HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230078

A Hybrid Intelligence Approach to Training Generative Design Assistants: Partnership Between Human Experts and AI Enhanced Co-Creative Tools

Yaoli MAO [a,*,1], Janet RAFNER [b,*], Yi WANG [a], Jacob SHERSON [b,2]

[a] *Autodesk Research*
[b] *Center for Hybrid Intelligence, Department of Management, School of Business and Social Science, Aarhus University*

Abstract. The emergence of generative design (GD) has introduced a new paradigm for co-creation between human experts and AI systems. Empirical findings have shown promising outcomes such as augmented human cognition and highly creative design products. Barriers still remain that prevent individuals from perceiving and adopting AI, entering into collaboration with AI and sustaining it over time. It is even more challenging for creative design industries to adopt and trust AI where these professionals value individual style and expression, and therefore require highly personalized and specialized AI assistance. In this paper, we present a holistic hybrid intelligence (HI) approach for individual experts to train and personalize their GD assistants on the fly. Our contribution to human-AI interaction is three-fold including i) a programmable common language between human and AI to represent the expert's design goals to the generative algorithm, ii) a human-centered continual training loop to seamlessly integrate AI-training into the expert's task workflow, iii) a hybrid intelligence narrative to address the psychological willingness to spend time and effort training such a virtual assistant. This integral approach enables individuals to directly communicate design goals to AI and seeks to create a psychologically safe space for adopting, training and improving AI without the fear of job-replacement. We concertize these constructs through a newly developed Hybrid Intelligence Technology Acceptance Model (HI-TAM). We used mixed methods to empirically evaluate this approach through the lens of HI-TAM with 8 architectural professionals working individually with a GD assistant to co-create floor plan layouts of office buildings. We believe that the proposed approach enables individual professionals, even non-technical ones, to adopt and trust AI-enhanced co-creative tools.

Keywords. Communication, Co-creation, Human AI language, Partnership, Personalization, Tool adoption, Training generative AI assistants, Technology Acceptance Model

[1] *These authors contributed equally to this work.
[2] Corresponding Author: Jacob Sherson, sherson@mgmt.au.dk

1. Introduction

Generative Design (GD) has rapidly emerged as a powerful Artificial Intelligence (AI) enhanced design paradigm enabling human experts (e.g. architects) to augment their creativity and accelerate the design processes by suggesting new ideas and improving design quality in co-creation [1,2,3].

However, the adoption and integration of AI-support technologies into creative industries face significant challenges including the potential for job displacement, deskilling, concerns over the transparency and accountability of AI systems, and the need for highly personalized assistance [4]. Additionally, up to 90% of AI related projects fail to some extent in the real-world implementation stage [5] and in general only 10% of organizations are achieving significant financial benefits with AI [5]. Given the specialized skills involved in GD, introducing AI into professional workflows in the creative design field is likely to create tension.

These challenges can be divided into two areas to conquer: 1) human-centered continually evolving interaction with AI and 2) holistic development and deployment frameworks taking into account organizational aspects of the introduction of the technology at professional work settings. The former can be addressed by incorporating principles from human-centered AI (HCAI) such as an emphasis on user control[6], mutual learning from the field of Hybrid intelligence (HI) [7], and active learning and feedback loops from the field of Interactive Machine Learning (IML) [8]. The latter is addressed particularly well by the HI framework which presents an integrated way of deploying human-centered AI solutions with appropriate information system management methodologies to optimize business, societal and human values [9].

In this paper, we investigate how the HI approach helps human experts build a partnership with a GD personal assistant in design co-creation. We define partnership as a human expert's willingness to contribute to the co-creative tool during and after co-creation. Our contribution to human-AI interaction is three-fold. First, we have developed a novel grammar-based methods for constructing common language between human and algorithm allowing for the explicitation of individual experts' "design goals" and a method for feeding these in real-time into the generative algorithm. Second, we apply the human-centered AI interaction design principles to seamlessly integrate the AI-training into the expert's task workflow. These two algorithmic and human computer interaction advances enable individuals to directly communicate design preferences and goals to AI and gradually grow an accumulated and personalized design knowledge library. Finally, we address the willingness to spend time and effort training such a virtual assistant by embedding the process in an HI narrative designed to create a psychologically safe space for co-creation without the fear of job-replacement that is so often an underlying perception of rapidly advancing AI. We present our exploratory findings through a newly developed Hybrid Intelligence Technology Acceptance Model (HI-TAM).

2. Related Work

2.1. Generative Design Tools - The Technology

Generative Design (GD) [10] is a design process where designers utilize the power of artificial intelligence to explore large design space to deliver high-quality designs that

balance multiple design objectives.

Various technologies can be used for implementing GD including simulation, optimization (e.g., genetic algorithm), deep learning models and a combination of those (e.g. [10,11]). GD has been applied in many domains especially architecture design [12,13] and product design [14]. Tools have been developed to support GD processes, including the GD toolset for Autodesk Revit[1], the Refinery toolkit for Dynamo[2], and Grasshopper for Rhinoceros 3D[3].

A GD process allows for computational expression of design goals through a parametric model and automatic generation of numerous design options, in contrast to traditional design processes where designers must internalize all design goals and constraints to create a single solution. The GD process is human-AI collaborative in nature as the algorithm can report back to the user promising design options for further analysis and refinement; the user can also revise their input parameters. Research along this line has been so far mostly focused on representation of the design space, generation and evaluation of solutions, search algorithms and visualization of design options, with little discussion on the human-AI collaborative and user personalization.

In this study, we implemented a prototype system demonstrating a typical GD workflow in a simplified design problem, with an intention to study human-AI co-creation behaviour in a controllable research setting and discover useful insights that can be translated to improve workflows in practical GD software.

2.2. Virtual Assistants - The Application

Broadly, the goals of a virtual personal assistant is to provide support to users in a personalized and context-aware manner, thereby enhancing their productivity, satisfaction, and overall well-being [15,16]. As an early example, the Microsoft Paperclip, also known as Clippy, was a virtual assistant introduced in the late 1990s to assist with tasks such as creating and formatting documents in Microsoft Office. It could be accessed by clicking on a small paperclip icon in the Office application window. However, its implementation was widely criticized for being intrusive, annoying, and unprofessional due to unsolicited messages and its cartoonish appearance and behavior [17]. With the advancement of automated test and voice recognition and processing [18], virtual assistants and chatbots have proliferated recently in both the commercial and private spheres providing helpful input and innovative interactions but always within quite restricted domains.

In contrast, the recently launched ChatGPT seems to provide human-like conversation and assistance as a virtual assistant. However, in terms of output credibility there are still significant pitfalls [19] and in terms of user personalization, in its own words "As an AI language model, I do not have the ability to adjust to individual preferences in the way that humans do... I can be trained on large datasets of text to learn how people typically communicate, which can inform my responses to some extent." This lack of information about individual users' preferences, needs, or context can lead to generic or irrelevant responses that do not address the user's specific concerns or objectives. Furthermore, because ChatGPT does not have a memory of previous interactions with a particular user beyond the current session ("I do not store or transmit any personally identifiable information unless specifically instructed to do so by the user"), it may not be able to provide a consistent and coherent conversation or maintain a sense of continuity in interaction over time.

The importance of user feedback and preferences in designing virtual assistants, as well as the need for more advanced techniques such as personalized recommendation systems and user modeling, is highlighted by the limitations of ChatGPT. ChatGPT's inability to communicate residual uncertainty in its responses creates algorithmic overconfidence and a lack of transparency for the user. The study defines a GD virtual assistant as an AI system trained to assist human experts by generating design solutions based on their inputs and preferences. The HI narrative is introduced as part of our HI approach to clarify that the AI in question is fallible in the beginning and can only improve with the user's continual training and feedback.

2.3. HI-TAM - The Analysis Method

Researchers in the field of information systems management have developed models to understand factors influencing technology acceptance, including the widely used Technology Acceptance Model (TAM) which links user acceptance of technology to perceived usefulness and ease of use [20]. The TAM has been applied to various contexts such as mobile apps and e-commerce systems. An extension of TAM, the AI-TAM, has been proposed to evaluate user acceptance and collaborative intention in human-in-the-loop AI applications [21]. Human-in-the-loop AI involves integrating human input and feedback into AI systems to improve their accuracy and efficiency [22]. The AI-TAM includes constructs related to human-AI interaction such as functioning, quality, trust, familiarity, and collaborative intention. Despite adding the collaborative variable, the scenario investigated in the AI-TAM development work involved single-shot interactions with a fully developed product, which is far from the continuous mutual learning relationship with an HI system.

Although originally perceived as a useful framework for understanding the early stages of product design, the existing TAM approach has been heavily criticized in relation to product design [23]. However, subsequent literature such as the New Product Development TAM (NPD-TAM) [24] as well as case studies of developing smart payment card and adopting virtual assistants [25,26] have added no new variables to provide insight into the dynamic development and training process. In order to capture the dynamic mutual learning process of HI systems, we here introduce the HI-TAM (see Figure 1 for an overview and Table 1 for detailed variable definitions). Taking the AI-TAM as a point of departure, we added the process variables of *user control*, *AI output transparency*, *perceived partnership* and replaced the output variables *collaborative intention* and *behavioral intention* with *willingness to train*, *willingness to co-develop* and *willingness to adopt*. These variables were inspired by fundamental principles of HI (mutual learning), HCAI (pursuit of high levels of automation and control simultaneously), and IML (continuous interactivity) as well as the Co-Creative Framework for Interaction Design (COFI) which emphasizes the importance of establishing a partnership between end users and AI support tool [27]. The links represented in Figure 1 are supported by exploratory data analyses described in Section 4.1 and any omitted variables should be validated in further studies.

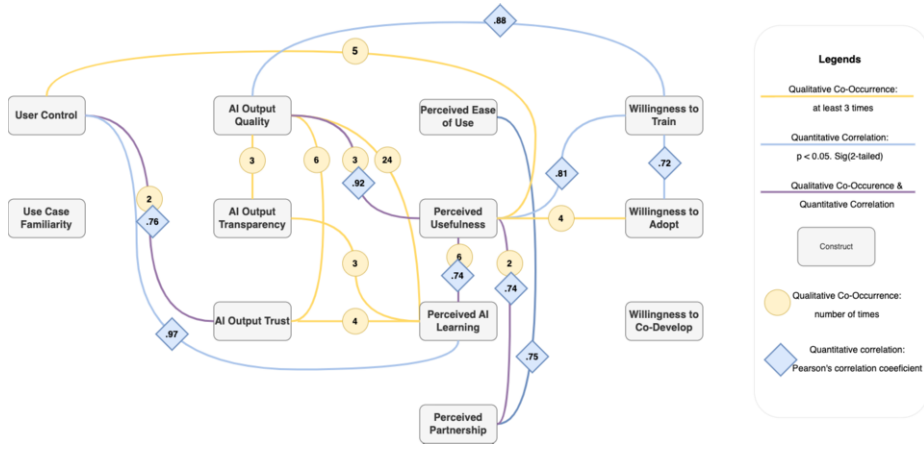

Figure 1. HI-TAM: Hybrid Intelligence Technology Acceptance Model adapted from the AI-TAM [21] incorporating key aspects from HI including AI transparency and user control, in order to support both virtual assistant training and general human-AI mutual learning. Qualitative co-occurrences and quantitative correlations were based on N=8.

3. Methods

To explore the proposed HI approach and the underlying HI-TAM, we conducted a one-shot case study [28] using a descriptive and correlational study design. Individual participants (i.e. architectural professionals) were introduced to a hypothetical scenario that is similar to their day-to-day work, where they were tasked with designing a floor plan layout of a typical office building. They were also introduced to work with a GD assistant through an HI approach. Both qualitative and quantitative measures were collected.

3.1. Hybrid Intelligence Approach Design

We now describe the 3 components of the *Hybrid Intelligence (HI) Approach* (see Figure 2 for an overview and Supplementary Materials[4] for details of user interface, system implementation and instruction narrative).

3.1.1. Programmable Common Language for Representing Human Experts' Design Goals to GD Assistant

In building layout design, a substantial amount of design requirements are about types of spaces and their spatial and topological relationships. We thus use sentences of the following syntax to represent design goals:

$$\textit{design\_goal} \rightarrow \textit{subject}\ \ \textbf{unary\_relation} \mid \textit{subject}\ \ \textbf{b}\textit{inary\_relation}\ \ \textit{object}$$

$$\textit{subject} \rightarrow \textbf{space\_type}$$

$$\textit{object} \rightarrow \textbf{space\_type}$$

In our prototype, **unary_relation** includes `at the center`, `on west`, `on east`, `on south`, `on north` and **binary_relation** includes `are close to`, `are away from`,

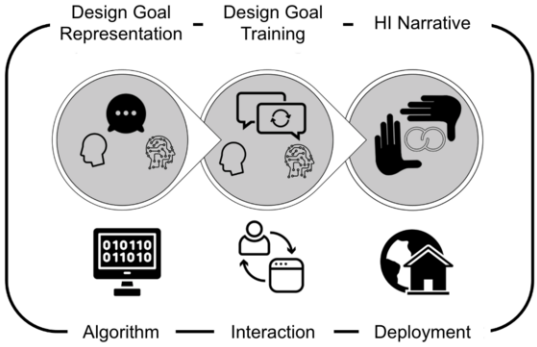

Figure 2. Illustration of the 3 Hybrid Intelligence (HI) components: programming a common language for humans and algorithms to interact, designing the interface for continual learning loops, and presenting the adoption within a broader framing of HI creating a psychological safe space for co-development.

`surrounded by`, `share the same door orientation with`. We also consider an office building setting for the layout design task, so **space_type** include `meeting room`, `office`, `open_desk_space`, `lunch_space`, etc. We compute an objective goal satisfaction score (as opposed to a subjective goal satisfaction score rated by participants) of each design goal by mapping each **unary_relation** and **binary_relation** to a function, which takes a layout configuration as input and returns a real number in the range of $[0, 1]$ as output. The closer the value is to 1, the more satisfied the design goal. The *objective goal satisfaction* of a layout configuration given a set of design goals is the product of the satisfaction scores within the goal set.

3.1.2. Continual Learning Loops for Training GD Assistant with Design Goals

Leveraging the programmable design goals as a starting common language between GD assistant and a human user, we created a training mechanism for users to iterate on their design goals and the tool-generated designs in feedback loops through the following steps: i) A user reacts to a tool-generated design by marking spatial objects their like or dislike. ii) The tool carries on the ”conversation” by prompting the user with a popup window to select reason(s) for their likes or dislikes. The selected reasons are added into the tool as design goals following the programmable language. iii) The user repeatedly marks likes and dislikes and select reasons until they feel sufficient. They can also revise any added design goals if they detect any conflicts among them. iv) Upon user request, the tool is invited to generate another round of designs, taking all the updated design goals from the previous rounds into consideration. v) The user selects one preferred design and repeats steps i) through iv).

This training mechanism is based on two major inspirations. One is the typical design critique process that architecture students would learn in their design studio and professional architects would practice at their daily work [29]. The second is IML where the system is tightly coupled with the human in the loop of model training and thus resulted in “more rapid, focused, and incremental model updates than in the traditional machine-learning process” [30,8,31]. Here are a few examples of how we translate these human-in-the-loop design opportunities into the training mechanism design: defining new constraints inspires expressing design goals, correcting errors in the training data inspires marking dislikes on the design, fine tuning parameters inspires adjusting previous design

goals. These design considerations are also in line with the principles outlined in the *Guidelines for Human-AI Interaction* which seeks to guided interaction over time beyond one shot usage, including learning from user behavior, updating and adapting cautiously, encouraging granular feedback and conveying the consequences of user actions [32].

3.1.3. HI Narrative For Nurturing Partnership Between Human Experts and GD Assistant

According to [9], one critical issue that limits companies from successfully adopting AI solutions is employees' fear towards job automation and replacement, and thus requires a thoughtful deployment of these solutions into the professional work context. In their case study [9], a HI corporate narrative was created to onboard employees to an AI-supported editing tool and facilitate their adoption willingness through tool customization based on their preferences instead of following a standardized rigid workflow. We created our HI narrative by following the proposed HI-TAM as the guiding design principles and taking inspirations from [9]'s narrative design. Specifically, our narrative introduces the GD assistant as a partner and emphasizes that the goal was to train GD assistant sufficiently towards building a partnership instead of achieving the best quality design. Partnership was defined as "the distribution of sub-tasks in an integrated and customizable workflow between you and the tool". Participants were instructed to keep customizing the GD assistant "by telling it about your preferences until you feel that you have trained it enough".

3.2. Measurement

For each HI-TAM construct, we included both a qualitative code that describes the participants' subjective experience and a quantitative measure to evaluate the construct with a numerical range, adapted from existing instruments (Table 1).

Table 1.: Constructs with respective qualitative codes and quantitative measures

Construct	**Qualitative Codes**	**Quantitative Measures** (Data Range)
Familiarity	Both familiarity towards AI, and familiarity towards AI and Generative Design tools.	I don't use it - Expert (1-6) • Proficiency with machine learning or artificial intelligence for automated design. • Proficiency with generative design tools such as Grasshopper or Dynamo for automated design.
User Control	The level of control and autonomy the user feels they have over the tool. It also includes comments about input features that the user appreciates or feels are lacking. This code could be applied to any mention of the user's ability to direct or influence the tool's behavior or output.	*Adapted from controllability[33,34,35] (1-5)* • I am able to let it do its work. • I do not feel out of control in working with it.
AI Output Quality	The system provides accurate and complete information. This construct refers to subjective and objective evaluations of the quality of concrete, identified outputs of the system.	*Adapted from ability[36] (1-5)* • It is very capable of performing its job. • It has specialized capabilities that can increase our performance.
AI Output Transparency	The degree to which the output generated by an artificial intelligence system is understandable and interpretable to the human user. It pertains to the system's ability to provide clear and coherent feedback or explanations to the user on how it arrived at a particular output	*Adapted from comprehensibility[33,37,35] (1-5)* • I understand its inputs and outputs. • I'm familiar with it.

AI Output Trust	The user finds it predictable and trustworthy, takes the assertions as valid and true.	*Adapted from predictability[33,37,35] (1-5)* • I am certain about how my interaction with it affects the generated designs. *Adapted from integrity[36]* • Sound principles seem to guide its behaviors. *Adapted from trust* • It is reliable. • I can trust it.
Perceived Ease of Use	Using the system helps the user's work, to achieve tasks, and to make better choices. The user finds the system useful. This construct refers to subjective evaluations of the system as a whole. This includes when the tool provides a positive output which is surprising, creative, or novel to the user.	*Adapted from TAM - perceived ease of use[36]. (1-5)* • Working with it is easy for me. • The interface is difficult to understand.
Perceived Usefulness	Using the system does not require a lot of mental effort, the system is easy to use and understand.	*Adapted from TAM - perceived usefulness [36] (1-5)* • I feel more involved in the generative design process working with it compared to other generative design tools. • I feel responsible for marking my likes and dislikes working with it. • It is clear to me why I need to work with it. • Working with it makes me feel in control of the design process. • I think it is useful. • I think it is useless to train it with my likes, dislikes and reasons.
Perceived Partnership	The user's subjective perception of the degree to which an AI is perceived as a collaborative and cooperative partner in the interaction. In other words, it describes how much the user feels that the AI assistant is working with them, rather than just functioning as a tool that responds to their commands.	• Overall, how much do you feel that you are building a meaningful partnership i.e. task distribution and integration and tool customization? (1-5)
Perceived AI Learning	The user's perception of an AI's ability to learn from their interactions and adapt to their needs and preferences. It refers to the degree to which the user believes the AI assistant is capable of improving its performance over time by learning from the user's behavior, feedback, and input. This code covers anything that is descriptive of the process over several inputs.	• To which extent did you feel that your interactions with the tool yielded improved designs? (1-5) • Please rate one a scale of 1 to 5, how much do you see your contributed reasons are reflected in the newly generated designs by the tool? (Subjective goal satisfaction (1-5)) • (Objective goal satisfaction: see details in 3.1.1 (0-1))
Willingness to Train	The user's willingness and openness to invest time and effort in training an AI tool to better understand their needs and preferences. It refers to the degree to which the user is willing to provide input and feedback to the AI tool, with the goal of improving its performance and adapting it to their specific needs.	• Would you invest your time in training to use this tool in your work practice? (1-5) • (The number of design goals.)
Willingness to Adopt	The user's intention and openness to use and incorporate the AI tool into their workflow. This also includes for alternative or future uses of the tool.	• Would you use such a tool in your work practice? (1-5)
Willingness to Co-Develop	The user's willingness and openness to actively provide feedback and input to help shape the development and improvement of the AI tool. This also includes suggestions for feature improvements.	• Please list as many suggestions as possible to improve the tool. (The number of suggestions.)

3.3. Study Procedure and Data Collection

The study took place online over Zoom and each participant were compensated $60 for their one-hour participation. Figure 3 shows major steps of the task workflow in experiencing HI. During the task, individual participants were encouraged to think aloud and asked to evaluate the iterated designs in interaction throughout by rating a numerical score of subjective design goal satisfaction. Additionally, GD assistant logged objective design goal satisfaction and the number of design goals. Participants' verbal and non-verbal behaviors were also recorded and then transcribed. Additionally, participants also completed pre-task survey questions (e.g. demographics, general experience in architectural design and with generative design tools and AI, as well as AI replacement fear) and

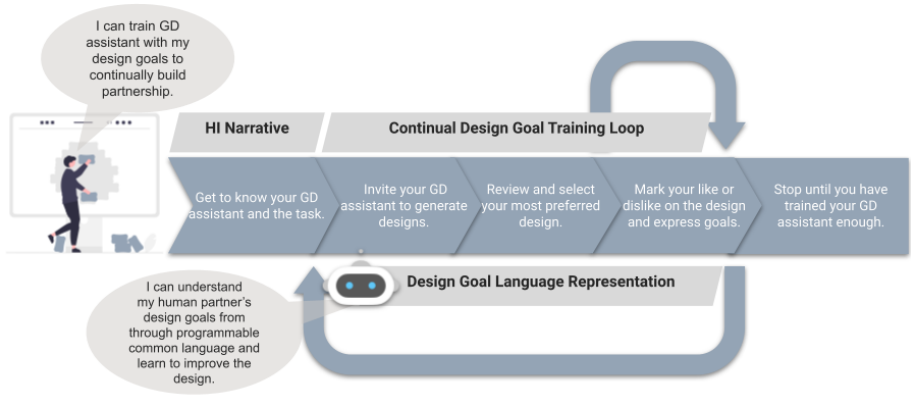

Figure 3. Overview of the interaction workflow with the 3 components of the HI approach through the eyes of a human expert vs. GD assistant.

post-task survey based on various constructs of the underlying model (e.g. willingness to adopt, willingness to train, providing as many suggestions as possible to improve GD assistant).

A total of 8 participants were recruited through e-mail, with the following inclusion criteria: age between 18 and 65, working professionals with at least 3 years of architectural design experience, and fluency in English. They were between the age of 31 and 38, with 3-16 years of architectural design experience (mean = 8.6, median = 8), and 3 of them were female. All had previous experience working with GD tools such as Grasshopper or Dynamo and half of them had experience with machine learning or artificial intelligence for automated design. Almost all of them (7 out of 8) expressed that in 10 years from now, a moderate amount of their current tasks in architectural design could be done by a machine instead of themselves, such as designing room layout, collaborative design review, checking building code compliance. All participants provided informed consent prior to the experiment.

3.4. Analysis

In general, we combine and compare all the qualitative and quantitative findings to seek complementary validation and explanations. We used Reflexive Thematic Analysis to examine user perceptions and interactions with the tool [38]. All authors participated in iterative coding rounds. Initial codes were identified from the transcripts, including known codes from survey questions such as "partnership" and "user control" and five constructs from the AI-TAM [21]. Collaborative and behavior intention codes were too vague to differentiate our data so they were replaced with willingness to train, adopt, and co-develop; codes for AI output trust, transparency, and perceived AI learning were also added to account for key aspects of HI. Relevant excerpts were coded with one or two of the most appropriate codes, and co-occurrences were mapped to determine linked constructs. Excerpts were labeled positive or negative. For example, P1 stated "They are not quite consistent in a way I like" referring both to the negative output quality of the output as well as negative AI output trust.

For quantitative measures, we first evaluated the reliability of any constructs with at least 3 questions using the same scale using Cronbach's alpha [39]. We obtained a min-

imum alpha value above ($\alpha > .7$), indicating a high internal consistency among items within the same construct and thus the survey measurement can be considered reliable. We also computed Pearson's correlation [40] to explore relationships among these constructs in comparison to the qualitative co-occurrences. Furthermore, we conducted a hierarchical clustering [41] of willingness to adopt, train and co-develop to explore the overall types of the resulted "partnership" profiles using Ward's method with square Euclidian distance as the distance or similarity between participants.

4. Results

4.1. HI-TAM

In this section, we present preliminary findings on the qualitative and quantitative links among constructs of the HI-TAM (see example codes and detailed results in Supplementary Materials). In total there were 106 relevant excerpts from the transcripts, ranging from 7 to 21 per participant. The coding identified the strongest ties between the constructs. To be a strong qualitative co-occurrence, it needed at least three instances across two participants. The quantitative findings reported significant statistical correlations.

4.1.1. Qualitative Co-Occurrences

The most common co-occurrence identified was between AI output quality and perceived AI learning (24 instances), observed by 6 out of 8 participants. Co-occurrences between AI output trust and AI output quality (6), perceived AI learning and perceived usefulness (6), perceived usefulness and user control (5), and perceived usefulness and willingness to adopt (4) were also identified. Positive constructs were nearly twice as frequent as negative ones, with most co-occurrences being positive-positive or negative-negative, meaning that when construct one had a positive connotation, construct two also had a positive connotation. Exceptions included willingness to co-develop co-occurrences which were often positive-negative as negative comments were linked to suggestions for improvement.

4.1.2. Quantitative Correlations

From the quantitative analysis, all the statistical correlations were positive, meaning one construct increased with another. For example, user control was strongly associated with perceived AI learning ($r = .97$, $p < .001 **$) and AI output trust ($r = .76$, $p = .031*$) suggesting users' control over the tool might play an important role in how they predict and trust GD assistant and how they assess it as capable to learn and adapt through interaction. We also found correlations between previous constructs from AI-TAM and new constructs from HI-TAM. In particular, the greater ease of use was associated with perceived partnership ($r = .75$, $p = .033*$); the greater perceived usefulness was also associated with greater AI output quality($r = 0.92$, $p = .001 **$), greater perceived AI learning($r = .74$, $p = .037*$), greater perceived partnership($r = .74$, $p = .037*$) as well as more willingness to train($r = 0.81$, $p = .015*$). Additionally, AI output quality was associated with willingness to train ($r = .88$, $p = .004 **$). Moreover, more willingness to train was associated with more willingness ($r = .72$, $p = .046*$) to adopt, indicating participants' willingness to train and their willingness to adopt might have influenced or

even reinforced each other when guided by the HI approach (asterisks indicate significance level: $***p < .001, **p < .01, *p < .05$).

Notably, quantitative and qualitative findings supported links between perceived usefulness and AI output quality, perceived AI learning, and perceived partnership as well as user control and AI output trust.

4.2. Three Types of Partnership

The hierarchical clustering resulted in 3 distinct clusters of participants that vary according to four quantitative measures of the three partnership constructs regarding GD assistant, 1) willingness to adopt measured by the post-task willingness to use it in work practice, 2) willingness to train measured by the number of design goals participants contributed during interaction and the post-task willingness to invest time in training to use it in work practice, 3) willingness to co-develop measured by the number of suggestions proposed to improve it after the task. Figure 4 displays the average statistics for each cluster profile contours. For each cluster, we describe their profile statistics and the most common qualitative code patterns shared across participants. By relating qualitative and quantitative findings, we hope to explore and reason further hypotheses about individual differences when experiencing the HI approach, and point future directions for researching and improving the HI approach.

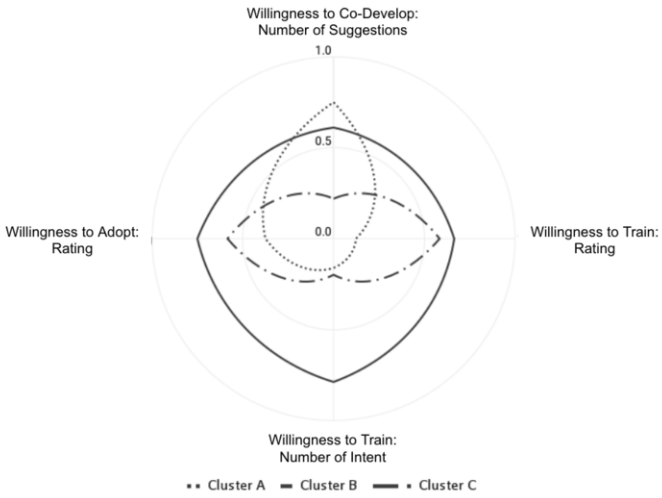

Figure 4. Mean of each partnership measure by cluster: using normalized measure values.

Cluster A (P5, P8) represents those who are the least willing to adopt, the least willing to train but the most willing to co-develop the tool. Both participants had notable occurrences of user control (3, 5 times) and perceived usefulness (4, 7 times), mostly with a negative sentiment. In particular, P8 who had the most mentions of willingness to co-develop (6 times) and linked willingness to co-develop with perceived usefulness, and user control. He provided a suggestion of how the system could be changed to increase control: "Have you used the feature in DALL-E where you can lock part of the image and regenerate? Yeah, I think that would be useful. That way you feel more in control of what

you are doing." They both found it more pressing to improve user control and usefulness of GD assistant in their suggestions before they can invest in training or adopting it.

Cluster B (P1, P2, P7) represents those who are most willing to adopt, most willing to train and also highly interested in co-developing GD assistant. They shared the most of perceived AI learning (10, 9, 9 times) with a strong positive sentiment and AI output quality (18, 9, 3 times) with a mixed sentiment. As an example, P7 held his conflicted input to GD assistant responsible for the resulted negative output and expressed his duty to train GD assistant with a strong belief in its learning capability: "I don't know if my rules are all following each other. Some of them might be contradictory. But it seems like it is learning...So I guess I should tell it. Tell the tool which one is my favorite even out of these." All of them speculated GD assistant's learning and reflected on how their training input can improve the interaction and the output quality, which might in return held them optimistic and responsible for both adopting and improving the tool.

Cluster C (P3, P4, P6) represents those with a moderate level of willingness to adopt or train while the least willingness to co-develop. They shared the most mentions of perceived AI learning (2, 5, 6 times) with a mostly positive sentiment and perceived usefulness (4, 7, 2 times) with a mixed sentiment. Moreover, they all hoped to appropriate GD assistant for alternative uses other than creative design. In particular, P6 was mindful of how GD assistant was learning from his behavioral input but expressed his concerns of its artistic expression limitation: "There are two aspects in architectural design: the scientific part also the functional part can be definitely done with machine learning, with program. It's just the artistic part...It's kind of tricky because it's just like a poem. It's so personal." Therefore he suggested GD assistant could be better used in its strength, "the scientific part also the functional part", for producing or checking designs in compliance with building code: "So a lot of those rules can be applied to the master plan, and can probably be used to generate like a basic layout without concept. A messy analysis can all be done by a machine and you work on the more artistic part". In summary, all of them acknowledged GD assistant's learning capability to some extent but had strong opinions of how human experts and AI should handle different design tasks. To them, GD assistant can never design or learn like humans do regardless of the amount of training investment. Thus it made sense to them to appropriate it for other contexts rather than improving it further.

5. Discussion

Firstly, it is important to reiterate that 7 out of 8 participants expressed that in 10 years from now, a moderate amount of their current tasks in architectural design could be done by a machine instead of themselves. This finding underscores the importance of investigating effective methods for integrating human and AI approaches in GD. In this paper, we set out to investigate to what extent the HI approach helps human experts to build partnership in design co-creation. We define the concept of partnership as a human expert's willingness to contribute to the creative tool, a GD personal assistant, during and after co-creation. We here discuss the the results from our study and implications for future work. Given the small scale of our study, all findings should be empirically tested further.

Given that the participants were asked to report their goal satisfaction with the generated layouts after each round, it is unsurprising that AI output quality and perceived AI

learning were the as the most frequently identified constructs. This co-occurrence suggests that human experts in architectural design highly value the ability of the GD assistant to generate high-quality design solutions while also rapidly adapting to their personal preferences. The link between AI output quality and AI output trust can be explained by the fact that the participants were generally happier with the outcome of AI output when it consistently reflected their design goals. Future iterations of this GD assistant could allow users to see the factors that are contributing to its output, such as the specific design preferences and goals that have been learned over time and which design goals it is weighting most when generating the design. In an effort to combat algorithmic overconfidence, future designs could display GD assistant's level of confidence in the generated design as a means of the algorithm communicating to the end user. Bi-directional communication has also been shown to increase perceived partnership [27].

Although the GD assistant described in this study is designed to track the user's design goals, building a personalized knowledge library that can be used to inform future design projects, there were several instances where participants were unclear if they their design goals were "contradictory" and they "confused" the assistant. Providing a way for the GD assistant to highlight if users are inputting contradictory information could improve perceived AI transparency. P7 exemplifies this with the following excerpt "I have said this before. That [design feature] should be closer to the window. And Now I'm saying...I think I'm contradicting myself."

From the quantitative analysis, the high and positive correlation found between willingness to train and willingness to adopt in this study could be a promising indicator of a "pathway to adoption". In other words, positive experiences training a virtual GD assistant could lead to a greater likelihood of adoption.

Both our quantitative and qualitative results supported a link between perceived usefulness and AI output quality, perceived AI learning, and perceived partnership. This is supported by the the final AI-TAM [21] which combined constructs of AI Output Trust and and AI Output Quality into the "super construct", explainable AI (XAI), correlating with perceived usefulness (.74). While the link between AI output trust and perceived usefulness was not supported in our analysis, perceived usefulness was linked to AI output transparency, and AI perceived learning. As the addition of AI output quality and perceived AI learning allowed the participants feedback to be more granular, it is unsurprising that there were fewer instances of perceived AI trust as the codes are conceptually related.

Our study's results suggest that perceived ease of use, as described in the AI-TAM, may have been absorbed into more fine-grained constructs such as AI output quality, AI output transparency, AI output trust, and perceived AI learning in our HI-TAM. Similarly, perceived usefulness may be able to be broken down into more granular metrics to evaluate and inform AI design.

Perceived partnership is one of the new constructs we added to the HI-TAM given its importance in human-AI co-creative systems [27]. Our HI-TAM shows quantitative correlations and qualitative co-occurrences between perceived partnership and perceived usefulness. Interestingly, there were no significant correlations or strong co-occurrences between perceived partnership and willingness to co-develop. We maintain that the construct of partnership is important to the HI-TAM and postulate the missing link between perceived partnership and willingness to adopt may be due to the fact that the tool is still in prototype form and may not exhibit enough qualities to warrant partnership. In

our analysis we also identify 3 distinct clusters of participants that vary according to 4 quantitative measures of the three partnership constructs (willingness to adopt, willingness to train and willingness to co-develop). While our sample is small, these initial findings shed light on the possibility that different modes of training may be necessary for different types of users.

6. Conclusion and Future Work

In conclusion, our study sheds light on the potential of the HI-TAM to inform the design of GD assistants that facilitate a co-creative partnership between human experts and algorithms. Opportunities for future work include improving the functionality, user experience, and integration of the current GD assistant prototype with existing design workflows and processes. Another avenue is to validate the effectiveness of the HI narrative by comparing it to a control group. The HI narrative could be further improved to address human experts' concerns about AI adoption and job displacement. Additionally, scalability and generalizability of HI in GD assistants and AI systems to other domains and contexts could be investigated.

Endnotes

1. www.autodesk.com/products/revit/overview
2. dynamobim.org/refinery-toolkit/
3. www.rhino3d.com/6/new/grasshopper/
4. Supplementary Materials: https://drive.google.com/drive/folders/1LKAl1i-mh9rqpScYFs9OTcJ-KVArGCbC?usp=sharing

References

[1] Kazi RH, Grossman T, Cheong H, Hashemi A, Fitzmaurice G. DreamSketch: Early Stage 3D Design Explorations with Sketching and Generative Design. In: Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology. UIST '17. New York, NY, USA: Association for Computing Machinery; 2017. p. 401–414. Available from: https://doi.org/10.1145/3126594.3126662.

[2] Keshavarzi M, Hotson C, Cheng CY, Nourbakhsh M, Bergin M, Rahmani Asl M. SketchOpt: Sketch-Based Parametric Model Retrieval for Generative Design. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. CHI EA '21. New York, NY, USA: Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3411763.3451620.

[3] Demirel HO, Goldstein MH, Li X, Sha Z. Human-Centered Generative Design Framework: An Early Design Framework to Support Concept Creation and Evaluation. International Journal of Human–Computer Interaction. 2023:1-12.

[4] Rafner J, Dellermann D, Hjorth A, Verasztó D, Kampf C, Mackay W, et al. Deskilling, Upskilling, and Reskilling: a Case for Hybrid Intelligence. Morals & Machines. 2022;1(2):24-39.

[5] Boston Consulting Group. Study Finds Significant Financial Benefits with AI; 2020. https://www.bcg.com/press/20october2020-study-finds-significant-financial-benefits-with-ai.

[6] Shneiderman B. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS). 2020;10(4):1-31.

[7] Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid Intelligence. Business & Information Systems Engineering. 2019;61:637-43.
[8] Amershi S, Cakmak M, Knox WB, Kulesza T. Power to the people: The role of humans in interactive machine learning. Ai Magazine. 2014;35(4):105-20.
[9] Rafner J, Bantle C, Dellermann D, Söllner M, Zaggl MA, Sherson J. Towards Hybrid Intelligence Workflows: Integrating Interface Design and Scalable Deployment. In: HHAI2022: Augmenting Human Intellect. IOS Press; 2022. p. 310-3.
[10] Shea K, Aish R, Gourtovaia M. Towards integrated performance-driven generative design tools. Automation in Construction. 2005;14(2):253-64. Education and Research in Computer Aided Architectural Design in Europe (eCAADe 2003), Digital Design. Available from: https://www.sciencedirect.com/science/article/pii/S0926580504000809.
[11] Oh S, Jung Y, Kim S, Lee I, Kang N. Deep generative design: Integration of topology optimization and generative models. Journal of Mechanical Design. 2019;141(11).
[12] Caetano I, Santos L, Leitão A. Computational design in architecture: Defining parametric, generative, and algorithmic design. Frontiers of Architectural Research. 2020;9(2):287-300. Available from: https://www.sciencedirect.com/science/article/pii/S2095263520300029.
[13] Nagy D, Villaggi L, Benjamin D. Generative urban design: integration of financial and energy design goals in a generative design workflow for residential neighborhood layout. In: Symposium on Simulation for Architecture and Urban Design. vol. 3; 2018. .
[14] Alcaide-Marzal J, Diego-Mas JA, Acosta-Zazueta G. A 3D shape generative method for aesthetic product design. Design Studies. 2020;66:144-76. Available from: https://www.sciencedirect.com/science/article/pii/S0142694X19300791.
[15] Kepuska V, Bohouta G. Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In: 2018 IEEE 8th annual computing and communication workshop and conference (CCWC). IEEE; 2018. p. 99-103.
[16] Dubiel M, Halvey M, Azzopardi L. A survey investigating usage of virtual personal assistants. arXiv preprint arXiv:180704606. 2018.
[17] Maedche A, Morana S, Schacht S, Werth D, Krumeich J. Advanced user assistance systems. Business & Information Systems Engineering. 2016;58:367-70.
[18] Yuan X, Chen Y, Zhao Y, Long Y, Liu X, Chen K, et al. Commandersong: A systematic approach for practical adversarial voice recognition. In: 27th {USENIX} Security Symposium ({USENIX} Security 18); 2018. p. 49-64.
[19] Borji A. A Categorical Archive of ChatGPT Failures. arXiv preprint arXiv:230203494. 2023.
[20] Davis FD. Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS quarterly. 1989;13(3):319-40.
[21] Baroni I, Calegari GR, Scandolari D, Celino I. AI-TAM: a model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. Human Computation. 2022;9(1):1-21.
[22] Zanzotto FM. Human-in-the-loop artificial intelligence. Journal of Artificial Intelligence Research. 2019;64:243-52.
[23] Salovaara A, Tamminen S. Acceptance or appropriation? A design-oriented critique of technology acceptance models. Future interaction design II. 2009:157-73.
[24] Diamond L, Busch M, Jilch V, Tscheligi M. Using technology acceptance models for product development: case study of a smart payment card. In: Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct; 2018. p. 400-9.
[25] Song YW, et al. User acceptance of an artificial intelligence (AI) virtual assistant: an extension of the technology acceptance model; 2019.
[26] Gunadi D, Sanjaya R, Harnadi B. Examining the acceptance of virtual assistant-Vanika for university students. In: 2019 3rd International Conference on Informatics and Computational Sciences (ICICoS). IEEE; 2019. p. 1-4.
[27] Rezwana J, Maher ML. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. ACM Transactions on Computer-Human Interaction. 2022.
[28] Creswell JW, Creswell JD. Research design: Qualitative, quantitative, and mixed methods approaches. Sage publications; 2017.
[29] Oh Y, Ishizaki S, Gross MD, Do EYL. A theoretical framework of design critiquing in architecture studios. Design Studies. 2013;34(3):302-25.
[30] Fails JA, Olsen Jr DR. Interactive machine learning. In: Proceedings of the 8th international conference

on Intelligent user interfaces; 2003. p. 39-45.
[31] Dudley JJ, Kristensson PO. A review of user interface design for interactive machine learning. ACM Transactions on Interactive Intelligent Systems (TiiS). 2018;8(2):1-37.
[32] Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems; 2019. p. 1-13.
[33] Shneiderman B, Maes P. Direct manipulation vs. interface agents. interactions. 1997;4(6):42-61.
[34] Hartmann M. Challenges in Developing User-Adaptive Intelligent User Interfaces. In: LWA. Citeseer; 2009. p. ABIS-6.
[35] Oh C, Song J, Choi J, Kim S, Lee S, Suh B. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems; 2018. p. 1-13.
[36] Mayer RC, Davis JH. The effect of the performance appraisal system on trust for management: A field quasi-experiment. Journal of applied psychology. 1999;84(1):123.
[37] Jameson AD. Understanding and dealing with usability side effects of intelligent processing. Ai Magazine. 2009;30(4):23-3.
[38] Braun V, Clarke V. Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health. 2019;11(4):589-97.
[39] Brown JD. The Cronbach alpha reliability estimate. JALT Testing & Evaluation SIG Newsletter. 2002;6(1).
[40] Freedman D, Pisani R, Purves R. Statistics Fourth Edition. WH Norton & Company New York; 2007.
[41] Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. The quantitative methods for psychology. 2015;11(1):8-21.