketches with extensive shading and annotations (Level 5). Several studies have used these taxonomies to evaluate design sketches and explore how they are used in the early stages of the design process [33–35].

Recently, researchers have started to integrate neural network algorithms and computer simulation to predict the functionality of 2D sketches generated via deep generative design models**.** Cunningham and Tucker [36] present a Validation Neural Network (VNN) that integrates a physics computer simulation. Frequently, during the initial stages of the design process, designers use their experience and domain

knowledge to ensure that generated ideas are relevant to the design problem. For example, experts have been used to evaluate and screen crowdsourced ideas [7,26]. Similarly, crowds have been used to evaluate the perceptual attributes of new designs [37]. For instance, in the previous study of this work, the authors implement a crowdsourcing method to recruit 983 raters and explore the perceived functionality of low-fidelity 2D sketches [7]. The results of the study reveal that participants perceived sketches generated via a deep generative model as more functional than human created sketches. Moreover, the results indicate that the perceived functionality of human generated sketches was negatively affected by explicitly presenting them as human generated sketches. Finally, the study reveals that participants were not able to accurately distinguish between the human created sketches and the computer generated ones.

While previous studies support the use of human raters to evaluate new ideas [26,37], the difference in the functionality evaluation of 2D sketch ideas between raters and computer simulation has yet to be explored. If computational tools are to co-create new products and solutions alongside designers, their capability to produce not only novel, but also functional ideas needs to be explored. Hence, in this work, the authors expand on their previous study and explore the functional characteristics of 2D sketches created by humans and sketches generated via a deep generative design model, using both computer simulation and crowdsourcing methods.

2.3 Designers' biases

The ability to generate creative ideas is an insufficient condition for innovation because decision-makers need to not only to generate, but also select creative ideas for innovation to occur [8]. Unfortunately, human bias can have a direct impact on the screening and selection of ideas [38]. Studies indicate that decision-makers can experience ownership [9], complexity [39], and even creativity biases [40]. Several studies have shown that the gender and risk attitudes of decision-makers can bias their selection of ideas [9,41]. Similarly, the educational level and experience of individuals has been related to their risk attitudes [11]. When evaluating the expertise of crowds, Burnap et al. [42] reveal that educational level and mechanical aptitude (e.g., domain knowledge) of raters was correlated to their capability to accurately evaluate design solutions. Besides gender, educational level, and experience, age is another factor that could affect designers' decision-making when interacting with deep generative design tools. Studies have indicated that age can affect technology adoption, revealing that younger individuals value the usefulness of technology more than older individuals [15]. This digital divide between generations is attributed to the fact that younger generations are exposed to digital technologies earlier in their life than older generations [43]. Moreover, studies indicate that technology acceptance and perceived usability are affected by age [13,14].

Besides decision-makers' biases towards creative ideas, researchers have recognized that individuals can be biased towards automated systems (i.e., *Automation bias*) [44,45]. One of the factors that contribute to *Automation bias* is the trust given to

automated support systems. This trust is the product of humans' perception of these systems as having superior analytical capabilities than their human counterpart [46]. For example, the results by Dzindolet et al. [47] indicate that participants expected an automated support system to outperform the human system in a visual detection task. Studies on *Automation bias* focus on safety and automation aids, and not directly on decision-makers' biases towards early stage conceptual design tools. Hence, as designers are increasingly integrating computational tools into the design process, their possible biases towards computer generated ideas, compared to human created ideas, need to be explored. Also, more research is needed to understand the possible biases and the effects that individuals' demographic characteristics and domain knowledge have on their perceived functionality of 2D design sketches.

In light of existing knowledge gaps, this work implements computer simulation and crowdsourcing methods to explore the functional characteristics of 2D design sketches generated via a deep learning generative model, compared to human created sketches. The computer simulation enables the virtual physics-based evaluation of sketches to perform their intended function. The crowdsourcing method enables the evaluation of the perceived functionality (i.e., perception of how likely design sketches will perform a given function) of computer generated sketches, compared to the perceived functionality of human created sketches. As a result, the possible effects of individuals' age, gender, educational level, and domain knowledge on their perceived functionality are quantified. Moreover, the integration of computational simulation and crowdsourcing methods allows for the comparison of the functional characteristics of

the sketches against their perceived functionality. In this work, the term '*sketch'* is used to mean a low-fidelity, rough 2D drawing representation of an idea with no shading or annotations.

3. RESEARCH QUESTIONS

This work aims to test the following hypotheses and address research questions (RQ):

RQ1: Do individuals' gender, age, educational level, or domain knowledge affect their perceived functionality of 2D computer and human generated sketches?

RQ2: Do individuals' gender, age, educational level or domain knowledge affect their bias towards the perceived functionality of computer or human generated sketches (i.e., labeling effect)*?*

RQ3*:* Does the functional evaluation of computer simulation correlate to humans' perceived functionality of 2D human and computer generated sketches?

The authors hypothesize that (h_1): individuals' perceived functionality of 2D computer and human generated sketches is correlated with their age, gender, educational level, and domain knowledge. The authors hypothesize that the perceived functionality of male raters is different from those of female raters. In addition, they hypothesize that raters' perceived functionality will be positively corrected to their age, educational level, and domain knowledge. These hypotheses are grounded in research that reveals that individuals' demographic characteristics and domain knowledge level can relate to their decision-making process, technology adaptation, and evaluation of

design ideas [9,11,15,42] (see section 2.3). Testing this hypothesis will allow the authors to address ***RQ1***. The hypothesis can be mathematically expressed as:

For,

$$PF = \beta_0 + \beta_1 (G) + \beta_2 (Age) + \beta_3 (Gender) + \beta_4 (EduL) + \beta_5 (Dk) + \varepsilon \quad (1)$$

(h_1) h_o: $\beta_i = 0$ vs. h_a: $\beta_i \neq 0 \ for \ i \in \{1-5\}$

Where,

- PF is the individual's perceived functionality of 2D sketches.
- β_1 is the coefficient terms for the categorical variable for either computer generated or human generated.
- β_2 is the coefficient terms for the variable of the individual's age.
- β_3 is the coefficient terms for the variable of the individual's gender.
- β_4 is the coefficient terms for the variable of the individual's educational level.
- β_5 is the coefficient terms for the variable of the individual's domain knowledge.

Moreover, following ***RQ2***, the authors hypothesize that (h_2): individuals' bias towards the perceived functionality of 2D sketches is correlated with their age, gender, educational level, and domain knowledge. That is, these factors will confound the effects of explicitly presenting the 2D sketches as computer generated or human created on the individual's perceived functionality (i.e., with a label as in Fig. 1). The authors hypothesize that raters' bias towards the perceived functionality of the sketches will differ based on their gender, age, educational level, and domain knowledge. This hypothesis is grounded in research that reveals that decision-makers' biases are

correlated with their demographic characteristics and experience level [10,40,41], and it is expressed as:

For,

$$PF_* = \beta_0 + \beta_1 (G) + \beta_2 (\overline{PF}) + \beta_3 (Age) + \beta_4 (Gender) + \beta_5 (EduL) + \beta_6 (Dk) + \varepsilon \quad (2)$$

(h_2) h_o: $\beta_i = 0$ vs. h_a: $\beta_i \neq 0 \ for \ i \in \{1-6\}$

Where,

- PF_* is the individual's perceived functionality of 2D sketches explicitly presented as either computer or human generated (i.e., with a label).
- β_1 is the coefficient terms for the categorical variable for either computer generated or human generated.
- β_2 is the coefficient terms for the average perceived functionality of 2D sketches presented without labels.
- β_3 is the coefficient terms for the variable of the individual's age.
- β_4 is the coefficient terms for the variable of the individual's gender.
- β_5 is the coefficient terms for the variable of the individual's educational level.
- β_6 is the coefficient terms for the variable of the individual's domain knowledge.

Finally, the authors hypothesize that (h_3): individuals' perceived functionality of 2D sketches is positively correlated with the functional evaluation of a computer simulation of the same sketches. This hypothesis is motivated by studies that indicate the benefits of using human raters to evaluate and select crowdsourced ideas [26,48]. Testing this hypothesis will enable the authors to address ***RQ3***. The hypothesis is expressed as:

(h_3) ho: $\rho_{\overline{PF},CS}^{g} = 0$ vs. ha: $\rho_{\overline{PF},CS}^{g} > 0 \; \forall \; g \in \{\text{computer generated, human generated}\}$

Where,

- $\overline{PF}$ is the average perceived functionality of 2D sketches.
- CS is the computer simulation's evaluation of the 2D sketches functionality.

4. CASE STUDY

To address the previous research questions and test the hypotheses, a case study in which 2D boat sketches generated by humans and a deep generative model were presented to raters recruited via a crowdsourcing platform and evaluated using a physics computer simulation.

4.1 Dataset of 2D sketches

For this case study, the *Quick, Draw!* dataset was utilized [49]. This dataset was acquired by Google via the *Quick, Draw!* game. In this game, individuals are asked to draw a specific object within 20 seconds (e.g., "*draw a boat in under 20 seconds*"). For this case study, a total of 132,270 human created boat sketches were used as a training dataset for the *Sketch-RNN* algorithm [19]. The model generated by the *Sketch-RNN* algorithm (see Ha and Eck's [19]) was used to generate 250 new boat sketches. From these 2D boat sketch datasets, 50 computer and 50 human sketches were randomly selected for evaluation. Figure 1 show some of the human and computer generated boat sketches used.

4.2 Crowdsourcing

In this work, Amazon Mechanical Turk (AMT) was used as the crowdsourcing platform to recruit raters. AMT has been previously used to evaluate the output of deep generative models [3,4]. Moreover, AMT has established itself as a valuable tool for behavioral research since studies have found no significant differences in the response consistency between internet users and laboratory participants [50,51]. Compared to other crowdsourcing platforms, AMT provides the benefits of (i) low cost, (ii) large rater pool access, and (iii) large rater pool diversity [51]. In this work, a total of 1,187 raters were recruited to evaluate a set of boat sketches, which expand the number of participants from the previous study by 204 individuals [7]. The raters were compensated $0.20 for their participation in the experiment. Only raters with a 90% satisfaction rate were allowed to participate in this experiment. Similarly, participants were only allowed to take the questionnaire once. Other quality assurances were set in place, which are explained in the following section.

4.3 Questionnaire

For this work, a between-subject experiment was implemented to test the effect that labeling the sketches as either human or computer generated had on participants' response, and disentangle this effect from any possible confirmation bias (e.g., individual rate the sketches based on his/her previous response). Once the participants consented to be part of the experiment, they were randomly assigned to one of the 25 conditions of the questionnaire. Each condition contained questions regarding a unique set of eight different 2D boat sketches. Each set of images was composed of: (i) 2 human

and (ii) 2 computer generated sketches without a label, as well as (iii) 2 human and (iv) 2 computer generated sketches with a label. At the beginning of the experiment, participants completed a short questionnaire regarding their age, gender, educational level, and physics knowledge (i.e., domain knowledge). Participants' physics knowledge was assessed via two questions (i.e., (i) *How experienced are you with the law of physics that allow boats to float on water?*, based on a 7-point Likert scale; (ii) *Please select the law of physics that explains why boats float?,* with choices: *the law of quantum mechanics, the law of buoyancy, first law of thermodynamics, none of the above, do not know*). Subsequently, participants were presented with instructions on how to complete the assessment of the 2D boat sketches, as shown in Fig. 2. For quality control purposes, the response of participants who spent less than 10 seconds on the instruction page was not considered for analysis since it is assumed that they did not read the instructions carefully. After the instruction page, participants were introduced to the five questions shown in Table 2, similar to [7]. A 7-point Likert scale was used for questions Q1, Q2, Q4, and Q5. Along with the demographics and physics knowledge assessment, questions *Q1* and *Q2* allow the authors to address the research question ***RQ1***; while questions *Q4* and *Q5* address the research question ***RQ2***. Question Q3 was used in the previous study of this work to test the capability of the participants to accurately distinguish between the human and computer generated sketches [7].

On questions Q4 and Q5, participants were shown 2 human generated and 2 computer generated boat sketches with their respective labels as shown in Fig. 1. While for questions Q1 and Q2, a different set of 2 human and 2 computer generated sketches

without labels were presented. For all the questions, the sketches were presented in a random order. Furthermore, question Q3, implemented an additional image for quality control purposes. Participants who did not correctly answer this control question were excluded from the analysis.

4.4 Computer simulation

To evaluate the functional characteristic and physical properties of the 2D boat sketches presented to the human raters, computer simulation similar to the one employed in [36] was used in this work. The simulation was implemented in Unity [52]. Unity has several desirable characteristics suitable for physics simulations. For example, it has a robust native physics engine and can support custom physics packages. Because of these characteristics, researchers have used Unity to perform physics simulations not only for validation purposes [36] but also for educational purposes [53,54]. Figure 3 shows the computer simulation environment in Unity.

The objective of the simulation environment was to evaluate the capability of the boat sketches to perform their intended function. To achieve this, two different scores (i.e., *Speed score* and *Float score*) were calculated for each of the boat sketches, similar to [36]. The *Speed score* was calculated based on the time each boat took to reach the objective (see Fig. 3). The upper limit of the *Speed score* was set to 10. A *Speed score* closer to 10 means that a boat reached the objective in less time (i.e., faster), compared to a boat that had a score less than 10. In the simulation environment, the same constant propelling force with equal magnitude and direction was applied to all of the boats evaluated. The direction of the force was chosen in order

to move the boat from left to right towards the objective. Once a boat reached the objective, the simulation ended. The simulation environment was designed to resemble the 2D environment presented to the participants on the instruction page of the questionnaire (see Fig.2). The *Float score* was calculated based on the average distance throughout the simulation between the water level of the environment without the boat, and the boat's lowest point while in the water (see Fig. 3). This score helped account for the differences in time each boat took to reach the objective. Moreover, the *Float score* was given a range between 1 and 0. A *Float score* close to 1 means that the boat's lowest point was on average, closer to the water level than the *Sink line* throughout the simulation. If a boat's lowest point hit the *Sink line*, it was assumed that the boat sunk; hence, a *Float score* of 0 was given, and the simulation was ended. The range of values for the *Float* and *Speed* scores were selected to facilitate the design of the simulation environment.

For the 2D boat sketches to interact with the simulated environment, collision detection was applied along the line segments of the boat sketches, as [36]. Also, for simulation purposes, it was assumed that the line segments of the boat sketches were all made out of the same material, which had a constant density. Consequently, the mass of a boat was proportional to the number and length of its line segments. Hence, the net acceleration of a boat was inversely proportional to its mass and directly proportional to the magnitude of the net force applied to it, following Newton's second law. For simulation purposes, it was assumed that the only forces that interacted with

the boats were the force of gravity, the drag force from the water particles, and the constant propelling force applied to the boats.

5. RESULTS AND DISCUSSION

After filtering participants based on their response to the quality control question and time spent reading the instructions, the data of only 748 participants (48.1% females) are used in this work. On average, the participants spent 430.4 seconds (SD= 328.8 secs) to complete the questionnaire. Table 3 shows the summary statistics for the participants' response to the demographics and physics questions, while Table 4 shows the summary statistics for questions Q1, Q2, Q4, and Q5. Moreover, Fig. 1 shows the boat sketches that were perceived as the most functional (leftmost column), the least functional (rightmost column), and having average functionality (center column). In this work, an alpha level of 0.05 is used to test the statistical significance of the results.

5.1 Reliability and validity

The inter-rater reliability of participants' responses was assessed via Cronbach's alpha. The reliability of the raters' response (i.e., Q1-Q5) on each of the 25 different conditions of the questionnaire was calculated. The results indicate that on average, participants' responses had a Cronbach's alpha of 0.813 (SD= 0.048). These results reveal that participants' responses were more consistent when evaluating certain sets of images (range= [0.708-0.894]). Overall, the Cronbach's alpha indicates acceptable

inter-rater reliability (>0.7) [55]. This reveals that in general, participants showed consensus in their responses.

Moreover, an analysis was performed to test the validity and fidelity of the Unity physics simulation used. An experiment was designed to test the effects that a boat's overall density had on its *Speed score* and *Float score*. For this experiment, the same boat design was implemented, and the independent variables were the boat's overall mass and dimensions. Both the mass and dimension variables were set to two levels (i.e., high and low) (e.g., 2x2 factorial design). The high values were set to two times that of the low values. The scores indicate that increasing the mass of a boat, while maintaining its dimensions constant (i.e., increased density), negatively impacted the ability of the boat to float and move (Δ*Float score*: Dimension-low= -0.04, Dimension-high= -0.02; Δ*Speed score*: Dimension-low= -1.77, Dimension-high= -0.75). In contrast, increasing the dimensions of a boat, while maintaining its mass constant (i.e., reduced density), positively impacted the ability of the boat to float and move (Δ*Float score*: Mass-low= 0.02, Mass-high= 0.44; Δ*Speed score*: Mass-low= 0.39, Mass-high= 1.40). These simulation results are in line with the law of buoyancy, supporting the ability of the Unity simulation used in this work to recreate the physics of boats floating and moving on water.

5.2 RQ1: Perceived functionality of sketches

To test the hypothesis (h_1) and explore the possible confounding effects of participants' age, gender, educational level, and domain knowledge on their perceived functionality, a linear regression analysis was performed. Two models were fitted

following Eq.1, one using the participants' response of Q1 as the dependent variable, and a second using Q2. In both models, the variable of participants' age and response on the first physics questions were considered to be on an interval scale (i.e., age: [18-76], first physics questions: [1-7]), while the remaining variables were considered to be on a nominal scale (i.e., categorical variables, see Table 5 notes). For these categorical variables, the first level was used as a reference category (see Table 5 notes). Moreover, the data of participants that prefer not to report their gender identity or selected the choice of "Other" (i.e., *Prefer not to say*: 4, *Other*: 1, see Table 3) were not analyzed to reduce imbalance between the levels of the factor of *Gender* (see Table 5 notes). Table 5 shows the summary statistics and the estimates for the standardized coefficients for the regression model using Q1 and Q2 as the dependent variable. The results indicate that on average, the human generated sketches were perceived as less functional (i.e., less likely to float and move) than the computer generated sketches (Q1: β_1= -0.396, $t_{(1)}$= -*11.11*, *p-value*<0.001; Q2: β_1= -0.385, $t_{(1)}$= *-10.74*, *p-value*<0.001). Additionally, the results reveal that participants' age (Q1: β_2=0.01, $t_{(1)}$ = *5.967*, *p-value*<0.001; Q2: β_2=0.004, $t_{(1)}$= *2.935*, *p-value*=<0.001) and their response on the first physics question (Q1: $\beta_{5.1}$=0.058, $t_{(1)}$= *5.188*, *p-value*<0.001; Q2: $\beta_{5.1}$=0.048, $t_{(1)}$= *4.326*, *p-value*<0.001) were positively correlated with the perceived capability of the boat sketches to float and move. Nonetheless, the models were able to explain only 5.9% ($F_{(9,2982)}$=23.35, p-value<0.001) of the variability in Q1, and 4.6% ($F_{(9,2982)}$=17.81, p-value<0.001) of the variability in Q2, which are small effects according to [56].

The previous results provide enough evidence to reject the null hypothesis (h_1) since participants' perceived functionality of the boat sketches was related to their age and domain knowledge, and not only to the source of the sketches. Even though the findings of this work indicate that the perceived functionality of computer generated and human created sketched ideas will depend on the individual demographic characteristics and domain knowledge, the result supports the capability of deep generative design tools to generate ideas that are perceived as functional. These results indicate that deep generative design tools could potentially assist in creative tasks such as ideation, and that individuals' demographic and domain knowledge relate to their perceived functionality of 2D computer and human generated sketches (***RQ1***).

5.3 RQ2: Perceived functionality bias

To test the hypothesis (h_2) and explore the possible confounding effects that age, gender, educational level, and domain knowledge have on participants' bias towards the perceived functionality of the boat sketches, a linear regression analysis was performed. Two models were fitted following Eq.2, one using participants' response on Q4 as the dependent variable, and a second using Q5. In the first model, the average response on Q1 was used as an independent variable, while for the second model the average response on Q2 was used. Because this work implemented a between-subject design, using these variables as independent variables allow the authors to explore how participants' perceived functionality of the sketches presented with a label differed from the average perceived functionality of the same sketches presented without a label.

Similarly, only the data from participants that selected the gender identity of female or male was used in the analysis (see Table 6 notes).

Table 6 shows the summary statistics and the estimates for the standardized coefficients for the regression models using Q4 and Q5 as the dependent variable. The results indicate that on average, the sketches were perceived as less functional when explicitly presented with a label (Q4: β_0= -2.291, $t_{(1)}$= -14.51, *p-value*<0.001; Q5: β_0= -2.184 , $t_{(1)}$= -13.19, *p-value*<0.001). Nonetheless, the human generated sketches were perceived as less functional than the computer generated when explicitly presented with a label (Q4: β_1= -0.203, $t_{(1)}$= -5.341, *p-value*<0.001; Q5: β_1= -0.226, $t_{(1)}$= -5.875, *p-value*<0.001). In addition, the results indicate that participants' age (Q4: β_3= 0.008, $t_{(1)}$= 5.567, *p-value*<0.001; Q5: β_3= 0.007, $t_{(1)}$= 4.557, *p-value*<0.001), and participants' experience with the law of physics that explains why boats float (Q4: $\beta_{6.1}$= 0.049, $t_{(1)}$= 4.775, *p-value*<0.001; Q5: $\beta_{6.1}$= 0.066, $t_{(1)}$= 6.238, *p-value*<0.001) were correlated with an increased perceived functionality of the sketches. Similarly, participants who did not correctly answer the second physics questions, on average perceived the boat sketches (i.e., computer and human generated) as more likely to move when a label was present (Q5: $\beta_{6.2}$= 0.099, $t_{(1)}$= 2.535, *p-value*= 0.011). Finally, the results indicate that participants with higher educational level (i.e., *Degree: 3* and *Degree: 4*, Table 6) perceived the boat sketches as less likely to float when presented with a label. Nonetheless, the models were able to explain only 1.7% ($F_{(10, 2981)}$= 65.64, *p-value*<0.001) of the variability in Q4, and 1.4% ($F_{(10, 2981)}$= *53.62*, *p-value*<0.001) of the variability in Q5, which are small effects according to [56]. These results indicate that participants' perceived functionality of the

sketches was negatively affected by explicitly presenting them with a label. However, this effect was more prominent on the human created sketches. Also, the results indicate that age, educational level, and domain knowledge confounded the effects that presenting the sketches with labels have on participants' perceived functionality. These findings provide enough evidence to reject the null hypothesis (h_2).

While previous studies have shown that individuals' demographic characteristic and experience level, can influence their decision making [9,46,47], they did not explore the possible bias decision-makers may have towards the functionality of computer and human generated sketches nor the confounding effects of their demographic characteristics and domain knowledge. In this work, the results reveal that the perceived functionality of sketches was negatively biased by the fact that they were explicitly presented as either computer or human generated (i.e., with a label). However, this bias was more significant for the human generated sketches. Moreover, individuals' age, educational level, and domain knowledge influenced their biases towards the perceived functionality of computer and human generated sketches (***RQ2***). This indicates that during the evaluation and screening process of new design sketches, individuals' perceived functionality of sketches may be subject to *Automation* bias.

5.4 RQ3: Correlation between human and simulation functionality evaluation

To test the hypothesis (h_3), the computer simulation introduced in section 4.4 was used to evaluate the capability of the boat sketches to float and move. A Pearson product-moment correlation coefficient was computed to assess the relationship

between the evaluations of the computer simulation and the average perceived functionality of the boat sketches. The results indicate that the simulation *Float score* was positively correlated with participants' response in Q1 (ρ= 0.3, *p-value*= 0.002). Similarly, the *Speed score* had a positive correlation with participants' response in Q2 (ρ= 0.5, *p-value*<0.001). In addition, the results indicate that both the *Speed score* and *Float score* had a strong positive correlation (ρ= 0.83, *p-value*<0.001). Similarly, participants' response between Q1 and Q2 (ρ= 0.82*p-value*<0.001), and Q4 and Q5 (ρ= 0.91, *p-value*<0.001), were strongly correlated. The correlation of the simulation scores can be explained by the fact that a boat that has less buoyancy will encounter more resistance from the water particles due to its larger contact area (i.e., drag or fluid friction). Hence, under a constant force, the magnitude of a boat's acceleration will be less than a boat that has more buoyancy. The correlation between participants' responses reveals that when evaluating the sketches, they may be considering this relationship as well. Moreover, an independent-samples one-tailed t-test was conducted to compare the *Float score,* and *Speed score* between the computer and human generated sketches. The t-test results indicate that, on average, the *Float score* ($t_{(98)}$ = 2.44, *p-value*= 0.02) and *Speed score* ($t_{(98)}$= 2.58, *p-value*= 0.01) of the computer generated boat sketches (*Float*: M= 0.905, SD= 0.017;*Speed*: M=3.427, SD=0.485) were greater than the human created sketches used to train the deep generative model (*Float*: M= 0.896, SD= 0.022;*Speed*: M=3.136, SD=0.636).

The simulation results provide enough evidence to reject the null hypothesis (h_3), indicating that participants' perceived functionality of the boat sketches were similar to

the functional evaluation given by the computer simulation. These findings help address ***RQ3***, indicating that the functional evaluation of computer simulation correlates to humans' perceived functionality of 2D human and computer generated sketches. These results support the value of using human raters to evaluate the functionality of 2D sketched ideas, which are in line with previous studies that have shown the benefit in using expert raters and crowds to evaluate new design ideas [26,37]. Moreover, the simulation results support the results in section 5.2, indicating that the boat sketches generated by the deep generative model were more likely to float and move than the human created sketches used to train the model. These findings support the capability of deep generative models to not only generate new sketched ideas but sketches that are functional.

6. CONCLUSIONS AND FUTURE WORKS

Recent advancements in technology have allowed designers to implement computational tools to automatically generate large pools of new design ideas. Nonetheless, an idea needs to meet its intended functionality and be useful in order to be considered creative. Therefore, if computational tools are to co-create ideas and solutions alongside designers, their capability to produce not only novel, but functional ideas, needs to be explored. Furthermore, the ability to generate creative ideas is an insufficient condition for innovation because decision-makers need to not only generate, but also select creative ideas for innovation to occur. However, several studies indicate that demographic characteristics and experience level of decision-makers can influence and bias their selection of ideas. As designers are increasingly integrating

computational tools to assist in the design process, their possible bias towards computer generated and human created ideas, as well as the potential confounding effects of individuals' demographic characteristics and domain knowledge, need to be explored. In order to fill this knowledge gap, this work implemented a crowdsourcing method to explore the perceived functional characteristics of 2D design sketches created by humans and 2D design sketches generated by a deep learning generative model (i.e., computer generated). This work also explored the underlying influence of individuals' demographic characteristics and domain knowledge on their perceived functionality of sketches. Finally, a computer simulation method was implemented to test the capability of the sketches to perform their intended function. The integration of computational simulation and crowdsourcing methods allows for the comparison of the functional characteristics of the sketches against their perceived functionality. In summary, the results of this work indicate that:

1. Computer generated sketches were perceived as more functional than the human generated sketches. Additionally, participants' age and domain knowledge were positively correlated with their evaluations.
2. The perceived functionality of sketches was negatively affected by explicitly presenting them with a label. However, this effect was more significant for the human created sketches, and was confounded by participants' age, educational level, and domain knowledge.
3. Participants' perceived functionality of sketches was positively correlated with the functional evaluation of the computer simulation.

4. Sketches generated by the deep generative model were, on average, more likely to float and move, compare to the human created sketches used to train the model.

The results reveal that participants perceived the 2D boat sketches generated by a deep generative model (i.e., computer generated) as more likely to float and move than the human created sketches used to train the model. Also, the results indicate that this perception was correlated with participants' age and experience with the law of physics that explains why boats float. Furthermore, the computer simulation results also indicate that the 2D computer generated boat sketches were more likely to float and move, compared to the human created sketches. These findings support the capability of deep generative models to generate functional sketched ideas. As deep generative design tools become more efficient at creating novel and functional ideas, researchers argue that they will foster designers' creativity and help in creative tasks [2,20]. Also, the results reveal that participants' perceived functionality of the boat sketches were similar to the functional evaluation given by the computer simulation. These findings support the value of using human raters to evaluate the functionality of rough 2D design sketches.

The results of this work also revealed that participants' perceived functionality of sketches was negatively biased by explicitly presenting them as either computer or human generated (i.e., with a label). This effect was correlated with participants' age, educational level, and domain knowledge. However, this bias effect was more significant for the human generated sketches than for the computer generated sketches. The human-computer interaction community has recognized that *Automation bias* can

affect individuals' perception of automated system's capabilities [44,47]. The results of this work reveal that participants were more negatively biased towards human created sketches. This indicates that during the evaluation and screening process of new design sketches, individuals' perceived functionality of sketches may be subject to *Automation* bias.

While this work provides evidence that supports the capabilities of deep generative design tools and their potential to assist designers in creative tasks, several limitations exist. For example, although the results indicate that participants' perceived functionality of the computer generated sketches was greater than the human created sketches, the practical significance of these differences (i.e., ΔQ1= 0.72 or 10.28%, ΔQ2= 0.67 or 9.57%) needs to be explored. Moreover, the effect of presenting the sketches with and without labels on participants' perceived functionality cannot be disentangled from a possible order or fatigue effect. This is because all the questions that contained sketches with labels were presented after the questions that contained sketches without labels. In addition, while studies have found no significant differences in the response consistency between internet users and laboratory participants, the crowdsourcing method and experimental protocol implemented in this work (e.g., sequence of the questionnaires) could have impacted the validity of the responses. Future works should implement other methods and experimental designs to disentangle possible order or fatigue effects. In addition, while this work only used low-fidelity, rough 2D boat sketches (since these are typically the primary communication source of ideas in the early stages of the design process [6]), future work should explore the effects that the

fidelity of the sketches has on individuals' perceived functionality. Previous studies have shown that sketch fidelity and design complexity (e.g., task complexity) can impact evaluators' responses [34,42]. Moreover, the visual features and characteristics of the sketches should be further explored to understand why computer generated boat sketches were perceived and were more likely to float and move than the human created sketches.

ACKNOWLEDGMENT

This research is funded in part by DARPA HR0011-18-2-0008 and NSF NRI # 1527148. Any opinions, findings, or conclusions are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] Chandrasegaran, S. K., Ramani, K., Sriram, R. D., Horváth, I., Bernard, A., Harik, R. F., and Gao, W., 2013, "The evolution, challenges, and future of knowledge representation in product design systems," Comput. Des., **45**(2), pp. 204–228.

[2] Liapis, A., Yannakakis, G. N., Alexopoulos, C., and Lopes, P., 2016, "Can Computers Foster Human User's Creativity? Theory and Practice of Mixed-Initiative Co-Creativity," Digit. Cult. Educ., **8**(2), pp. 136–153.

[3] Burnap, A., Lui, Y., Pan, Y., Lee, H., Gonzalez, R., and Papalambors, P., 2016, "Estimating and Exploring the Product Form Design Space Using Deep Generative Models," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf.,North Carolina, USA, pp. 1–13.

[4] Dering, M. L., and Tucker, C. S., 2017, "Generative Adversarial Networks for Increasing the Veracity of Big Data," IEEE Inter. Conf. on Big Data (BIGDATA), Boston MA, USA, pp. 2513–2520.

[5] Boden, M. A., 2004, "The creative mind: Myths and mechanisms, " Second edition, Routledge.

[6] Kazi, R. H., Grossman, T., Cheong, H., Hashemi, A., and Fitzmaurice, G., 2017, "DreamSketch: Early Stage 3D Design Explorations with Sketching and Generative Design," Proc. 30th Annu. ACM Symp. User Interface Softw. Technol., Quebec City,

Canada, pp. 401–414.

[7] Lopez, C. E., & Tucker, C. S., 2018, "Human validation of computer vs human generated desing sketches," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Quebec City, Canada.

[8] Rietzschel, E. F., Nijstad, B. A., and Stroebe, W., 2006, "Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection," J. Exp. Soc. Psychol., **42**(2), pp. 244–251.

[9] Toh, C. A., Strohmetz, A. A., and Miller, S. R., 2016, "The Effects of Gender and Idea Goodness on Ownership Bias in Engineering Design Education," J. Mech. Des., **138**(10), p. 101105.

[10] Toh, C. A., Patel, A. H., Strohmetz, A. A., and Miller, S. R., 2015, "My Idea Is Best! Ownership Bias and its Influence on Engineering Concept Selection," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Boston MA, USA, pp. 1–10.

[11] Zheng, X., and Miller, S. R., 2017, "Risky business: The driving factors of creative risk taking," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Clevelan OH, USA.

[12] Thomson, M. E., Önkal, D., Avcioğlu, A., and Goodwin, P., 2004, "Aviation risk perception: A comparison between experts and novices," Risk Anal., **24**(6), pp. 1585–1595.

[13] Arning, K., and Ziefle, M., 2007, "Understanding age differences in PDA acceptance and performance," Comput. Human Behav., **23**(6), pp. 2904–2927.

[14] Wang, Y.-S., Wu, M.-C., and Wang, H.-Y., 2009, "Investigating the determinants and age and gender differences in the acceptance of mobile learning," Br. J. Educ. Technol., **40**(1), pp. 92–118.

[15] Venkatesh, V., Morris, M. G., Davis, G. B., and Davis, F. D., 2003, "User Acceptance of Information Technology: Toward a unified view," MIS Q., **27**(3), pp. 425–478.

[16] Orsborn, S., Cagan, J., and Boatwright, P., 2009, "Quantifying Aesthetic Form Preference in a Utility Function," J. Mech. Des., **131**(6), p. 061001.

[17] Reid, T. N., Gonzalez, R. D., and Papalambros, P. Y., 2010, "Quantification of Perceived Environmental Friendliness for Vehicle Silhouette Design," J. Mech. Des., **132**(10), p. 101010.

[18] Schmidhuber, J., 2015, "Deep Learning in neural networks: An overview," Neural Networks, **61**, pp. 85–117.

[19] Ha, D., and Eck, D., 2017, "A neural representation of sketch drawings.," Preprint arXiv:1704.03477.

[20] Chen, Y., Tu, S., Yi, Y., and Xu, L., 2017, "Sketch-pix2seq: a Model to Generate Sketches of Multiple Categories.," Preprint arXiv1709.04121.

[21] Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L., 2017, "Learning Representations and Generative Models for 3D Point Clouds," Preprint arXiv1707.02392.

[22] Boden, M., 2001, "A guide to recurrent neural networks and backpropagation," Electr. Eng., (2), pp. 1–10.

[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y., 2014, "Generative Adversarial Nets," Proc. Adv. Neural Inf. Process. Syst. 27 (NIPS), Montreal, Canada, pp. 2672–2680.

[24] Dosovitskiy, A., Springenberg, J. T., Tatarchenko, M., and Brox, T., 2017, "Learning to Generate Chairs, Tables and Cars with Convolutional Networks," IEEE Trans. Pattern Anal. Mach. Intell., **39**(4), pp. 692–705.

[25] Theis, L., Oord, A. V. D., & Bethge, M., 2015, "A note on the evaluation of generative models," Preprint arXiv1511.01844.

[26] Poetz, M. K., and Schreier, M., 2012, "The value of crowdsourcing: Can users really compete with professionals in generating new product ideas?," J. Prod. Innov. Manag., **29**(2), pp. 245–256.

[27] Wang, G. G., and Shan, S., 2007, "Review of Metamodeling Techniques in Support of Engineering Design Optimization," J. Mech. Des., **129**(4), p. 370.

[28] Dering, M., and Tucker, C., 2017, "A Convolutional Neural Network Model for Predicting a Product's Function, Given Its Form," J. Mech. Des., **139**(11), pp. 1–14.

[29] Buxton, B., 2010, "Sketching User Experiences: Getting the Design Right and the Right Design," Morgan Kaufmann.

[30] Goel, V., 1997, "Sketches of thought," Des. Stud., **18**(1), pp. 129–130.

[31] Rodgers, P. A., Green, G., and McGown, A., 2000, "Using concept sketches to track design progress," Des. Stud., **21**(5), pp. 451–464.

[32] Van Der Lugt, R., 2005, "How sketching can affect the idea generation process in design group meetings," Des. Stud., **26**(2), pp. 101–112.

[33] Yang, M. C., 2009, "Observations on concept generation and sketching in engineering design," Res. Eng. Des., **20**(1), pp. 1–11.

[34] Macomber, B., and Yang, M. C., 2011, "The role of sketch finish and style in user responses to early stage design concepts," ASME Int. Design Eng and Technical Conf., Washington DC, USA, pp. 567–576.

[35] Häggman, A., Tsai, G., Elsen, C., Honda, T., and Yang, M. C., 2015, "Connections Between the Design Tool, Design Attributes, and User Preferences in Early Stage Design," J. Mech. Des., **137**(7), p. 071101.

[36] Cunningham, J., and Tucker, C. S., 2018, "A Valination Neural Network (VNN) metamodel for predicting the performane of deep generative desings," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Quebec City.

[37] Ren, Y., Burnap, A., Papalambros, P., 2013, "Quantification of perceptual design attributes using a crowd," Proc. of the 19th Int. Conf. on Eng. Design, Seoul, Korea, pp. 19–22.

[38] Toh, C. A., Miele, L. M., and Miller, S. R., 2016, "Which One Should I Pick ? Concept Selection in Engineering Design Industry," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Boston MA, USA, pp. 1–10.

[39] Cox, D., and Cox, A. D., 2002, "Beyond first impressions: The effects of repeated exposure on consumer liking of visually complex and simple product designs," J. Acad. Mark. Sci., **30**(2), pp. 119–130.

[40] Mueller, J. S., Melwani, S., and Goncalo, J. A., 2012, "The bias against creativity: Why people desire but reject creative ideas," Psychol. Sci., **23**(1), pp. 13–17.

[41] Toh, C. A., and Miller, S. R., 2014, "The role of individual risk attitudes on the selection of creative concepts in engineering design," Proc. ASME Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf., Buffalo NY, USA, pp. 1–10.

[42] Burnap, A., Gerth, R., Gonzalez, R., and Papalambros, P. Y., 2017, "Identifying experts in the crowd for evaluation of engineering designs," J. Eng. Des., **28**(5), pp. 317–337.

[43] W, I. J., Nap, H. H., De Kort, Y., and Poels, K., 2007, "Digital game design for elderly users," Proc. 2007 Conf. Futur. Play. Futur. Play '07, pp. 17–22.

[44] Parasuraman, R., and Manzey, D. H., 2010, "Complacency and bias in human use of automation: An attentional integration," Hum. Factors, **52**(3), pp. 381–410.

[45] Mosier, K. L., and Skitka, L. J., 1996, "Human decision makers and automated decision aids: Made for each other?," Automation and Human Performace, Erlbaum.

[46] Lee, J. D., and See, K. A., 2004, "Trust in Automation: Designing for Appropriate Reliance," Hum. Factors J. Hum. Factors Ergon. Soc., **46**(1), pp. 50–80.

[47] Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A., 2002, "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," Hum. Factors J. Hum. Factors Ergon. Soc., **44**(1), pp. 79–94.

[48] Le, Q., and Panchal, J. H., 2011, "Modeling the Effect of Product Architecture on Mass-Collaborative Processes," J. Comput. Inf. Sci. Eng., **11**(1), p. 011003.

[49] Jongejan, J., Rowley, H., Kawashima, T., Kim, J., and Fox-Gieg., N., 2016, "The Quick, Draw! - A.I. Experiment.," https://quickdraw.withgoogle.com/.

[50] Buchanan, T., 2000, "Psychological Experiments on the Internet," Academic Press.

[51] Mason, W., and Suri, S., 2012, "Conducting behavioral research on Amazon's Mechanical Turk," Behav. Res. Methods, **44**(1), pp. 1–23.

[52] Unity, 2017, "Unity - Game Engine," https://www.unity3d.com.

[53] González, J. D., Escobar, J. H., Sánchez, H., De La Hoz, J., and Beltrán, J. R., 2017, "2D and 3D virtual interactive laboratories of physics on Unity platform," Journal of Physics: Conference Series, **935**(1), p. 012069.

[54] Ballu, A., Yan, X., Blanchard, A., Clet, T., Mouton, S., and Niandou, H., 2016, "Virtual Metrology Laboratory for e-Learning," Procedia CIRP, **43**, pp. 148–153.

[55] Cortina, J. M., 1993, "What is coefficient alpha? An examination of theory and applications.," J. Appl. Psychol., **78**(1), pp. 98–104.

[56] Cohen, J., 1988, "Statistical power analysis for the behavioral sciences,"Second Edition, Routledge.

Table Caption List

Table 1. Summary of existing studies on deep generative model evaluation

Reference	*Object Classification evaluation*	*Functionality evaluation*	*Crowdsourcing method*	*Effects of raters' attributes**
[6][24]	X			
[3][4] [20][37]	X		X	
[28][36]		X		
[7]	X	X	X	
This work		X	X	X

**Effects of raters' demographic characteristics and domain knowledge on their evaluation*

Table 2. Questions presented to participants

Q1: Please evaluate the following boat sketches based on how well they will **float** in the 2D environment shown below.
Q2: Please evaluate the following boat sketches based on how well they will **move** from point A (left) to point B (right) when a force is applied in the 2D environment as shown below.
Q3: Please classify the following sketches as *human-generated* (drawn by a person) or *computer-generated* (drawn by a computer).
Q4: Please evaluate the following computer and human generated boat sketches based on how well they will **float** in the 2D environment shown below.
Q5: Please evaluate the following computer and human generated boat sketches based on how well they will **move** from point A (left) to point B (right) when a force is applied in the 2D environment as shown below.

Table 3. Summary statistics for demographics and physics questions

		Frequency	Percentage		Frequency	Percentage
Gender				**Educational Level**		
Female		360	48.1	Less than high school degree	1	0.1
Male		383	51.2	High school graduate (high school diploma or GED)	60	8.0
Other		1	0.1	Some college but no degree	140	18.7
Prefer not to say		4	0.5	Associate degree in college (2-year)	67	9.0
				Bachelor's degree in college (4-year)	331	44.3
Age by gender				Master's degree	123	16.4
<20	Female	71		Doctoral degree	14	1.9
	Male	74		Professional degree (JD, MD)	12	1.6
	Other	1				
	Total	146	19.5	**Physics question 1 (Fig.2)**		
				1 (Do not understand how boats float)	64	8.6
21-30	Female	92		2	72	9.6
	Male	128		3	97	13.0
	Prefer not to say	2		4	143	19.1
	Total	222	29.7	5	168	22.5
				6	128	17.1
31-40	Female	82		7 (I can clearly explain why and how boats float)	76	10.2
	Male	113				
	Prefer not to say	1		**Physics question 2 (Fig.2)**		
	Total	196	26.2	The law of quantum mechanics	29	3.9
				The law of buoyancy	538	71.9
41-50	Female	60		Frist law of thermodynamics	30	4.0
	Male	38		Node of the above	54	7.2
	Prefer not to say	1		Do not know	97	13.0
	Total	99	13.2			
>51	Female	55				
	Male	30				
	Total	85	11.4			

Table 4. Summary statistics for Q1, Q2, Q4, and Q5

	Computer generated			Human generated		
	μ	***median***	***σ***	***μ***	***median***	***σ***
Q1	5.12	6	1.65	4.40	5	1.92
Q2	5.02	5	1.63	4.35	5	1.83
Q4	5.04	5	1.58	4.12	4	1.90
Q5	4.98	5	1.58	4.14	4	1.78

Notes: Responses are on a 7-point Likert scale (range [1-7])

Table 5. Summary statistics of linear regression model for Q1 and Q2

Variable	Model for Q1				Model for Q2			
	Standardized β	Std. Error	t-value	p-value	Standardized β	Std. Error	t-value	p-value
Intercept	-0.381	0.104	-3.666	**<0.001**	-0.157	0.104	-1.504	0.133
Generated: Human	-0.396	0.036	-11.11	**<0.001**	-0.385	0.036	-10.74	**<0.001**
Age	0.010	0.002	5.967	**<0.001**	0.004	0.002	2.935	**<0.001**
Gender: Male	-0.041	0.037	-1.109	0.268	-0.006	0.038	-0.156	0.876
Degree: 2	0.061	0.071	0.848	0.397	-0.003	0.072	-0.049	0.961
Degree: 3	0.008	0.067	0.117	0.907	0.042	0.068	-0.625	0.532
Degree: 4	-0.098	0.115	-0.856	0.392	0.026	0.115	0.231	0.817
Physics Q1	0.058	0.011	5.188	**<0.001**	0.048	0.011	4.326	**<0.001**
Physics Q2: 2	-0.027	0.041	-0.659	0.510	0.005	0.041	0.133	0.895

Notes: The categorical variable *Generated* had two levels (1) *Computer* and (2) *Human*. The categorical variable *Gender* had two levels (1) *Female* and (2) *Male*. The categorical variable *Degree* was grouped into four levels: (1) *Less than high school degree* plus *High school graduate*, (2) *Some college but not degree* plus *Associate degree in college*, (3) *Bachelor's degree in college* plus *Master's degree*, and (4) *Doctoral degree* plus *Professional degree*. The categorical variable *Physics Q2* was grouped into two levels: (1) *The law of buoyancy*, (2) otherwise (see Table 3). *For the categorical variables level (1) is the reference for the dummy variables*. *Significance level codes (p-values): Bold <0.05.*

Table 6. Summary statistics of linear regression model for Q4 and Q5

Variable	Model for Q4				Model for Q5			
	Standardized β	Std. Error	t-value	p-value	Standardized β	Std. Error	t-value	p-value
Intercept	-2.291	0.158	-14.51	<**0.001**	-2.184	0.165	-13.19	<**0.001**
Generated: Human	-0.203	0.038	-5.341	<**0.001**	-0.226	0.039	-5.875	<**0.001**
Q1/Q2 Avg.	0.420	0.024	17.25	<**0.001**	0.388	0.027	14.57	<**0.001**
Age	0.008	0.002	5.567	<**0.001**	0.007	0.002	4.557	<**0.001**
Gender: Male	0.002	0.035	0.051	0.959	0.052	0.036	1.459	0.145
Degree: 2	-0.117	0.067	-1.739	0.082	-0.131	0.068	-1.915	0.056
Degree: 3	-0.136	0.063	-2.150	**0.032**	-0.125	0.064	-1.947	0.052
Degree: 4	-0.236	0.108	-2.189	**0.028**	-0.147	0.109	-1.349	0.177
Physics Q1	0.049	0.010	4.775	<**0.001**	0.066	0.010	6.238	<**0.001**
Physics Q2: 2	-0.001	0.039	-0.020	0.984	0.099	0.039	2.535	**0.011**

Notes: The categorical variable *Generated* had two levels (1) *Computer* and (2) *Human*. The categorical variable *Gender* had two levels (1) *Female* and (2) *Male*. The categorical variable *Degree* was grouped into four levels: (1) *Less than high school degree* plus *High school graduate*, (2) *Some college but not degree* plus *Associate degree in college*, (3) *Bachelor's degree in college* plus *Master's degree*, and (4) *Doctoral degree* plus *Professional degree*. The categorical variable *Physics Q2* was grouped into two levels: (1) *The law of buoyancy*, (2) otherwise (see Table 3). *For the categorical variables level (1) is the reference for the dummy variables. Significance level codes (p-values): Bold <0.05.*

Figure Captions List

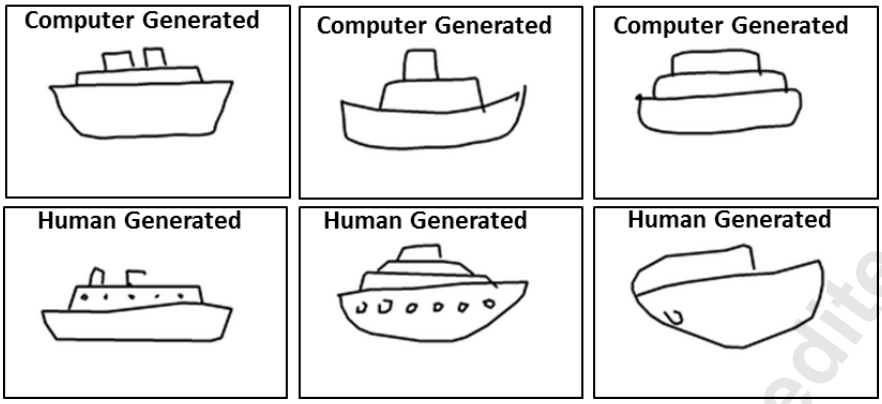

Fig. 1. Example of human and computer generated boat sketches

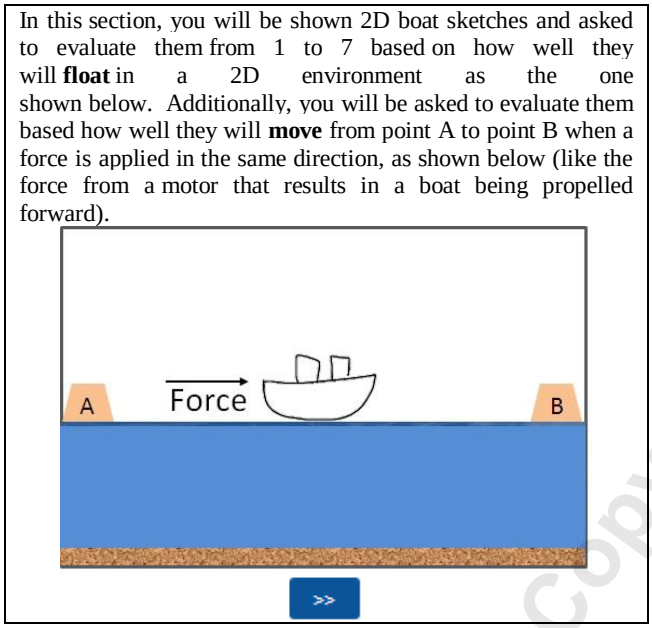

Fig. 2. Instruction page from questionnaire

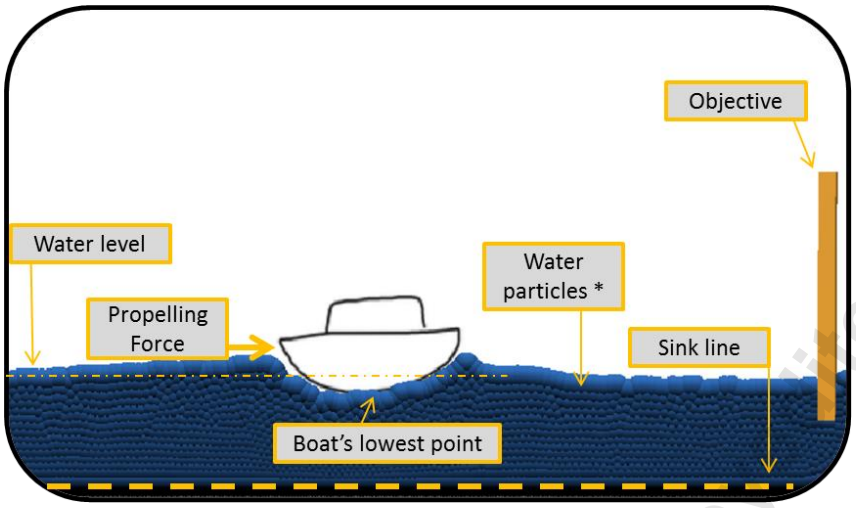

Fig. 3. Computer simulation environment in Unity

**Water particles are rendered to appear larger for visual representation.*