

Agency, Trust, and Interpretability of Generative Adversarial Networks (GANs)

Design of an interface that utilizes feature design, visualization, and explanations to build an AI novice's mental models for a GAN system.

Sakshi Gupta(preferred name: Syashi)

*Department of Graphic and Industrial Design College of Design
North Carolina State University April 29th, 2021*

*Submitted in partial fulfillment for the degree of Master of
Graphic Design*

Helen Armstrong, Committee Chair

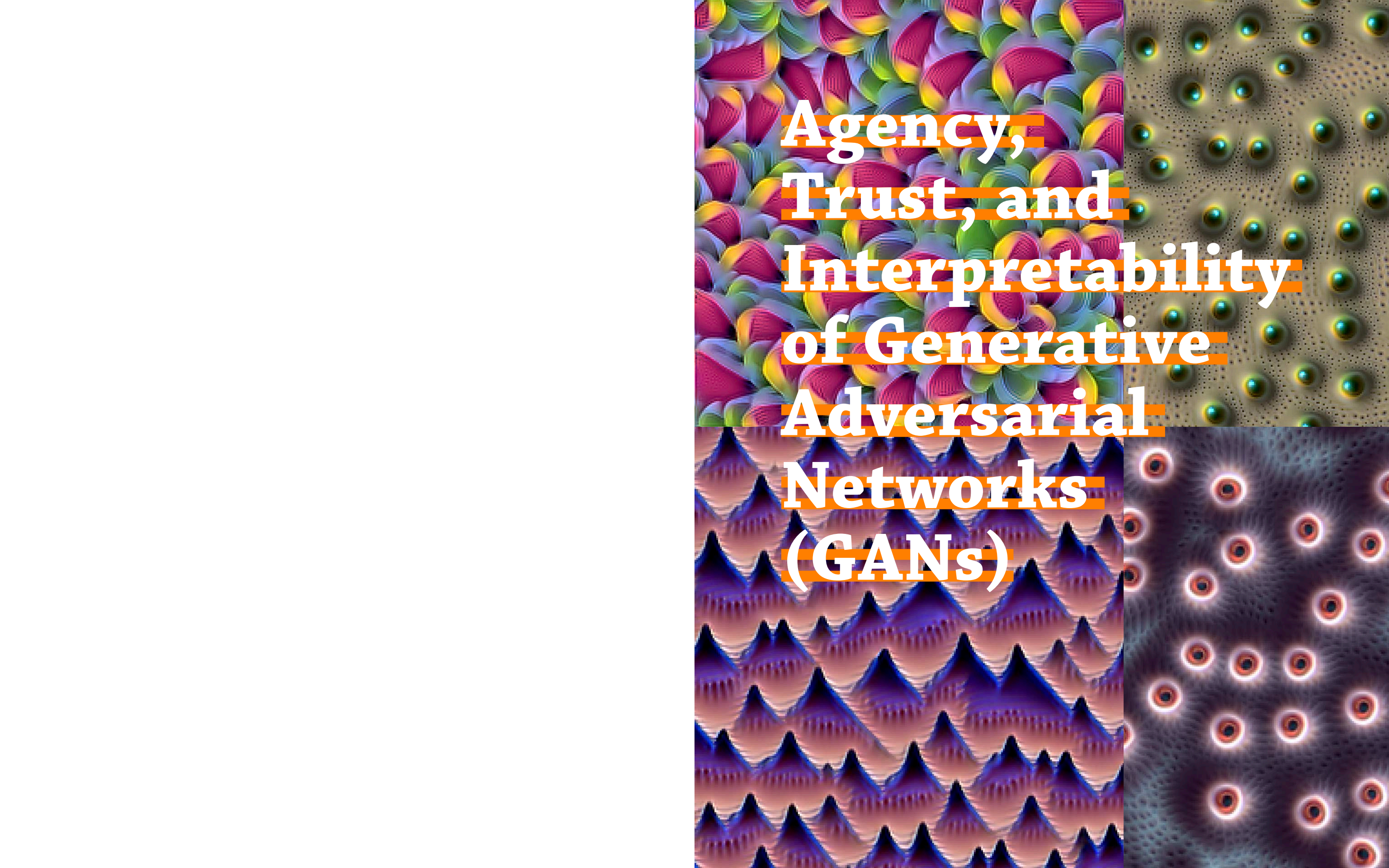
Associate Professor of Graphic Design

Matthew Peterson, PhD, Committee Member

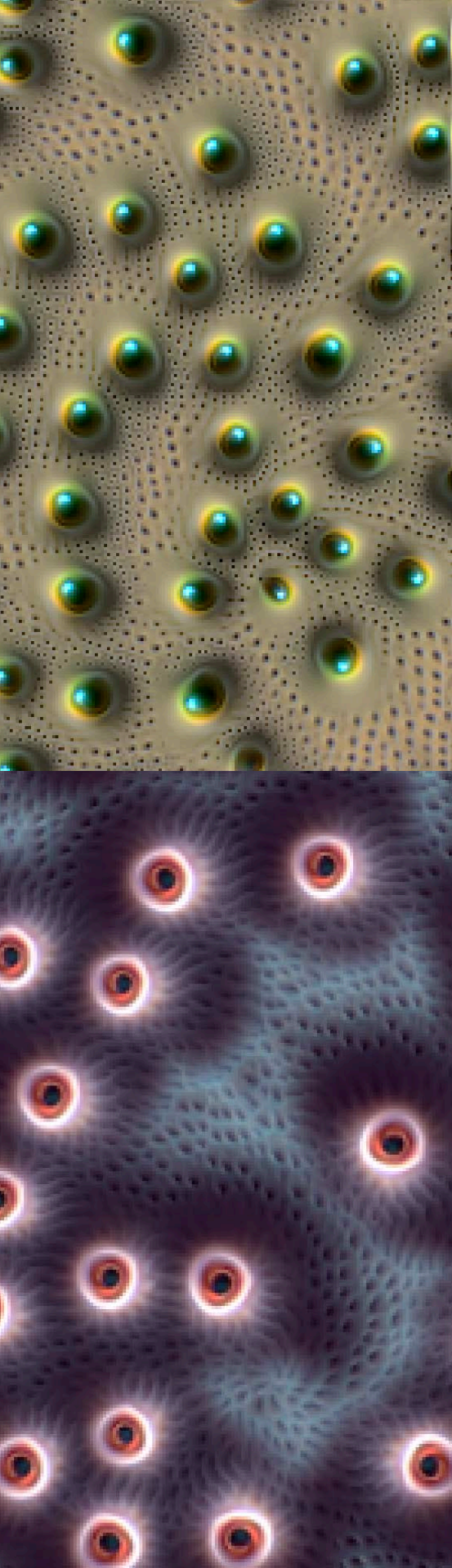
Assistant Professor of Graphic Design

Scott Townsend, Committee Member

Professor of Graphic Design

The background of the slide is composed of several distinct abstract patterns. On the left, there are vertical bands of colorful, fractal-like patterns in shades of purple, blue, green, and yellow. On the right, there is a grid of glowing, teal-colored spheres on a dark, textured background. Below the text, there is a pattern of glowing, red and white circular structures on a dark background.

Agency, Trust, and Interpretability of Generative Adversarial Networks (GANs)



SIGNATORY TITLE PAGE

MGD Statement Page Program Statement on the Master of Graphic Design Final Project

This document details a final project, which in design is commonly referred to as a graduate “thesis,” at North Carolina State University. The work was defined in a 3-credit course in a fall semester, and executed in a 6-credit course in the following spring semester. The Master of Graphic Design is a terminal professional degree with a research orientation, but like the MFA and MDes, it is not a primary research degree. This is a discovery-based investigation. Cash (2018) describes the process of building scientific knowledge as a cycle between theory building and theory testing. The theory building mode includes (1) discovery and description, (2) definition of variables and limitation of domain, and (3) relationship building (pp. 88–89). This investigation is restricted to the theory building mode. The theory testing mode includes (4) prediction, testing, and validation, and (5) extension and refinement (p. 89). While experts may have been consulted, this investigation does not entail any testing with human subjects, and it does not endeavor to prove anything; all assertions are tentative and speculative.

See: Cash, P. J. (2018). Developing theory-driven design research. Design Studies, 56, 84–119.

Image credits - Hexells by Alexander Mordvintsev, generative unfoldings <https://generative-unfoldings.mit.edu/works/hexells/view.html>

Thank You

To my parents, without you, I don't think I would have been able to do my masters. I can't imagine what you would have gone through to put me here and supporting me throughout my masters.

To my sister Swati. I don't appreciate you enough, but you stood by me during the toughest of times.

To my cohort and friends. You guys always had my back! From editing my grammar, lending voice for my prototypes to making me smile through the miserable times. You all honestly gave meaning to this place (Raleigh) for me.

To my amazing, helpful, and kind committee- Helen, Matt, and Scott. I am grateful for all your feedback that helped me improve my work. I owe you a lot, Helen. Your classes changed my life!

To my future team at Adobe. Especially Ryan and Rachel. You both gave me a chance last summer, and the learnings from my internship guided many parts of my research. I consider myself lucky.

Finally, thank you to the universe and all the positive energy that removed toxic people and blessed me with good ones.

I dedicate my research to my home and country, India. Quoting Neri Oxman, architect and designer- "You have to go away to come back home. You'll never really truly have a sense of home until you leave home." I realize how important home is to me, and I miss everyone back at home!

TABLE OF CONTENTS

Abstract

Introduction

Problem Space

2.1 Problem Statement16

2.2 Justification18

2.3 Annotated Bibliography19

2.4 Definition of Terms28

2.5 Assumptions and Limitations30

2.6 Precedents.....31

Investigation Plan

3.1 Conceptual Framework42

3.2 Research Questions44

3.3 Investigation Model.....45

3.4 Scenario.....47

Studies

4.1 Design of Digital Interface Features54

4.2 Visualization of Neural Networks70

4.3 Textual Explanations85

4.4 XAI Interface Design.....91

Discussion

5.1 Design Principles.....95

5.2 Future Work97

5.3 Conclusion99

References

6.1 Links101

6.2 Image Credit102

6.3 References.....103

ABSTRACT

ABSTRACT

Generative Adversarial Networks (GANs) are emergent neural network technology that allows users to generate new images, videos, or audio within a matter of minutes. Previous applications of AI were only able to classify, learn, and predict behaviors, but GANs enabled synthesis of content to produce something that did not previously exist. GAN technology affords robust generative creativity, but the technology is not without pitfalls. Automated output without active involvement of human input in the creative process can yield dull, even harmful results, that are often riddled with errors. Human manipulation at the level of a machine learning model can provide an astonishing amount of power and control over the generations of a GANs system but such manipulation is currently inaccessible and confusing for an AI novice. Drawing from research, this investigation gives an AI novice agency to explore with GAN systems via an explainable interface, understand system working and outputs (including erroneous ones), and notice visualized neural network internals to understand, manipulate, and constrain parts of generations. This investigation also conceptualizes utilizing the GANs system as a plug-and-play system (Application Programming Interface) on any software. The resulting Application Programming Interface (API) makes it simple for a user to flexibly utilize the functionality of both the API tool and software, set user intent, and establish context for the GAN system.

INTRODUCTION

“A single neuron in the brain is an incredibly complex machine that even today we don’t understand. A single ‘neuron’ in a neural network is an incredibly simple mathematical function that captures a minuscule fraction of the complexity of a biological neuron.” — Andrew Ng, scientist, Google Brain

INTRODUCTION

Living organisms are complex. Even a small part of humans—like a neuron in the brain—is biologically complex and intricately arranged. The human brain produces every thought, action, memory, feeling, and experience of the world. It took millions of years of evolution to get to this level of meticulous detail. Uncovering all the mysteries of brain function will take time. Creativity is a quality of the brain. To replicate and apply that creativity through technology is bound to invite many dialogues. With the design of a new machine learning framework called generative adversarial networks (GANs), the field of computational creativity—the domain of artificial intelligence that deals with creativity—expanded and grew multifold (Colton & Wiggins, 2012). GANs are very promising in terms of creation, and further research on this technology will only take us beyond to explore the unconventional with the AI systems.

First proposed in 2014, generative adversarial networks (GANs) can produce photorealistic images, often indistinguishable from reality. This ability to generate remarkably high-quality images has powered many real-world applications to synthesize realistic imagery (Bau et al., 2018). GANs can also produce exceptional quality audio and writing work apart from visuals. GANs are incredibly complex and can be daunting for an artificial intelligence (AI) novice. For an AI novice, it is challenging to understand the internal complexities of the technology as novices are oriented more towards results rather than internal neural linkages and workings (Kahng et al., 2019). While GANs promises an effective generation process, an AI novice’s interaction with and control of the generative models is limited. For an AI novice, it is challenging to interact with the models and guide them toward producing acceptable images as per user intentions. Interactive machine learning (iML) tries to make machine learning more accessible by involving users in the training process to create a more natural and powerful means of interacting with generative models (Dudley & Kristensson, 2018). Heim (2019) suggests that allowing user feedback iteratively in the process of output generation can help in correcting errors in the GAN models. Other than user feedback, users should be

INTRODUCTION

able to interact through an interface to practice control and constraint on GAN outputs. These interfaces should allow users to explore, visualize, use, and incorporate generative machine learning models into their creative work (Carter & Nielsen, 2017).

As users begin to interact with generative adversarial networks (GANs) more frequently, there is a need for ways to explain the computational system to them so that they know that the GAN generative process is reasonable. As systems become more complex and less interpretable, justifying system decisions and explaining the results becomes more crucial (Hoffman et al., 2019). A GAN system like other machine learning systems can produce outputs that can be biased, discriminatory, or unwanted with adverse effects. Hence, there is a push to create an explainable AI (XAI) system. An XAI can mitigate bias and malfunctions, develop novices’ mental models for system functioning, and engender appropriate trust. In addition to the system being more explicable, the resulting explanations should be interpretable by the user. In computer vision—an interdisciplinary scientific field that deals with how computers can gain high-level understanding from digital images or videos—there is an ongoing struggle to make system decisions interpretable by combining interfaces with the methods of neural network visualization. Although these methods of neural network visualization of the internals are meant for experts to investigate, improve, and fix problems in the models, combining visualizations with interface and explanations can also help an AI novice to better understand the system and its outputs. This combination can further enable an AI novice to iteratively tweak input towards the desired output goal, thereby promoting users’ agency(Olah et al., 2019; Olah et al., 2020).

There are many individual studies focusing on interpretable or explainable AI, providing agency in ML, feedback on ML systems, control and constraint on GANs, or interface design for ML. This investigation explores these elements in combination with each other, primarily for an AI novice, and prompts a question: How can the interface design of a GAN system facilitate AI novices’ interpretability, agency, and trust on a GAN system?

PROBLEM SPACE

- ∞ 2.1 Problem Statement
- ∞ 2.2 Justification
- ∞ 2.3 Annotated Bibliography
- ∞ 2.4 Definition of Terms
- ∞ 2.5 Assumptions and Limitations
- ∞ 2.6 Precedents

2.1 PROBLEM STATEMENT

A generative adversarial network (GAN) is a machine learning model in which two neural networks compete with each other to increase the accuracy of their predictions. Neural networks are complex models, and GANs consist of two neural network models, discriminator and generator (Creswell et al., 2018). GANs can achieve impressive results for many real-world applications of audio, text, and visuals, with many GAN variants emerging in sample quality and training stability. However, GANs' complex internal workings are often not visualized effectively or understood (Bau et al., 2018; Kahng et al., 2019). It is difficult for an AI novice to construct mental models of these internal sub-models, whose functioning is not even clear to the experts (Mohseni et al., 2020). Visualizing this internal functioning is difficult, and showing the structure of neural network connections does not impart any meaning to a user (Browne et al., 2018). GANs act as black boxes with observable input and outputs but with inscrutable internal processes. Furthermore, because AI novices use models that are pre-trained, it is difficult for them to see how neural networks make decisions leading to possible mistrust in the outputs (Browne et al., 2018; Samek et al., 2017).

There is a plethora of research on improving GAN models for better outputs, but few address the needs of users by giving them access to controlling and constraining the results via interaction with an interface (Heim, 2019; Sbai et al., 2018). Studies demonstrate that users want to interact with machines in a much richer collaborative manner than the current systems allow. When users build system reasoning, they can give generous feedback ranging from error correction to potential new features, which can lead to tremendous improvement in machine learning capabilities (Stumpf et al., 2007). If the design of an interface for GANs can allow more novice user engagement through agency and feedback on outputs while building user trust and reasoning, then this will directly help construct user mental models, in turn, enhancing the user's experience of GAN systems.

Some research projects bring users into the algorithmic loop so that users can add constraints or edit attributes on GANs through an interactive system (Sbai et al., 2018; Heim, 2019; Chrysos et al., 2020; He et al., 2018). Software like Runway ML, provides interfaces for users (mostly researchers) to train machine learning models and interact with them. The user base for Runway ML is experts and it does not allow for generating feedback. However other tools like CueFlik

and CueTip exist that demonstrate how users can exhibit agency by providing feedback, thus correcting errors and improving the systems (Shama et al., 2018; Fogarty et al., 2018; Shilman et al., 2006; Bryan et al., 2014; Tripathi et al., 2019; Putzu et al., 2020). Apart from these potentials for control and feedback in GANs, neural network visualization interfaces like Playground on tensor flow and GAN Lab visualize the working architecture of neural networks but do not explain any particular model working and output (Kahng et al., 2019). It is interesting to see how Runway ML allows creators with no coding experience to train and use their data models. None of this research, however, provides a refined explainable interface design through which a novice can interact with a GAN to utilize all these individual capabilities of agency, visualization of internals, and explainability. There is, however, potential for an API or add-on to be used alongside the software used by an AI novice for creative purposes. This will help AI novices who want to use pre-created models to fulfill their goals while practicing user agency, building trust, and developing reasoning of the GAN models through the explainable interface.

2.2. JUSTIFICATION

This investigation seeks design features of an interface for generative technologies as we move towards computational creativity or automation, ease of manual work, novel outputs, and generations of creative workflows through AI. Instead of establishing a whole interactive system for GANs and users, this investigation focuses narrowly on an AI novice’s user experience of agency and control in the GAN system.

The focal point of the investigation is on the design of an explainable interface that can mediate between the system and the user by providing fulfilling interactions. Users through this interface exert their sense of agency and control, while the system through the interface explains, visualizes, and provides outputs as the user intends. Design can help an AI novice interact with the system at a deeper level, and through this interaction, experience trust and reasoning for the system’s output. I am designing a visually relevant interface for an AI novice to use and recognize GAN capabilities within the boundaries of given XAI research. The outcomes of the investigation are ambitious but attainable and applicable in the near future with refinements in GAN technology.

2.3. ANNOTATED BIBLIOGRAPHY

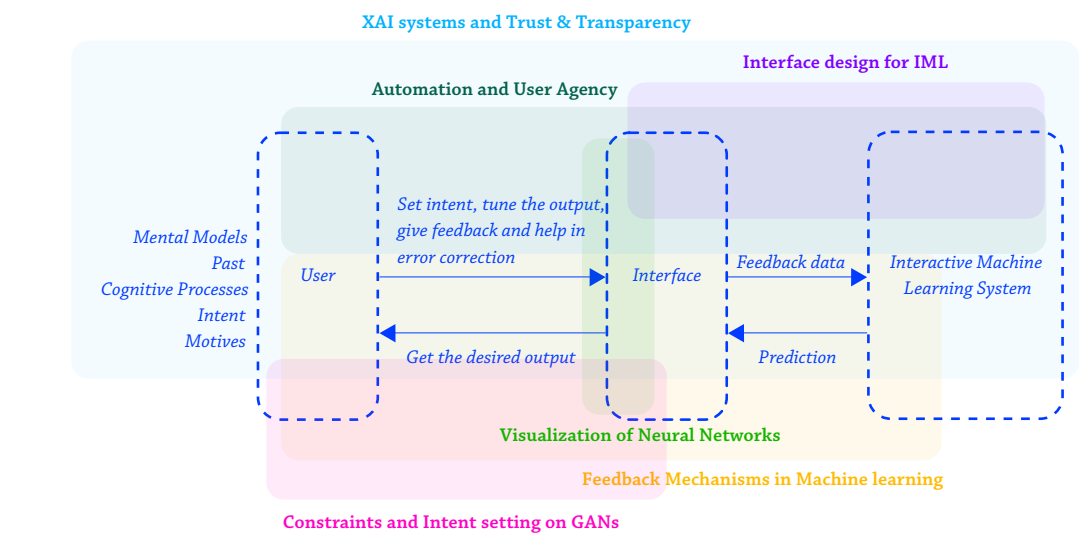


Figure 2.3.1 - Research papers referred for this research are clubbed under six different topics. This visualization shows systems’ components as covered within those six topics.

Topic 1: Interface design for Interactive Machine Learning

Interactive machine learning (iML) systems include an automated service, a user interface, and a learning component. A human interacts with the automated component via the user interface and provides iterative feedback to a learning algorithm (Boukhelifa et al., 2018). Appropriate design of interfaces is critical to the success of systems that include generative, uncertain, or predictive outputs. Interfaces form the foundation of mental models to support mental simulation and prediction for novel situations. (Browne et al., 2018). This topic provides insight on principles and frameworks in interface design in different domains of iML applications (Dudley & Kristensson, 2018; Kahng et al., 2019; Jasper et al., 2017).

Topic	Sub Topic	Title	Citation
Interface design for Interactive Machine Learning	Interface design principles	A review of User Interface Design for interactive Machine Learning	Dudley & Kristensson, 2018
	Interactive Visual representations	Interface Metaphors for Interactive Machine Learning	Jasper et al., 2017
	Interactive Visual representations	Critical Challenges for the Visual Representation of Deep Neural Networks	Browne et al., 2018
	Interface Design opportunities	User Interface Goals, AI Opportunities	Lieberman, H., 2009
	Interactive Visual representations	GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation	Kahng et al., 2019

Topic 2:- Feedback Mechanisms in Machine Learning

Feedbacks gathered from the users can help make the interactive machine learning systems better. Feedback can be in the form of the users defining their intent or explaining to the system problems in the given results or issues in the production environment (Putzu et al., 2020, Shama et al.,2018). Knowing how feedback affects user experience in a machine learning system can help design better feedback mechanisms (Tripathi et al., 2019). A part of this topic talks about

changes to the user interface that can impact the quality and quantity of feedback data and, therefore, system output accuracy (Schnabel et al., 2019).

Topic 3: XAI systems and Trust and Transparency

Greater transparency on the interface on system decisions and data collected potentially increases end-user control and improves the acceptance of complex algorithmic systems. Transparency can also promote user learning, help mitigate bias and oversights of algorithms (Springer et al., 2019). There has been a push for systems like explainable AI(XAI), adding human-in-the-loop, and progressive building of interface transparency. All of these help in building users’ trust and transparency in the machine learning system(Shih, 2018; Zhou et al., 2018; Mohseni et al., 2020). XAI aims to uncover and explain black box decisions of AI systems. This area inspects and tries to understand the steps and models involved in making decisions towards a particular system result (Wang et al., 2019).

Topic 4: Automation and User Agency

User agency refers to giving users an option to interact with the iML. This perspective integrates the human into the algorithmic loop. The goal is to use human knowledge and skills to improve the quality of automatic approaches (Holzinger et al., 2019). Giving users agency and priming them with the system’s behavior can help restore an appropriate sense of control and increase user acceptance of how the system processes (Goff et al., 2018). Furthermore, people want to interact with machine-learning systems in richer ways than anticipated, suggesting new input and output capabilities (Amershi et al., 2014). This rich collaboration between human-machine helps with problems of automation such as automation bias, occurring when human operators ignore other senses of information including their faculties, as they overly trust the automated system (Zerilli et al., 2019), and algorithmic omniscience, which means users over-accept system outputs (Hollis et al., 2018).

Topic	Sub Topic	Title	Citation
Feedback Mechanisms in Machine Learning	Feedback Mechanisms in different domains	Evaluation of Interactive Machine Learning Systems	Boukhelifa et al., 2018
	Feedback Mechanism Interventions	Shaping Feedback Data in Recommender Systems with Interventions Based on Information Foraging Theory	Schnabel et al., 2019
	Improving results through feedback	Adversarial Feedback Loop	Shama et al., 2018
	Improving results through feedback	CueFlik: Interactive Concept Learning in Image Search	Fogarty et al., 2018
	Error Correction through feedback	CueTIP: A Mixed-Initiative Interface for Correcting Handwriting Errors	Shilman et al., 2006
	Improving results through feedback	ISSE: An Interactive Source Separation Editor	Bryan et al., 2014
	Feedback type & User Experience	How Relevance Feedback is Framed Affects User Experience, but not Behaviour.	Tripathi et al., 2019
	Improving results through feedback	Convolutional neural networks for relevance feedback in content based image retrieval	Putzu et al., 2020

Topic	Sub Topic	Title	Citation
XAI systems and Trust and Transparency	Transparency through human in loop interaction	Beyond Human-in-the-Loop: Empowering End-Users with Transparent Machine Learning	Shih, 2018
	Transparent Machine learning - XAI	2D Transparency Space—Bring Domain Users and Machine Learning Experts Together	Zhou et al., 2018
	Transparent Machine learning - XAI	A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems	Mohseni et al., 2020
	Progressive building of Transparency	Progressive Disclosure Empirically Motivated Approaches to Designing Effective Transparency	Springer et al., 2019
	Transparent Machine learning - XAI	Metrics for Explainable AI: Challenges and Prospects.	Hoffman et al., 2019
	User centered XAI	Designing Theory-Driven User-Centric Explainable AI	Wang et al., 2019

Topic	Sub Topic	Title	Citation
Automation and User Agency	Human in the loop	Power to the People: The Role of Humans in Interactive Machine Learning	Amershi et al., 2014
	Human in the loop	Interactive machine learning: experimental evidence for the human in the algorithmic loop	Holzinger et al., 2019
	Automation Bias	Algorithmic Decision-Making and the Control Problem	Zerilli et al., 2019
	Algorithmic Omniscience	On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception	Hollis et al., 2018
	Users system acceptance	Agency modulates interactions with automation technologies	Goff et al., 2018

Topic 5: Constraints and Intent setting on GANs

Generative adversarial networks (GANs) can generate realistic images, videos, texts, and other kinds of media from the content used to train GAN models. They tend to reconstruct from training images. Using a new creative architecture to generate can lead to more creativity in the work and innovative forms of outputs (Sbai et al., 2018). GANs fall short in one key aspect of generation: controllability, the ability to control the semantics of the generated images in an interpretable,

deterministic manner. Adding controls allows for single variation like pose or eyebrows in a portrait(Chrysos et al., 2020). Other than establishing user control on the GANs, knowing user intent is also a requirement for a successful system. Magassouba et al. (1993) talks about ways in which multimodal language capturing from user interaction can establish user intent when the intent isn’t specified. Having users have control, motivation, and agency in the creative generation using GANs means that the originality of results would be on the author and hence can save an author from some copyright violations (Deltorn, 2017). In some of these research papers, it is interesting to note the visuals of the interface provided for user agency(Carter & Nielsen, 2017; Ghosh et al., 2019).

Topic 6:- Visualization of Neural Networks

GANs are made up of two convolutional neural networks and structurally are non-linear in structure. These neural networks are applied in a black-box manner, with no information about how they arrive at predictions available (Samek et al., 2017; Olah et al., 2020). Understanding and validating the decision process of an AI should be accessible to a user, as, with the developed recognition of the process, the user can tweak the input to get the expected goal(Olah et al., 2020). Interpretability of the system is a required tool for detecting flaws in the model and biases in the data, for verifying predictions, and for improving models(Samek et al., 2017). There are various ways to make richer interfaces with interpretability embedded in them, like via the use of feature visualization(Olah et al., 2019; Olah et al., 2020) or unit visualization (Bau et al., 2018).

Topic	Sub Topic	Title	Citation
Constraints and Intent setting on GANs	Adding constraints & control on GANs	DesIGN: Design Inspiration from Generative Networks	Sbai et al., 2018
	Intellectual Property thinking with GANs	Deep Creations: Intellectual Property and the Automata	Deltorn, 2017
	Adding Control in GANs	Unsupervised Controllable Generation with Self-Training	Chrysos et al., 2020
	Setting user intent	Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target-Source Classification	Magassouba et al., 2019
	Adding constraints & control on GANs	Constrained Generative Adversarial Networks for Interactive Image Generation	Heim, 2019
	Adding Control in GANs	Using Artificial Intelligence to Augment Human Intelligence	Carter & Nielsen, 2017
	Adding Control in GANs	Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation	Ghosh et al., 2019

Topic	Sub Topic	Title	Citation
Visualization of Neural Networks	Heatmap visualization	Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models	Samek et al., 2017
	Feature Visualization	The Building Blocks of Interpretability	Olah et al., 2020
	Feature Visualization	Feature Visualization	Olah et al., 2019
	Interpretable Unit Visualization	Gan Dissection: Visualizing and Understanding Generative Adversarial Networks	Bau et al., 2018

2.4. DEFINITION OF TERMS

Generative Adversarial Network (GAN): A GAN is a machine learning model in which two neural networks compete with each other to become more accurate in their predictions.

Interactive Machine Learning (iML): iML system comprises an automated service, a user interface, and a learning component. A human interacts with the automated component via the user interface and provides iterative feedback to a learning algorithm (Boukhelifa et al., 2018).

Human in the Loop: Interactive Machine Learning looks for “algorithms which interact with agents and can optimize their learning behavior through this interaction – where the agents can be humans”. This perspective basically integrates the human into the algorithmic loop (Holzinger et al., 2019).

Explainable Artificial Intelligence (XAI): An XAI system can be defined as a self-explanatory intelligent system that describes the reasoning behind its decisions and predictions. The AI explanations (either on-demand explanations or in the form of model description) could benefit users in many ways such as improving safety and fairness when relying on AI decisions (Mohseni et al, 2020).

Relevance feedback (RF): RF is a mechanism by which users flag search results that are relevant to the current search. By doing so, users can refine the scope of their search without explicitly describing what information they are seeking (Tripathi et al., 2019).

Machine Learning (ML): Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed.

Machine learning model: A machine learning model is a file that has been trained to recognize certain types of patterns. You train a model over a set of data, providing it an algorithm that it can use to reason over and learn from those data.

Neural Networks: A neural network is a network or circuit of neurons, or in a modern sense, an artificial neural network, composed of artificial neurons or nodes.

Black box: The term “black box” describes a system with clearly observable inputs and outputs, but with inscrutable internal processes (Browne et al., 2018).

Interpretability: In context of machine learning interpretability is the degree to which a human can understand the cause of a decision.

Interpretable AI: Inherently human-interpretable models due to their low complexity of machine learning algorithms (Mohseni et al, 2020).

A GAN system: A system that has different GAN models for different tasks.

2.5. ASSUMPTIONS AND LIMITATIONS

Limitations

The GAN technology that I am designing for is not currently fully functional or stable. However, my investigations grow increasingly useful and potent in the future as newer and more stable systems come into existence. The research pointed me in the direction where users can constrain or change outputs, and as a designer, I can intervene, but this does not ensure that the system technology described will work as expected. Another limitation of this investigation is that I lack expertise in the field of neural network technology. The GAN system used here is limited to image based models. Understanding the output is difficult because of the unpredictability of GAN outputs. In addition to unpredictability, the system generated results can be unstable, biased, discriminatory, and harmful because of problematic biased models. For this investigation, I will only be addressing the trust, visualization, and interpretability aspects of the interface design for GANs. There is a high likelihood of bad actors (e.g., creating fake images for political campaigns that can accelerate harm) using these generative technologies, which makes a case for examining the ethics of this system, but given the scope of the project, I will not address this issue.

Assumptions

I make several assumptions for this investigation. I assume that the AI novices want agency over the output in the hope and expectation of getting a better result quickly from the system. Also, a novice willingly spares some time to give feedback and set intent on the GAN system. Another of my assumptions is that the user is a novice with no knowledge of machine learning and data privacy. The user interacting with GAN technology is morally conscious and uses it sincerely, without trying to cause harm. The user is curious to know about the generated output and is questioning the given system actively. The data on which the GAN models is trained and tested is diverse and free from preprocessing bias. Additionally, in no way is the system using data for biometrics and facial recognition. All data provided by the user is locally stored, with the users having full authority. The final assumption is that all images that a user manipulates with GANs are appropriately credited and attributed.

2.6. PRECEDENTS

As a part of my investigation, I found and separated existing applications into three broader categories: agency, visualization for interpretability, and explainable interfaces. These categories also come up again as part of my studies and can help understand the umbrella structure of my build up.

Agency

LOBE.AI (Website: www.Lobe.ai, Figure 2.6.1, 2.6.2)

Lobe.ai is only available on beta, but there are already available work examples with lobe.ai and some interface examples which look easy to connect and understand. A user can use Lobe.ai to train models by making connections on the interface. For instance, the petal-generator example generates realistic petals after learning from many petal images. It makes it easy for even a non-expert to start training their own data set. Of note is the given agency and simplicity of the interface.

ANTHROPICS (Website: <https://www.anthropics.com/portraitpro/>, Figure 2.6.3, 2.6.4, 2.6.5)

Anthropics interface and the controls are very user friendly even though the software and its label aren't and can be termed discriminatory, especially the Portrait Pro that as soon as it detects the face terms it as either a female or a male. Interface components like sliders that allow for controlling outputs are present and can be used for the purpose of this investigation. The whole system isn't that interpretable. Anthropics on the webpage say it uses AI, though it does not specify the kind of AI. One can see the software working but still would want to know what each control means, why the slider ranges are as they are:- some sliders starting in the middle while others from the left end, and do default make the skin lighter? Understandably, Anthropics- Portrait Pro is a simple photo editing tool, while the GANs are complex and require a richer interaction-based interface. Interpretability and retractability will become essential, especially when generating images from scratch.

Another part of the Anthropics is landscape pro(Figure 2.6.5). It allows a user to define an area and make changes according to the selection. The software then adaptively gives options to make changes to the chosen area.

RUNWAY ML (Website: <https://app.runwayml.com/home>, Figure 2.6.6, 2.6.7)

Runway ML is an online tool for managing machine learning experiments and models for those experiments. It also supports an interactive interface for data scientists to interact and train models. The software has an expert user base, with users well versed in generative machine learning technology. It allows expert users’ agency over the data and its visualization through the interface.

LUMINAR (Website: <https://skylum.com/luminar>, Figure 2.6.8)

Luminar is versatile and highly interactive. Luminar is mostly for photo editors and is a tool that can be used alongside photo editing tools like lightroom through an API. The automated facilities provided by Luminar make it easy to use. Luminar asks users for feedback on the quality of output.

PIX2PIX (Website: <https://affinelayer.com/pixsrv/>, Figure 2.6.9)

Pix2pix requires user input to generate realistic images of cats and objects. The outputs aren’t perfect. The interface is a one-click generation with hardly any controls or options provided to make it easier for a user to change their input or provide assistance in their input sketch. Pix2pix is not a generative model, which means it does not have a latent space or a corresponding space of natural images. Instead, there is a neural network, called, confusingly, a generator –that takes as input the constraint image and produces as output the filled-in image.

MAKE GIRLS MOE (Website: <https://make.girls.moe/#/>, Figure 2.6.10)

Make Girls Moe is an anime face generator tool. It does not collect user data or even asks users to give feedback based on generated images. This software only generates girl anime characters. There is no extra user agency given other than choosing an option.

MAGENTA & MORE DEMOS FROM GOOGLE (Website: <https://magenta.tensorflow.org/studio>, Figure 2.6.11, 2.6.12)

The Magenta studio has audio and music-related machine learning models that can generate notes that are likely to follow your drum beat or melody. Give it an input file, and it can extend it by up to 32 measures. Magenta is an easy tool for musicians to use and access without understanding the technology behind it.

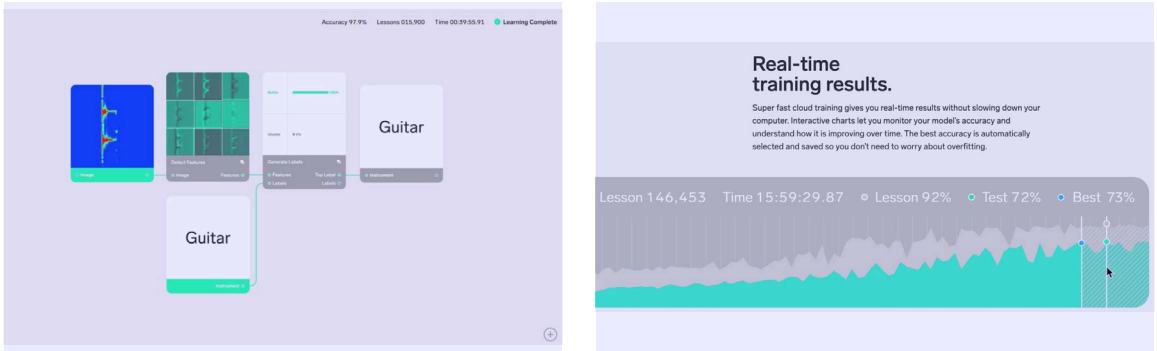


Figure 2.6.1- Lobe.ai interface providing users agency
Figure 2.6.2- Lobe.ai allows looking at training data results and controlling your models’ agency

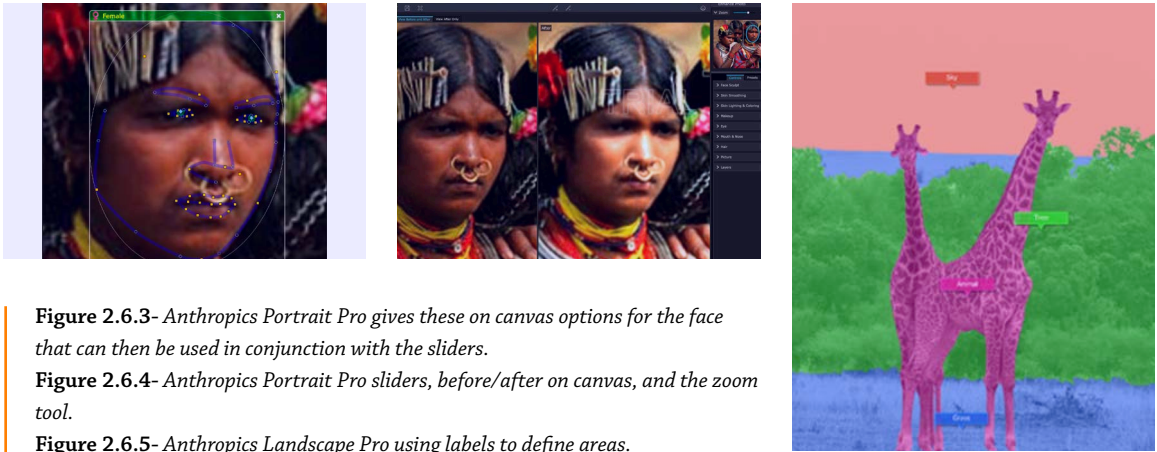


Figure 2.6.3- Anthropic Portrait Pro gives these on canvas options for the face that can then be used in conjunction with the sliders.
Figure 2.6.4- Anthropic Portrait Pro sliders, before/after on canvas, and the zoom tool.
Figure 2.6.5- Anthropic Landscape Pro using labels to define areas.

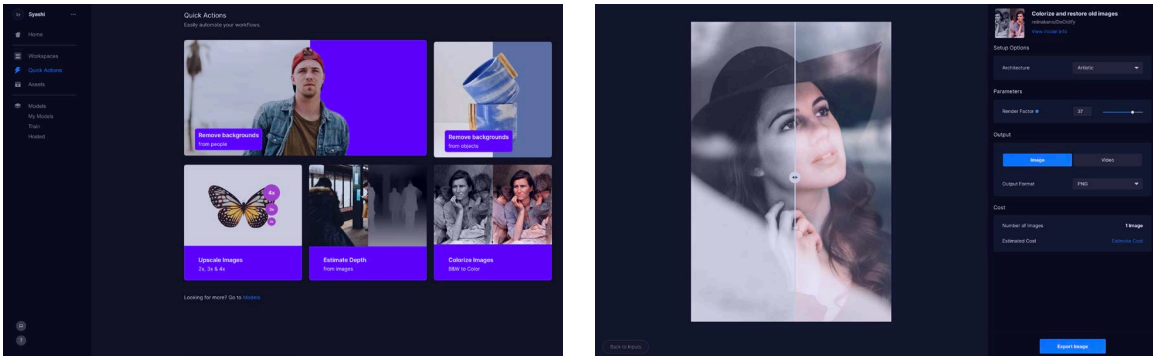


Figure 2.6.6- Runway ML allows users to choose tools according to their intent.
Figure 2.6.7- Hosting allows users to use the models they trained or use somebody else’s models and allow visitors to interact with hosted models.

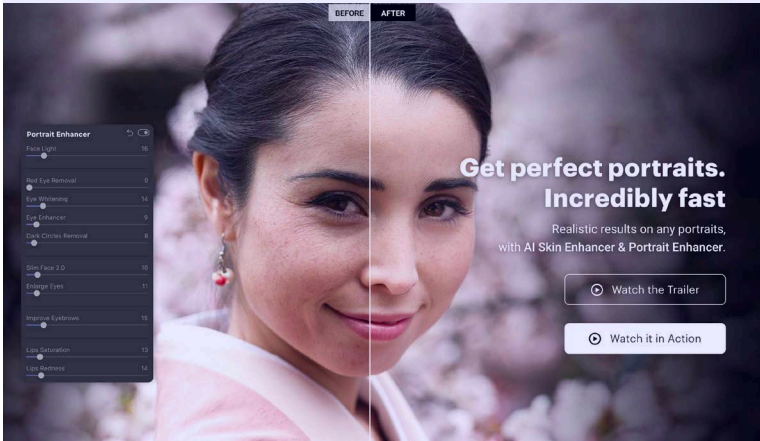


Figure 2.6.8- Luminar shows the before and after and the interface along with the result on their website

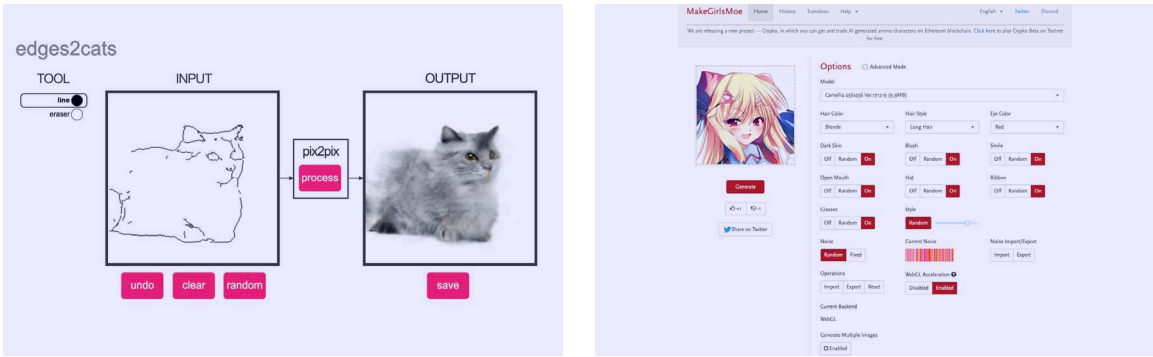


Figure 2.6.9- Pix2pix generates realistic images of the objects based on the drawings. This specific model edges2cats can generate realistic images of cats. There is just one simple button to process thereby limiting user’s agency.
Figure 2.6.10- MakeGirlsMoe is a fun interactive anime generation tool. It gives limited interface controls.

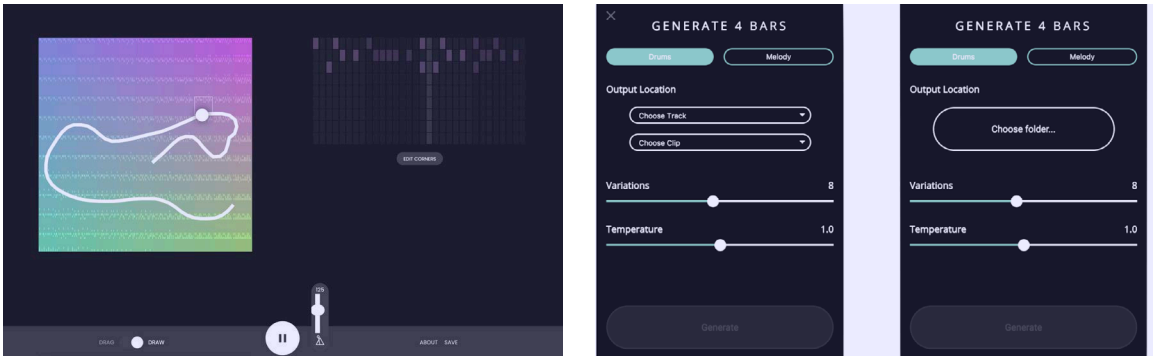


Figure 2.6.11 & 2.6.12- Magenta studio gives flexibility to a person to generate music. Although many details and user mental models are abstracted here.

Visualization for Interpretability

GAN Lab (Website: <https://poloclub.github.io/ganlab/>, Figure 2.6.13)

GAN Lab visualizes the GAN functionality. Users can understand what the model is doing, but the visualization isn't very approachable for a person with less to no knowledge of GAN.

DISTILL (Website: <https://distill.pub/>, Figure 2.6.14)

Distill is a new online interactive journal dedicated to an interactive explanation of machine learning. It is easy to approach Distill and dive into the concepts. Distill's approach gives precedence to how visuals can explain the working of the model on the interface interactively or via storytelling.

ConvNetJS (Website: <https://cs.stanford.edu/people/karpathy/convnetjs/>, Figure 2.6.15)

ConvNetJS provides interactive demos for the convolution network problems. GANs are also convolutional networks. Looking at some of the classification demos explains a lot about the internals of machine learning.

R2D3 & Explorables & Stitch Fix (Website: <http://www.r2d3.us/>, <https://explorabl.es/>, <https://algorithms-tour.stitchfix.com/>, Figure 2.6.16, 2.6.17, 2.6.18)

R2D3's creator Tony chu uses interactive storytelling to explain ML. Storytelling helps form users' mental models of the given machine learning model. Similar to R2D3 is a stitch fix interactive algorithm tour of their store model. Stitchfix's visualization goes in-depth into calculations that might steer a non-technical user away. This layer of transparency is helpful when trying to understand machine learning models

FiveThirtyEight (Website: <https://projects.fivethirtyeight.com/2020-election-forecast/>, Figure 2.6.19)

FiveThirtyEight's interactive visualization excellently shows the election forecast with the surrounding uncertain factors. FiveThirtyE-

ight lets a user choose one variable and see the impact of that variable on the election results. This way of visual interactivity helps in explaining the situation as is to the user without clouding the uncertainty portion of the data.

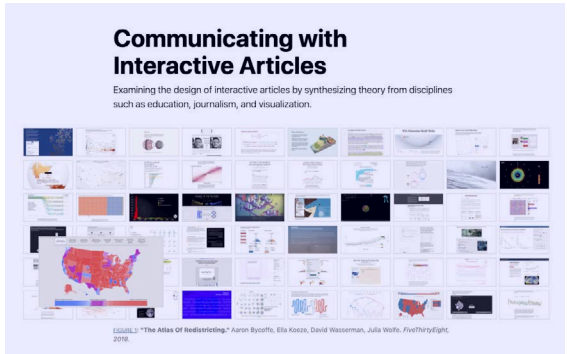
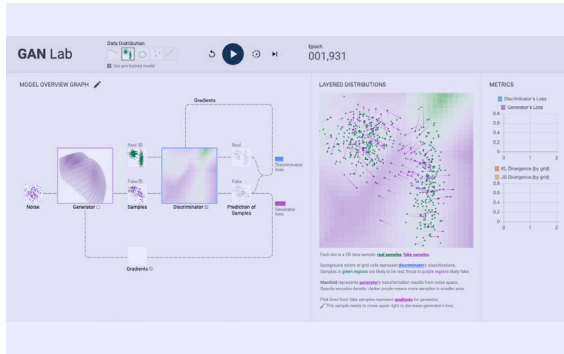


Figure 2.6.13- Gan Lab visualization for a specific data type one chooses

Figure 2.6.14- Distill Publication tries to explain machine learning black box through interactive visualizations

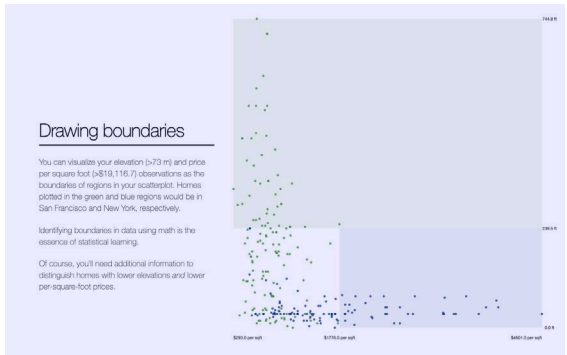
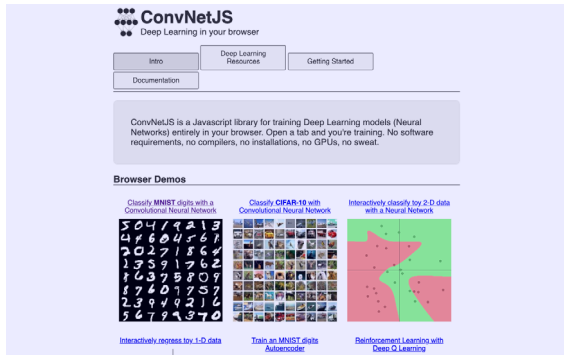


Figure 2.6.15- ConvNetJS helps to see and learn models working on the web using JS.

Figure 2.6.16- R2D3 as a storytelling explanations

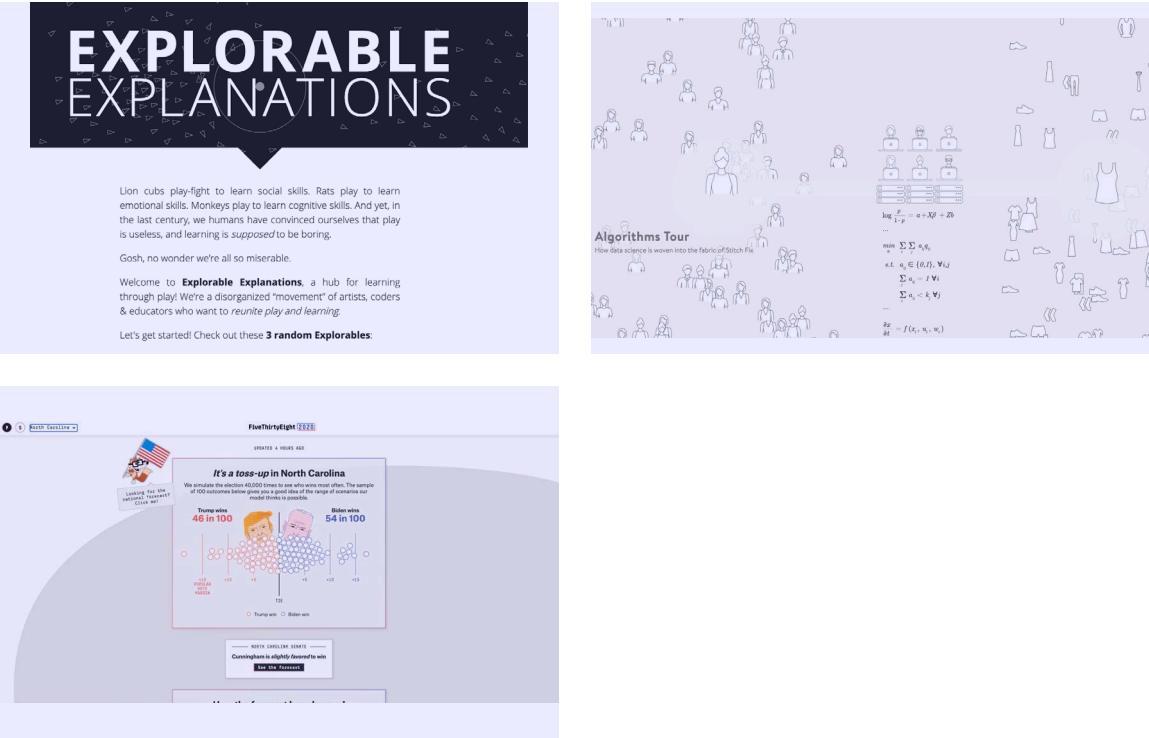


Figure 2.6.17- Explorable explanations
Figure 2.6.18- Stitchfix algorithm visualization
Figure 2.6.19- Five three Eight's Interactive charts explain the statistics and impacts of uncertainty to the user very well.

Explainable Interfaces

PROJECTS BY IF (Website: <https://www.projectsbyif.com/>, Figure 2.6.20)

Projects By If studio specializes in ethical and practical design. Sarah Gold, the founder of Projects By If studio, makes theoretical ideas of trust and explainability of systems come to reality through interface design patterns.

NMAIL (Website: <http://nmail.kaist.ac.kr/wordpress/index.php/category/research/neuro-inspired-intelligence/explainable-ai-interfaces/>, Figure 2.6.21)

NMAIL lab develops human-friendly AI interfaces that provide explanations to users.

PAIR by Google (Website: <https://pair.withgoogle.com/chapter/explainability-trust/>, Figure 2.6.22)

PAIR by Google provides key considerations for explainable AI systems alongside worksheets to research user trust for these systems. PAIR talks in the chapter about how explainability can help build trust and evaluation of the same.

FIDDLER (Website: <https://www.fiddler.ai/explainable-ai>, Figure 2.6.23)

Fiddler provides solutions for companies to help explain their AI model outcomes, which can lead to more trustable AI. The website itself doesn't lend trust for the services as it hides crucial information.

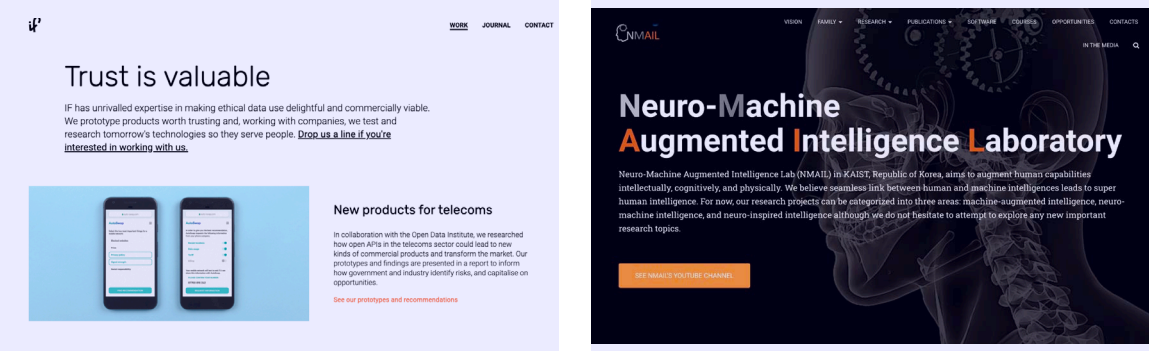


Figure 2.6.20- ProjectsbyIf
Figure 2.6.21- NMAIL Lab builds human-friendly interfaces that provide explanations

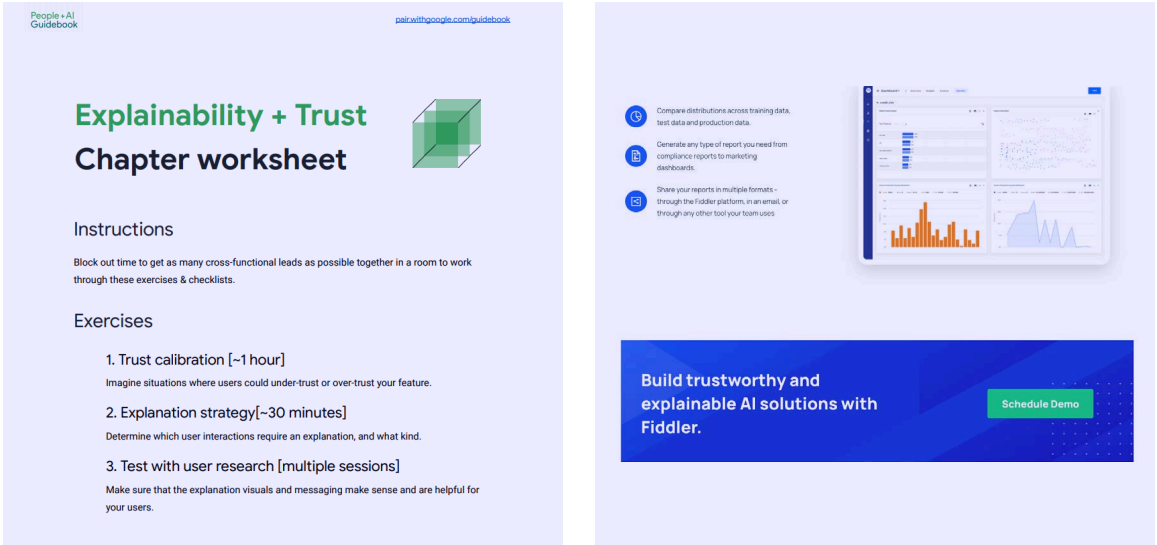


Figure 2.6.22- PAIR by google provides worksheets to help to imagine situations that could induce trust or mistrust and then resolving them by user research.

Figure 2.6.23- Fiddler’s website shows their methods of explaining the AI models.

INVESTIGATION PLAN

- ∞ 3.1 Conceptual Framework
- ∞ 3.2 Research Questions
- ∞ 3.3 Investigation Model
- ∞ 3.4 Scenario

3.1. CONCEPTUAL FRAMEWORK

To visualize this area of interactive machine learning—explainable AI(XAI), human-in-the-loop, and generative adversarial networks, I diagrammed frameworks based on my initial research. . As a framework for analyzing goal-oriented interaction of users with the environment through activity, I used an activity theory diagram as adapted by Davis (2012), and to position my investigation concerning digital technologies, I found digitally mediated activity theory (Blayone, 2019) (Figure 3.1.1). A user — in this case, an AI novice — has past experiences, perceptions, intent(motive), ways of reasoning, and emotions. The user is given agency to interact with the explainable AI system via interface and performs actions that facilitate engagement with the explainable interface in order to fulfill the user’s goals of increased efficiency, better quality, and intended output. Action or interaction with the system, as per activity theory, happens between the inputs or outputs and the user, for a user to make sense of the system and its output, give feedback, or control variables to get the desired results. Feedback can be in the form of relevance feedback, error correction, or quality assessment, depending on the machine learning model used by the user.

Further, the design of the XAI interface matters for an AI novice as intuition and interface can enhance the user experience of the system by increasing system accountability and improving end-users’ comprehension and reliance on the intelligent systems. As the designed interface is a critical aspect of user interaction with the system, I used XAI design and evaluation framework goals (Mohseni et al.,2020) in this framework.

The provided framework builds relationships between interface, user, cognition, agency, XAI, and feedback components and establishes these connections in the user-centered design domain. This framework considers the position of the user as a novice, suggests ways to engender appropriate user trust, develop user mental models of the system, extend user control, and enrich human-machine collaboration through the design of an interface.

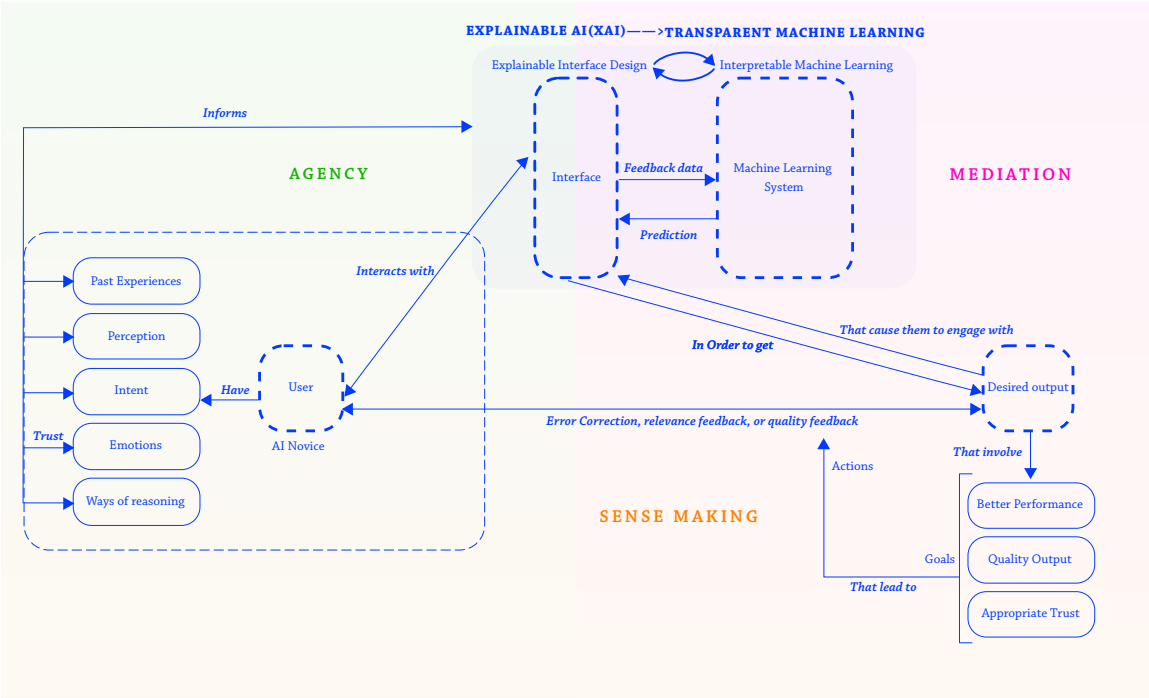


Figure 3.1.1- Conceptual Framework

3.2. RESEARCH QUESTIONS

This investigation is conducted in four exploratory studies, each of which informs a primary research question concerning a GAN system’s interface features. The four studies are individually guided by the four corresponding subquestions. These subquestions push the design inquiry forward into actionable explorations.

Primary Research Question

How can the interface design of a GAN system facilitate an AI novice’s interpretability, agency, and trust to build user mental models when using a GAN system?

Subquestions

How can the design of digital interface interactive features solicit user inputs to help an AI novice practice agency and control?

How can network visualizations reveal internal decisions of the system to help an AI novice interpret the GAN system and its decisions?

How can textual explanations communicate varying levels of information to facilitate AI novice’s trust in the GAN system?

How can feature visualization, textual explanations, and interface features together build mental models in AI novices so that they might practice agency in such a way that the output matches their intentions while allowing for exploration?

3.3. INVESTIGATION MODEL

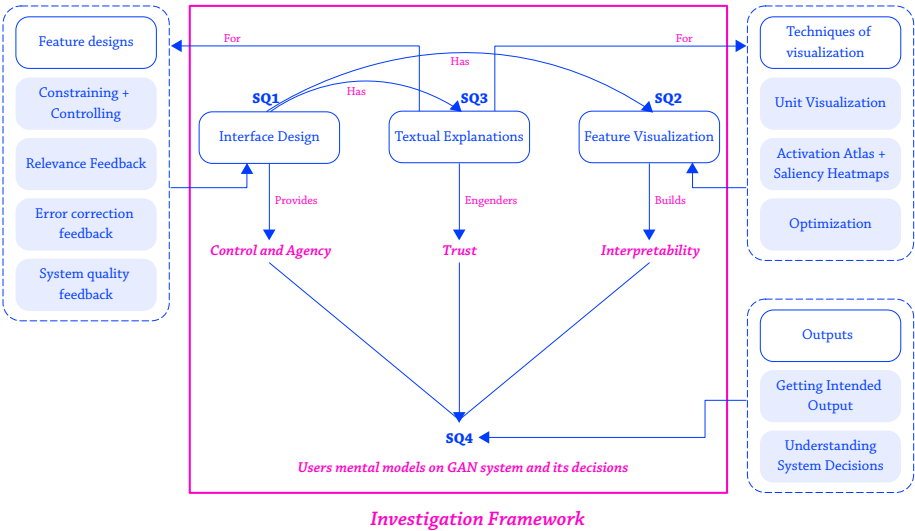


Figure 3.3.1- Investigation Framework

In my investigative model (Figure 3.3.1 & Figure 3.3.2) I am initially exploring three relationships: an interface’s features design providing control and agency (Study 1); a feature visualization building interpretability (Study 2); and textual explanations engendering trust (Study 3). In all studies an AI novice is imagined as the user. Finally, I demonstrate how a combination of interface design, explanations, and network visualizations together help users build their mental models on the GAN system and its decisions. Subquestions 1-3 are distinct studies while Subquestion 4 ties Studies 1-3 together. In my conceptual framework, I synthesized literature on agency practices (both control and feedback), interpretability, and trust. I seek to utilize those methods and explore various possibilities in the realm of interface design and visualizations that can then be evaluated by users via testing.

Sub-questions focus					
	A	B	C	D	
SQ1	Interaction features design	Constraining and controlling	Relevance feedback	Error correction feedback	System quality feedback
SQ2	Feature visualization	Unit visualization	Saliency heatmaps + Activation Atlas	Optimization	Activation Atlas + Units visualization
SQ3	Textual explanations	Building trust through explanations			
SQ4	SQ1+SQ2+SQ3	All together			

Figure 3.3.2- Investigation Framework Table

3.4. SCENARIO

Rose Wayne is a game designer and 3D artist working at a small studio in Raleigh. She is working on a fantastical game design project for kids and adults. She has to create scenes for the game’s initial level and characters or avatars for the whole gameplay. The game characters are realistic-looking animal chimeras. Designs for the game have to be distinctive from the ones already existing in the game market. Her initial sketches of chimera were illustration-based, and for the following iterations, stakeholders suggested she have a more realistic-looking chimera. She is limited in terms of ideas of what other kinds of characters can be generated and is looking for more examples of chimeras with good textural rendering. Rose wants to follow the studio’s aesthetics and style for game landscape backgrounds. As this is a fantasy game, she wants some landscape inspiration that can get her to explore in directions that her team had not previously considered.

Investing in making a character or landscape is time-consuming, so she turns to the internet for inspiration and ease of prototype. There Rose finds online APIs that can develop fantasy-based landscapes and character designs within minutes.

She uses a character generator API with Photoshop application that shows already created and editable segmentation maps to generate a mix of hyena and jackal, but she finds that the generated output is limited in terms of flexibility to change. It is hard to predict the system output if the 2D segmented shape is of animals that look similar. Rose has the flexibility to modify hair, eyes, and body parts, but the results don’t look photorealistic as soon as she edits or draws over them. Rose then looks for more options and variety in the generations from this API that can serve as a visual influence. This online software restricts multiple output generations or controlling ability in terms of the animals to crossbreed, randomize, or even interpolate. Rose dislikes the look of the chimera generated due to limited control in changing details, specifying constraints and intentions, and explainability of system decisions. Rose wants to understand why for one particular segmented drawing that Rose made, the system generated a gorilla or a monkey and not a fox head, as knowing this will help her tweak results to her intentions. She thinks that this API’s capabilities can be used to prototype multiple chimeras and user test prototypes.

Scenario 1

PERSONA: ROSE WAYNE
Age: 39
Profession: Character designer, game designer, and 3D artist

Rose wants to create animal character design and wild fantasy landscapes for the kids game design. This is for the first round of prototype testing and idea geaneration

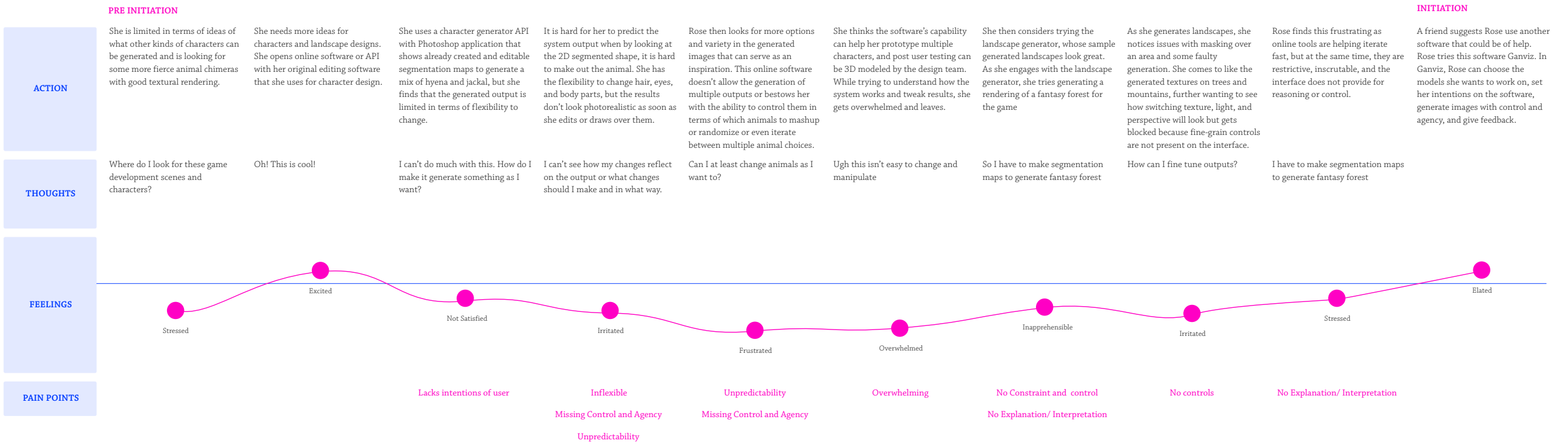


Figure 3.4.1- user journey map before using the system

Scenario 1

PERSONA: ROSE WAYNE
Age: 39
Profession: Character designer, game designer, and 3D artist

Rose wants to create animal character design and wild fantasy landscapes for the kids game design. This is for the first round of prototype testing and idea geaneration

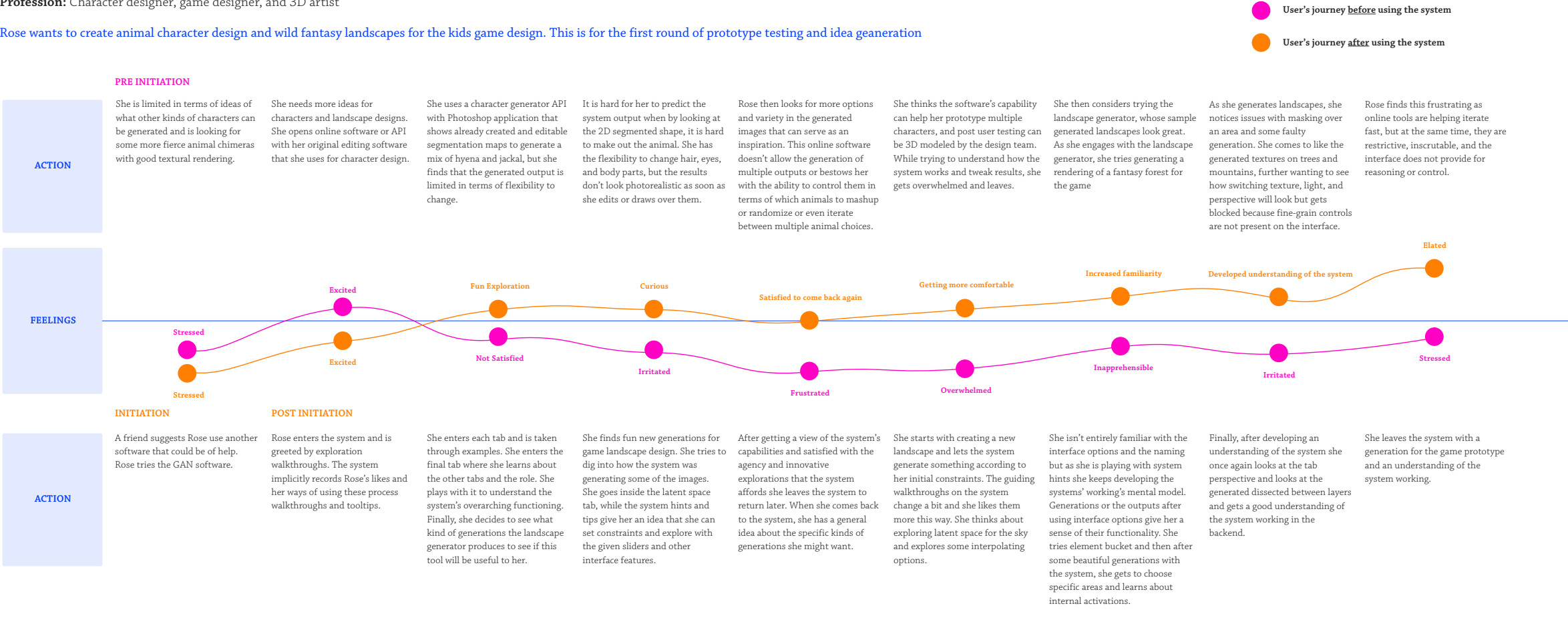


Figure 3.4.2- Comparing user journey maps of before using the system with after using the system

Afterward, prototypes can then be 3D modeled by the design team. While trying to understand how the system works and tweak results, she gets overwhelmed and leaves.

She then considers trying the landscape generator, whose sample-generated landscapes look great. As she engages with the landscape generator, she tries generating a rendering of a fantasy forest for the game. These generated images help Rose with some prototypes for user testing and idea generation. Rendering an entire landscape is a month’s work, and these generations can save crucial time. As she generates landscapes, she notices issues with masking over an area and some faulty generation. She likes the generated textures on trees and mountains, so she wants to see how switching texture, light, and perspective will look. However, she gets blocked because fine-grain controls are not present on the interface. There is no provision for adding or removing elements from the generated landscape. Rose finds this frustrating as API tools are helping her to iterate quickly, but at the same time, they are restrictive, inscrutable, and the interface does not provide for reasoning or control. Some outputs are impressive, while others are visually low quality and outrageous. She finds it hard to understand the operability of systems without getting overwhelmed (look at Figure 3.4.1).

A friend suggests Rose use another software that could be of help. Rose considers trying this API Ganviz. In Ganviz, Rose can choose the models she wants to work on, set her intentions, generate images with control and agency, and give feedback. With this API, Rose can interpolate different images, develop an understanding of those generated, and notice internal functioning. She can swap objects from dataset units based on her needs, add and remove elements from generations, and see the impact of her feedback on the outputs and visually understand how the system generates the output. There are multiple generations or few, according to her requirements. Within a couple of days, Rose understands the software and uses it to her benefit (look at Figure 3.4.2).

STUDIES

- ∞ 4.1 Design of digital interface features
- ∞ 4.2 Visualization of neural networks
- ∞ 4.3 Textual Explanations
- ∞ 4.4 XAI interface design

4.1 DESIGN OF DIGITAL INTERFACE FEATURES

Question

How can the design of digital interface interactive features solicit user inputs to help an AI novice practice agency and control?

Study 1

This study explores features for an interface design that helps extend agency to an AI novice to control inputs, provide feedback, and eventually pass control of the system outputs to the user. These designs can help evaluate whether an AI novice develops agency through these available interface features. This study was conducted in four phases and utilizes an introductory feature exploration of the interface.

Study 1a) Constraining & controlling input exploration (via the usage of input agency forms like sliders, uploadable/searchable images, and gestures).

Study 1b) Relevance feedback on the generated outputs.

Study 1c) Error correction feedback on the generated output by marking areas and additional description option provision.

Study 1d) System Quality feedback on the generated quality of output and looking at implicit and explicit feedback opportunities.

I began my visual explorations with an exploration into a more basic interface look. A user landing on the interface would expect some simple interactive options like canvas, tabs, and workspace. These need to be designed to be accessible to the novice user. Nomenclature for all interface options will be investigated again and concluded in Study 3. Within these visual explorations of the elementary interface, iterations helped decide the hierarchy that better-suited user needs. The final look of the landing screens are shown in Figure 4.1.1. Eventually, I explored ways an AI novice can put forth their intention or add inputs(Figure 4.1.2). As I proceeded through Study 1.1, Study 1.2, Study 1.3, and Study 1.4, the interface designs became more refined. Additionally, I referred to existing precedents that would help me build these interface controls. The high-fidelity wireframe designs can be accessed [from references section 6.1](#).

Segmentation- In digital image processing and computer vision, image segmentation is the process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects). The goal of segmentation is to simplify and/or change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. More precisely, image segmentation is the process of assigning a label to every pixel in an image such that pixels with the same label share certain characteristics.

Latent Space- Latent Space simply means a representation of compressed data. Whenever we graph points or think of points in latent space, we can imagine them as coordinates in space in which points that are “similar” are closer together on the graph.

Relevance Feedback(RF)- RF is a mechanism by which users flag search results that are relevant to the current search. By doing so, users can refine the scope of their search without explicitly describing what information they are seeking (Tripathi et al., 2019).

Basic interface controls:

∞ **Canvas:** There can be two canvases, one for editing and the other for output. Alongside the editing canvas is a canvas toolbox, ideal for editing purposes. Many of the on-canvas interactions and tools used have precedents present in 2D or 3D editing software.

∞ **Settings Pane:** Depending on the intention that the user sets, the settings pane will change. These options extend agency to the user to control generations.

∞ **Tabs:** My presented scenario explores chimera and landscape generation, thus, I explored different tab options and placements for both of these contexts. Finally, I decided on keeping three tabs as segmented, rendered, and latent space.

A few precedents for segmented maps present online are Chimera painter and Gaugan by Google and Nvidia, respectively. I also looked at Google Blog that shows that character representations and eventual image generation become easier and understandable by using segmented maps. Both the precedents let users generate images with one click, have a rudimentary interface, and control options presented to the user. As a result, I explored more on-canvas options and different tabs like the rendered tab that serves as an alternate option to creating images. Segmenting and rendering tabs both impart additional flexibility to the user to control generations.

User Input Controls:

An important part of interface exploration is thinking about how the user intent will be captured and used to ease user interaction with the interface. I considered these three tools, which can help capture user intent. The feasibility of each user input prompt was evaluated by finding precedents available.

∞ **Text prompts:** Functionality and playing with DALL.E guided some of the explorations of textual prompts and ways a user can easily set an intention for the use and get interface options that assist the user in their work.

Error Correction- Correcting errors in machine-generated images.

Chimera- Chimeras are animals composed of cells that originate from two (or more) different species. In the research lab, chimeras are created by introducing cells from one species into the developing embryo or fetus of another.

Open.ai has trained a neural network called DALL.E that creates images from text captions for a wide range of concepts expressible in natural language.

- ∞ **Voice prompts** : On similar lines to text prompts, NLP speech recognition can help tap into the opportunity of converting speech to text and then analyzing the textual expressions via DALL.E to set user intent.
- ∞ **Similar image prompt or Uploading image prompt**: There exist many GAN models like makeup transfer, style transfer, and content transfer. These models pick from one image and transfer content, style, or a variable to another.

Ideation Study 1a, 1b, 1c, and 1d

Study 1a) Constraining & controlling input exploration

Scenario:

Rose is a video game designer working on a fantasy game. She wishes to generate a chimera that flies and survives on both land and water. She intends this creature to be an avatar that evolves as users’ skills improve and advance through game levels.

Solution ideation:

Rose lands on the interface and enters the prompt for the chimera she wants to generate. The GAN system tracks Rose’s inputs - explains what it interprets from Rose’s prompts, and begins generating relevant imagery on the canvas. Rose can then correct or add to the GAN systems’ interpretation. Based on her suggestions, the system will respond accordingly and repeat the process of producing images.

Rose plays around with on-canvas buttons, anatomy-based settings, and slider options provided. She then decides to edit some body parts of the chimera, where she discovers the Segmented tab option, which allows her additional customization. Rose makes edits in the segmented tab to the chimera and succeeds in getting a generation as per her requirements. She decides to adjust the character position and places it in a different background setting. After she is satisfied, she saves her work.

Artbreeder, an online tool, offers a feature that traces the genealogy of an image. This genealogy can help trace image origins, and understand user’s wants and generate imagery from latent space accordingly.

Explorations:

Refer to [video 4.1.1](#) (check references 6.1) and figure 4.1.3 & 4.1.4

- ∞ **On-canvas options**: An ease of controlling generations directly from the canvas makes it easy for the user to work without getting distracted. Additionally, users with an apple pencil or pen tablet can select body areas with precision. On-canvas capabilities present for this exploration are

- Removing, adding, and categorizing body parts
- 2D to 3D and vice versa
- Drawing and erasing body parts
- Selecting body parts and changing positions
- Layering for segmented map

- ∞ **Settings Pane**: For this particular scenario, Rose can play with anatomy, lighting, positioning, and texture settings. The UI for the settings pane is adaptive and changes based on on-canvas UI manipulations, making controlling image generations easier. For example, when the user makes forward limbs selections on the on-canvas, options open up forward limb selections in the anatomy section (Figure 4.1.5). Rose can easily edit from both on-canvas and the settings available. The setting options explored here are

- Anatomy
- Lighting
- Positioning
- Texture

Study 1b) Relevance feedback on the generated outputs

Scenario:

Rose creates another project. She wants to generate a fantasy background landscape for the chimera. Rose has become accustomed to the interface and is aware of the user agency offered by the GAN system to vary the outputs, but she expects more variety for this particular task. She wishes to create different iterations and interpolations of the fantasy landscape and select those relevant to her.

Solution ideation:

Rose types in a prompt for a fantasy-based landscape, but because adding a description for a landscape can be difficult, she awaits generated results from the GAN system to explore. Afterward, Rose selects one of the images and then edits the image in segmentation mode that serves as a base for the system to understand the kind of generations Rose is imagining. Then the system continues to generate different landscapes based on a segmentation map. Rose also has the option to load more and view the root visualization, or a matrix of images produced from latent space near and similar to the selected landscape image. After choosing one image that closely aligns with her imagination, she can then tweak controls to edit the image. Each new landscape produced is created in real-time based on her feedback. Rose will continue to select the landscapes that are the most relevant to her and the system will generate varied landscapes until Rose is happy with her selection.

Explorations:

Refer to [video 4.1.1](#) (check references 6.1) and figure 4.1.6 & 4.1.7

∞ **Settings pane:** The settings pane for the landscape generator is different from that of the character generator. I did a couple of iterations to decide the minimum settings requirements for the study on relevance feedback. Similar to study 1, crossbreed here encompasses interpolating images and selecting relevant ones (Figure 4.1.8). Previously for chimera generators, it was only crossbreeding animals or fantastical creatures, but the landscape is

more convoluted and indescribable in words and needs more visuals to showcase what a user has in mind. It is important here to add uploadable images and interpolation areas.

Crossbreed: Options under crossbreed are interpolation, similar or different, and relevant imagery bookmarks. These options were thought of and combined under the banner name “crossbreed” after many explorations. The idea behind this is that a user can provide relevant imagery and interpolate them together. Users can choose between their inclination for content or style from an image, and the system generates accordingly. A user can let the GAN system generate something similar or different from its latent space. Finally, relevant imagery bookmarking serves both as an implicit and an explicit way of gathering relevance feedback.

Feature edit: Colors and natural element features in a landscape are editable, and I explored some of these options in the feature edit section (Figure 4.1.8).

∞ **Relevance feedback:** Recommendation systems gather implicit feedback from the user to recommend services or products that closely resemble their wants. Similarly, in a way, I was finding ways to collect relevant feedback from the user on the GAN system as this helps both the user and the GAN system. I explored:

Similar or different: The GAN system generates a different or a similar image to the one on the canvas according to the user’s setting. This further helps gather implicit relevance feedback.

Relevant imagery: Relevant imagery has precedent in Adobe Creative Suite’s tool swatches where relevant imagery can be bookmarked and later recalled for another project.

Explore Matrix: A matrix visualization provides a look into the latent space, where a user can look at all generations corresponding to a particular setting.



Root visualization: Root visualization can help trace image creation and origins. It is another way for a user to explore and find relevant images and implicitly give feedback.

Study 1c) Error correction feedback on the generated output.

Scenario:

Rose has many landscape images and 3D models from the studio’s previous game releases. She wants newly generated images to align closely with the studio’s design aesthetic and would love it if this GAN system can provide possible exploratory directions that her team hadn’t previously considered.

After generating landscapes, Rose changes the mode from 2D to 3D to play around with different angles and parts that could be useful for the game image background prototype. Some of them have blurred or cut off areas. Although Rose likes one of the system-generated landscape images, she is not able to use the image due to its poor image quality.

Solution ideation:

Rose starts by switching the mode for the landscape from 2D to 3D. As she turns the angles and moves positions on the canvas, she notices some areas are unsuitable or left blank. Rose uses the option of correcting errors by marking them. She also gets to upload images of landscapes that are similar to the generations she wants. These chosen images are then later used to train future models, which helps Rose prototype better in the future. Sometime later, the GAN system gets updated with re-trained models, and she can utilize these new images. Additionally, Rose can check her error correction feedback anytime she likes and make adjustments accordingly.

Explorations:

Refer to [video 4.1.1](#) (check references 6.1) and figure 4.1.9 & 4.1.10

∞ **2D-3D transition:** Most of the interactions that will change here are from the positioning perspective in 3D. 2D-3D transitions are



necessary as they imply generating a 3D landscape, which can help both automate many 3D processes and lends flexibility to the user to choose an area of generation and control it.

∞ **Error correction:** I explored ways to correct errors for generations. Ideally, this should work in both segmentation and rendered tab. One of the explorations was an on-canvas option to highlight and choose error correction. Another iteration was a provision of an error corrector tool on the canvas tools. Lastly, I explored including an already present error correction tool on the settings pane.

Study 1d) System Quality feedback on the generated quality of output and looking at implicit and explicit feedback opportunities.

Scenario:

After spending some time customizing the chimera to her liking, Rose is ready to export her image. She gets prompted for feedback on the system quality, which allows her to review her options.

Solution ideation:

Rose gives the rating and looks at the implicit feedback that the GAN system records. She observes that the last image she generates is included as a part of the system feedback. She gets an option to remove or look at a later date.

Explorations:

Refer to [video 4.1.1](#) (check references 6.1) and figure 4.1.11 & 4.1.12

∞ **Giving feedback**

∞ **Accessing older feedbacks**

Evaluation Study 1

A possible settiwng to evaluate design is to ask a game designer or an artist to execute a task set in a particular scenario. The evaluation would be accurate if the scenario setting and goal assigned to the user are similar to ones used in the studies. While the user is following through a particular task flow, assessment is on the following criteria:

- If a user can easily create, control output, feedback outputs, correct errors, and feedback system quality without external aid and assistance using the interface controls provided.
- If a user follows an identical logical flow as laid out in the task flow for the scenario.
- If the user develops an understanding of each interface feature option and knows what its functionality is.

** Evaluation doesn’t rely on system functioning but interface controls as being described sufficiently helpful in aiding users to perform tasks. System functioning is assumed to be optimal and glitch-free.

Observation Study 1

Overall, I drew several conclusions about what interface features a GAN system should have to extend users’ use of these features and provide user agency. I made design decisions on a hierarchy and visual look of elementary interface designs. The interface features provide users with tools that extend the flexibility of moving between generations seamlessly, selecting different control options, and utilizing the element of play that the interface offers. A user landing on the interface will move through specific steps to reach a particular goal. A provision of options like relevant imagery, segmentation map, on-canvas options, interpolation, positioning, and lighting give users the means to influence results in the way they want. Many GAN interfaces limit generation options and controls. I propose that assimilation of these control tools can increase user agency in relation to the GAN system. Crucially, the provision of intention setting either by images, text, or speech helps both the user and the GAN system together move forward to accomplish a goal. Under this interface, a user works in

tandem with the GAN system to produce exploratory output in line with the user’s intent, rather than a system-generated automated output that is undesirable. The shown hi-fidelity wireframes do not cover all possible interactions and options and instead target a particular task flow that presents the persona moving through the interface. Design explorations on converting 2D to 3D and logical flow for getting feedback can appear restrained because it was outside my scenario’s scope, but these specific interaction features provide a rich area for an inquiry to be taken up later if time permits. Another impediment is the visual design of the interface features themselves, as I created only high-fidelity wireframes. It can be contended that there are many ways to design and position a particular interface element, like an error correction, but this exploration concerns the user experience and task flow rather than a specific designed visual look. Realistically, the GAN system I am imagining will be vast as there are copious GAN models, but I narrowed it to the two kinds of generations; landscapes and animal chimeras. Lastly, all my iterations and task flows set for a user operate under the notion that system functioning is glitch-free and morally unobjectionable, so interactions are subject to change in a real-world scenario.

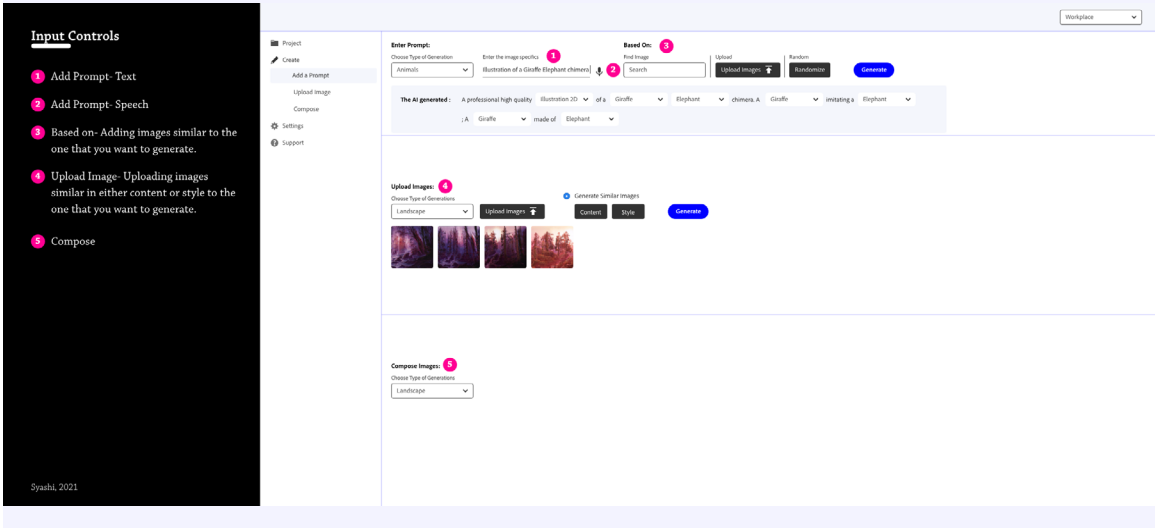
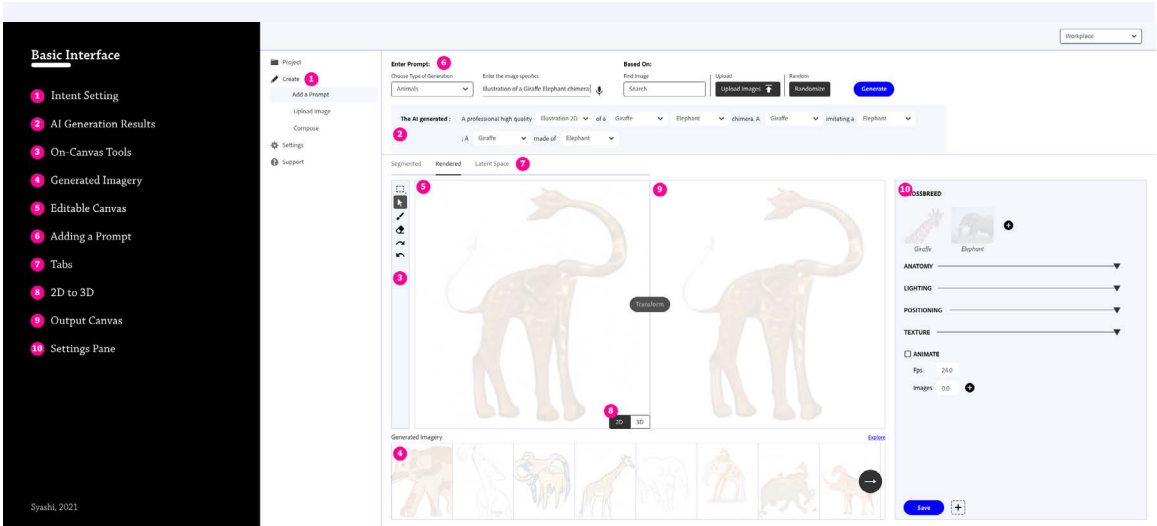


Figure 4.1.1- Basic interface look. The hierarchy and positioning of the element were decided after going through multiple design iterations.

Figure 4.1.2- Input controls available to the user. They can either add prompts, upload images, or compose images post selecting the generation type.

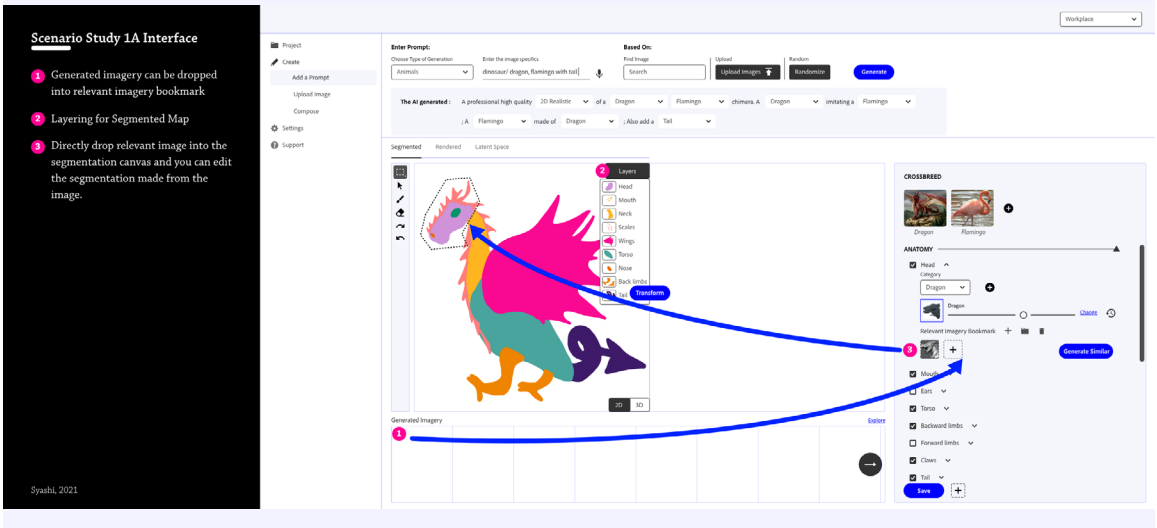
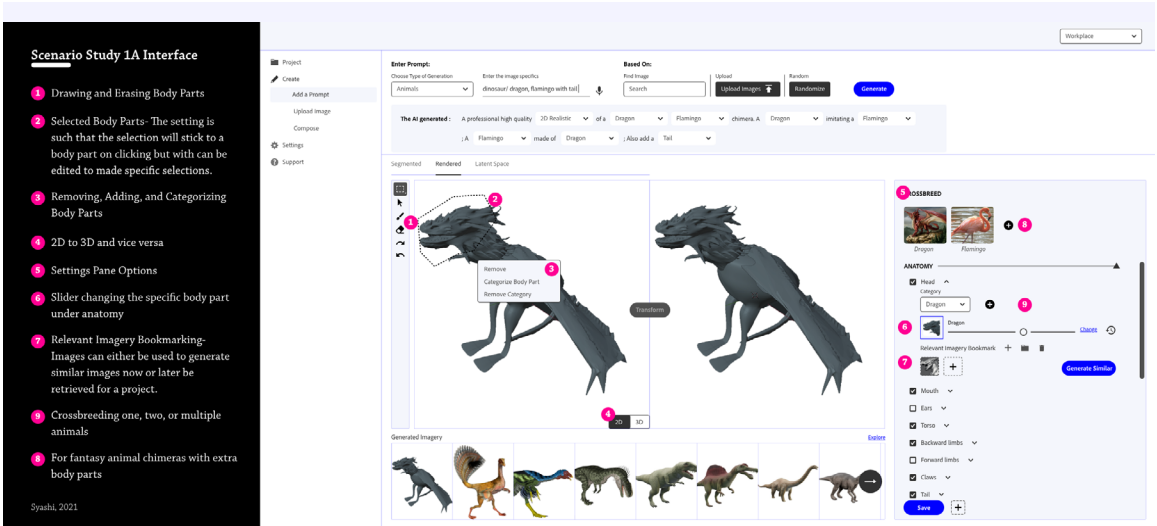


Figure 4.1.3- Study 1A interface. A user landing on the interface has access to this set of controls that can support them in completing a particular task flow described in scenario 1A.

Figure 4.1.4- Study 1A- interaction affordances and tools available for the user.

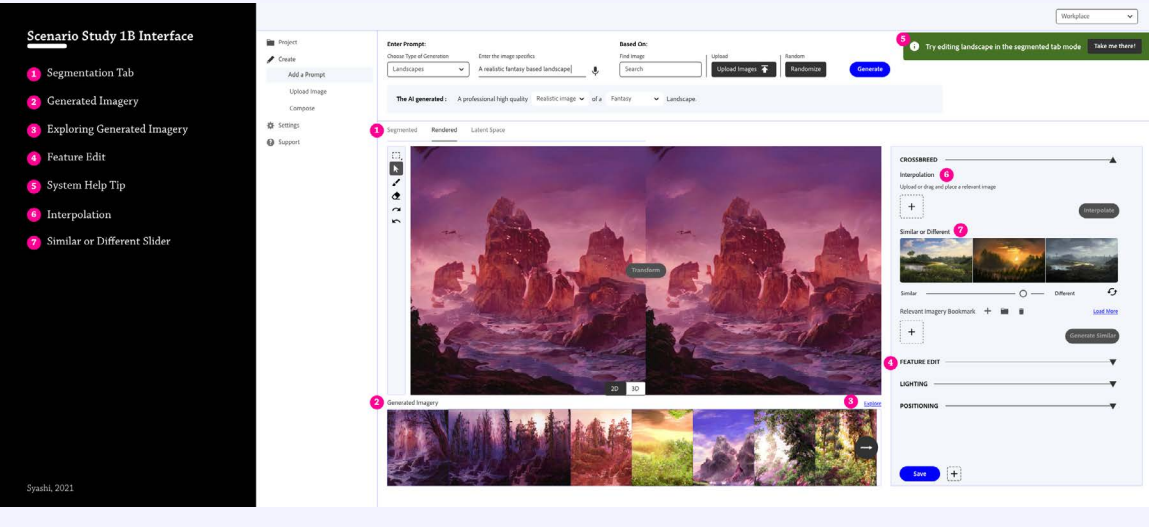
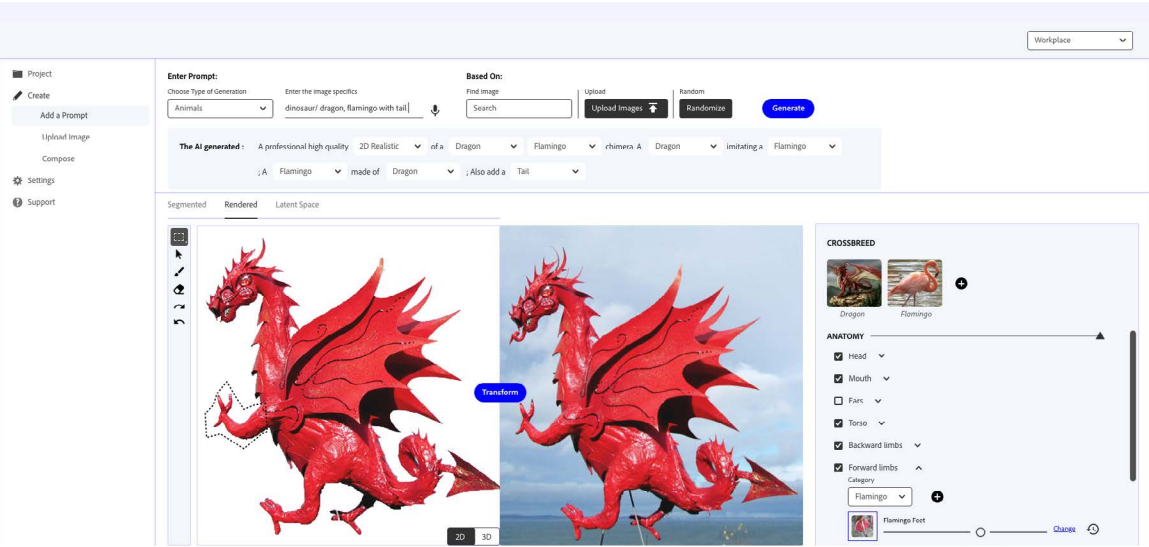


Figure 4.1.5- The UI provided is adaptive and so when Rose makes a selection on the canvas for the front limbs, the front limbs section opens on the settings pane to make changes.

Figure 4.1.6- Study 1B interface features. Users give relevance feedback, and the system generates images in real-time based on the given feedback.

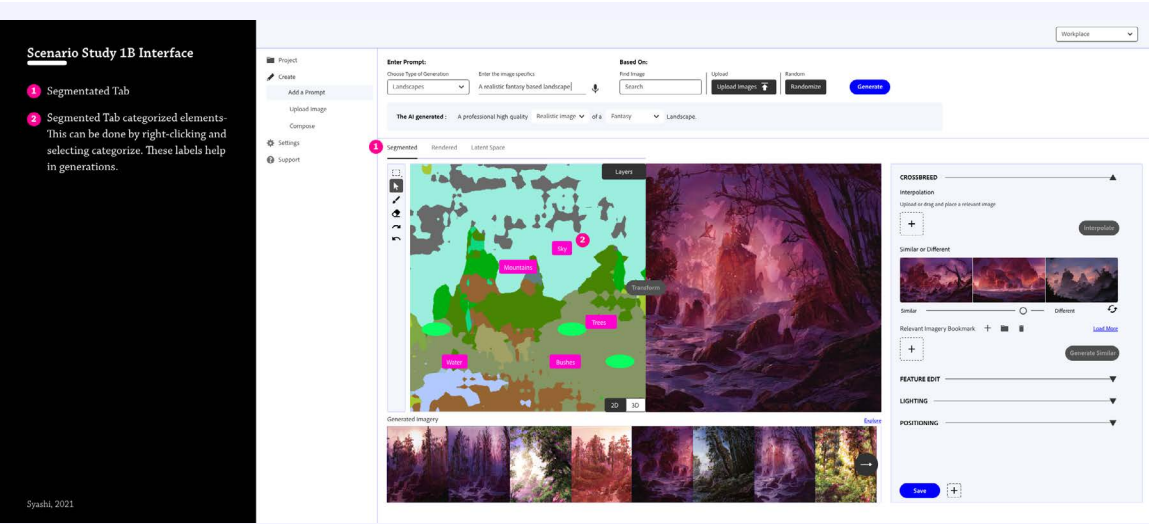


Figure 4.1.7- Users on a segmented tab can use layers to add new elements and then categorize them. Categorization and tags help the system understand and generate as per users' imagination.

Figure 4.1.8 - A user is offered options like interpolation, similar or different, and relevant imagery bookmarks under crossbreed. For feature edit, a user gets to edit color and natural elements in a landscape.

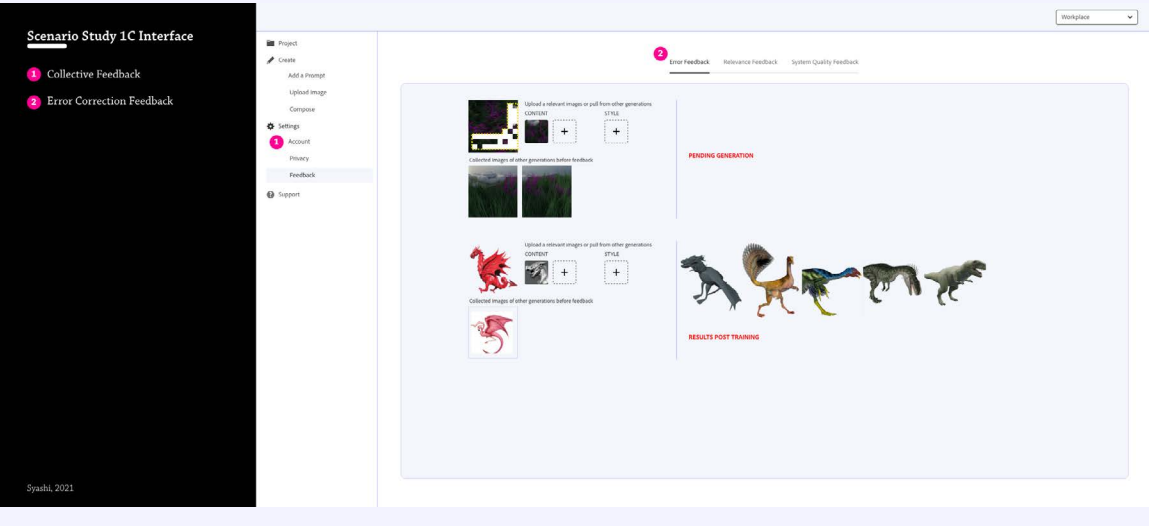
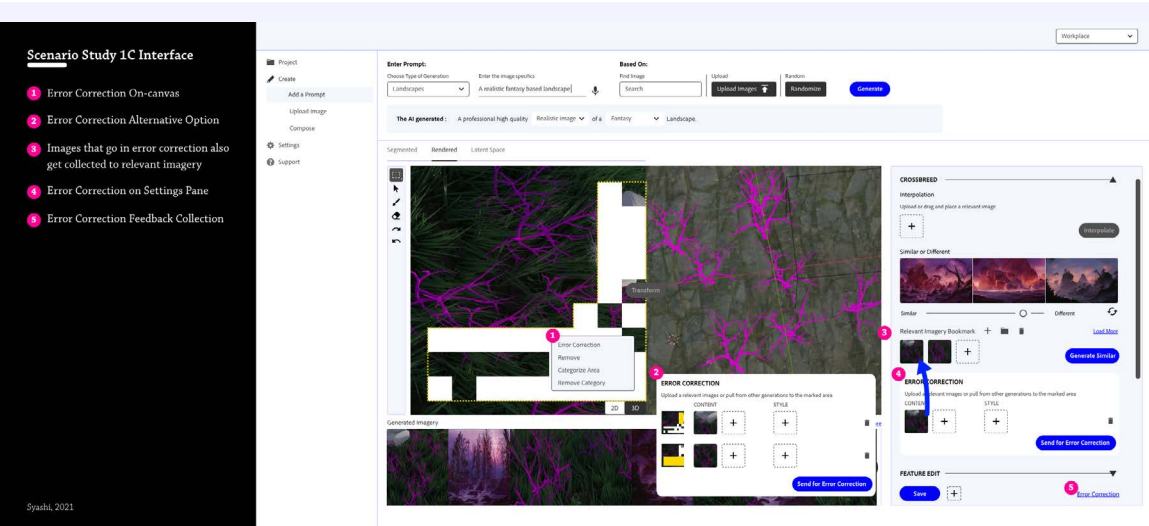


Figure 4.1.9- Study 1C interface. A user can correct errors by marking them and sending them as feedback. Users can later view their critique and its impact on the generations.

Figure 4.1.10- To create transparency both explicit and implicit feedback is shown to the user.

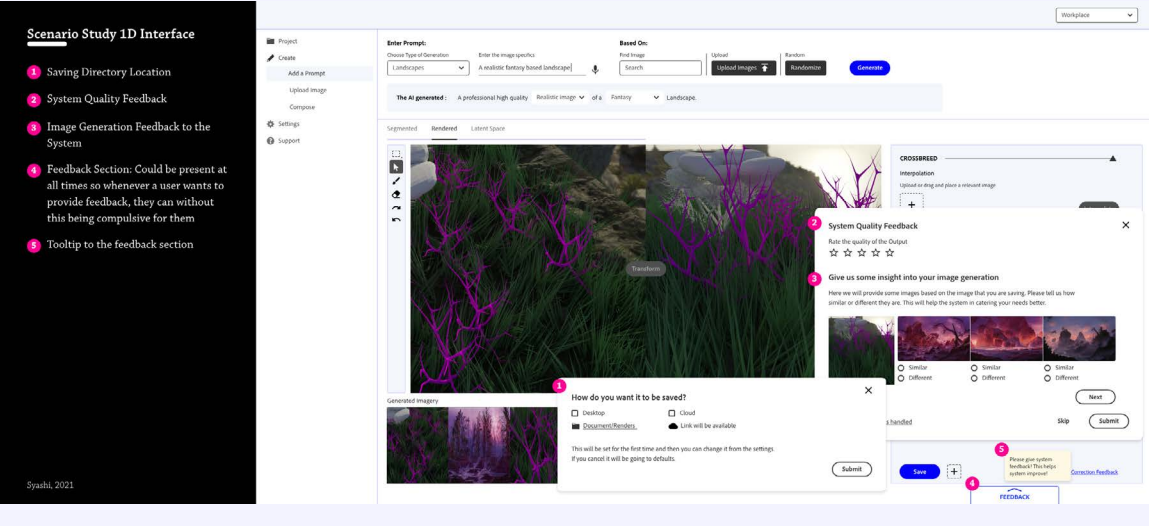
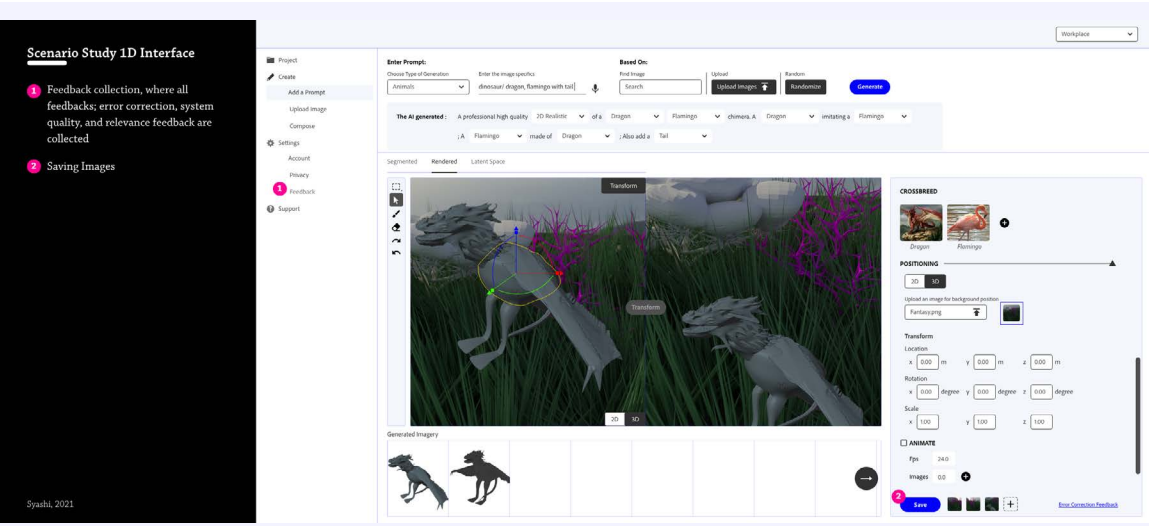


Figure 4.1.11- When a user clicks on the save button, they are requested to give feedback on the system quality.

Figure 4.1.12- Directory to save a file in can be found in system settings but while initially setting up user gets to pick saving location from the pop-up for feedback. More explorations for logic and interface designs for gathering system quality feedback will be required as it is a crucial part of the interface features.

4.2. VISUALIZATION OF NEURAL NETWORKS

Question

How can network visualizations reveal internal decisions of the system to help an AI novice interpret the GAN system and its decisions?

Study 2

In Study 2, I am exploring the visualization of GAN system inscrutable internals, targeting the black box of neural networks. This exploration can lead an AI novice user to develop interpretability of the GAN system and its outputs;—specifically for erroneous outputs. For this study, I wanted to test if the user can understand how the system reaches a particular output and if the user can practice agency to collaboratively explore with the GAN system. This study concludes by exploring systems’ internal visualization methods of feature visualization and unit visualization that can aid in a user developing interpretability of the system.

Study 2 is a step further from Study 1 as it helps interpret the system’s output. A curious user, in hopes of understanding, fixing their erroneous outputs, or playing with the GAN system, can try different visualizations that help make the GAN system interpretable. As mentioned in Carter et al., (2019) instead of painting with pixels and colors, a user can directly choose the content to paint on the canvas. Imagine a collage made by a user depicting things present in a scene, and the GAN system generates the output scene in the style of the user’s choosing. While some of these visualization techniques are employed by experts for understanding image classification systems, the same study by Carter et al., (2019) explains the possibility of generating images using these. Other research papers(Bau et al., 2020, Samek et al., 2017, Bau et al., 2018, Olah et al., 2020, Olah et al., 2019) show different visualization techniques for network visualization: heatmaps, saliency maps, unit visualization, activation atlas, and feature visualization. Further, Olah et al. (2018) talk about how interface designs utilizing these visualization techniques to explain systems working can extend the system’s explainability. In this study, I am exploring concepts around testing and designing for explainability to an AI novice to better suit their mental models.

I started exploring this study in 4 parts, and for each part, I created a scenario, which is a continuation from Study 1, where the character

***Saliency Heatmaps-** Saliency maps help us understand what a CNN is looking at during classification. Saliency maps are a part of feature visualization techniques.*

***Activation Atlas-** Showing the feature visualizations of the basis neurons gives us the global view of a network that we are seeking. In practice, however, neurons are rarely used by the network in isolation, and it may be difficult to understand them that way. As an analogy, while the 26 letters in the alphabet provide a basis for English, seeing how letters are commonly combined to make words gives far more insight into the concepts that can be expressed than the letters alone. Similarly, activation atlases give a bigger picture view by showing common combinations of neurons.*

***Unit Visualization-** Unit Visualization is a method to edit and correct errors in the GAN system presented in both Bau et al. (2020) and Bau et al. (2018). This method provides ways in which the GAN models can be dissected and errors from the models corrected while making the system interpretable*

Rose moves through the interface to complete particular tasks. In each of these parts, I am exploring a visualization technique. In these visualizations, exploration of latent space comes up multiple times, and as a part of the visualization, I will be exploring latent space as well. The high fidelity wireframes can be accessed through [link in references section 6.1](#).

The four parts of Study 2 are:

Study 2a) Unit Visualization

Study 2b) Saliency Heatmaps + Channel Activation

Study 2c) Optimization

Study 2d) Activation Atlas + Units Visualization + Optimization

Ideation 2a, 2b, 2c, 2d

Study 2a) Unit Visualization

Scenario:

Rose chose to generate a landscape image with forest and mountainous terrain. The generation looks neat and unlike anything, Rose has seen before. Rose wants to remove some elements and add in others. She wants to see how relevant imagery or her uploaded images influence the generation of the image. Rose wants to explore adding fantastical trees and rock elements to her landscape. She already interpolated and used the relevant imagery to generate similar images. The generated landscape image looks very different from anything she could imagine. Rose would prefer an extra capability to influence the results directly and explore more with the GAN system via element selection. She knows how to create and categorize a segmentation map, but that isn’t accurate, and she can’t choose specific trees or create ones that she wants to use. These generated images can be easily tested with the users as prototypes or given to the design team, who can visually understand and create landscapes based on these images.

***Feature Visualization-** Feature visualization can make the hidden layers of networks comprehensible.*

***Optimization-** Optimization can give us an example input that causes the desired behavior. It turns out that the optimization approach can be a powerful way to understand what a model is really looking for because it separates the things causing behavior from things that merely correlate with the causes. Optimization also has the advantage of flexibility. For example, to study how neurons jointly represent information, we can easily ask how a particular example would need to be different for an additional neuron to activate. This flexibility can also be helpful in visualizing how features evolve as the network trains. If we were limited to understanding the model on the fixed examples in our dataset, topics like these ones would be much harder to explore.*

Solution Ideation:

Rose lands on the interface and goes to the landscape generator, where she already has a base image that she had previously created with the GAN system. She goes directly to the segmentation tab and tries to add trees. Each rendered generation is either fun or weird and it motivates Rose to experiment more with GANs. These generations are something Rose hadn’t envisioned. Then she goes to the latent space tab and searches trees. Rose finds a map of trees in the latent space visualization. She hovers on the latent space visualization and she sees different kinds of trees that exist in the system in the table on the right. She can pick some trees and put them in her content bucket. As she goes back to the rendered area, she finds that she can paint with the contents directly on the canvas. She can see the different layers and their contents. She can move layers up or down, edit or create new layers, and view, hide, or delete layers. While coloring with elements, she can make minor adjustments to details like the number of trees and their size. Finally, pleased with the imagery she has successfully created with the trees, Rose thinks exploring rocks to include in the current landscape could make it a nice game scene prototype. On the rendered tab, she selects an area and selects categories as rocks. The UI automatically adapts to give her unit visualizations of different rocks, and simultaneously the generated imagery changes in different ways according to the system’s understanding. Rose has fun directly editing on the canvas with content and playing with GAN generations.

Exploration:

Refer to [video 4.2.1](#) (check references 6.1) and figure 4.2.1, 4.2.2, 4.2.3, & 4.2.4

∞ Latent space exploration

∞ Painting on canvas with content

- Elements bucket
- Layers Panel
- Setting Pane

Study 2b) Saliency Heatmaps + Activation Atlas

Both Saliency Heatmaps and Activation Atlas reveal the neural network internals.

Scenario:

Using only the segmentation maps as a knowledge base, Rose can’t understand the details. Although she can control and constrain image qualities, to know how outputs are generated she must understand how the system is functioning. She saw blurred and incomplete textures generated in the landscape generator. She can send for error correction, but she wants to understand why for a particular landscape, the surrounding area has water and not land. She notices that the generations have areas where details like these are off, and she is interested in knowing why and how the system generates these different weird images.

Solution Ideation:

Rose is working on landscape generation. She is happy with the generation, but some areas catch her attention. She isn’t able to understand why the system is generating water when she has not added or suggested any element of water. As she marks the area, she gets an expandable area that reveals an activation atlas for the image on the canvas. On the activation atlas, she can see the activated images for the ground have water on the sides. They are associated with rocky beaches with water. She understands high activations and lower activations areas shown to her. To change her generation she follows steps to switch higher activation areas to ones with dry sand grounds without any water body around. She goes through a similar process for the chimera generator, where she can see what the system has been imagining. Rose tries the compare and contrast features to understand the system generation and switch generations with ease to get the explorations from the system constraint in some ways while being open-ended in others.

Exploration:

Refer to [video 4.2.1](#) (check references 6.1) and figure 4.2.5, 4.2.6, 4.2.7, 4.2.8, 4.2.9, & 4.2.10

- ∞ Activation Atlas
- ∞ High Activation Areas
- ∞ Low Activation Areas
- ∞ Changing Activations
- ∞ Comparing and Contrasting elements

Study 2c) Optimization

Scenario:

For chimera generations, Rose creates a segmentation map but does not categorize limbs. The system generates very different kinds of limbs with talons and a different texture. After categorizing, she tries to generate again. These generations are both unconventional and fun. She wants to understand how the system is getting its output and if she can activate and explore more with the GAN system.

Solution Ideation:

After looking at the activation atlas and optimization options in the settings pane, Rose is a little confused about what these tab options are. She tries the last tab called tab perspective. Here the layers and tabs are explained. These explanations help Rose understand the different options she has. Finally, Rose goes back to her problem with the chimera generation. After creating a segmentation map, Rose didn’t categorize limbs, and she wants to see even more possibilities for the limbs. She marks the limb area and then finds that the system shows her closest activations and corresponding database images for the marked area. Rose can then choose specific activations that she wants to emphasize and see what the GAN system generates based on that. After looking at the database images, she emphasized Flamingo legs for her chimera. Rose now understands how to use these activations to interpolate to find newer possibilities of generations. Interpolation can be done in steps, and Rose can pick and play with the ones closest to her liking. This lets Rose work together with the system to explore possibilities that inspire the unconventional.

Exploration:

Refer to [video 4.2.1](#) (check references 6.1) and figure 4.2.11 & 4.2.12

- ∞ Tab Perspective
- ∞ Optimization
- ∞ Interpolation in Optimization

Study 2d) Units visualization + Activation Atlas

Scenario:

Rose gets different generations as she keeps refreshing on the rendered tab for the same segmentation map. She particularly finds the kind of generations the GAN system makes interesting as she had not seen them before. After a particular generation by the system that looks fun to explore, Rose wants to understand the reason for the generation and assess and explore elements with the GAN system.

Solution Ideation:

Rose goes through an exploration where the system layers are revealed to her. She goes through the activation atlas, latent space, and unit visualization to get to the generation she wants. This is an exploration of both visual form and interactions that Rose can go through to utilize both of the latest methods of visualization. Post-exploration, Rose can play with more interpolation and animation.

Exploration:

Refer to [video 4.2.1](#) (check references 6.1) and figure 4.2.13

- ∞ Interpolation in Latent Space
- ∞ Moving and working between different tabs and settings pane.

Evaluation Study 2

A possible setting to evaluate GAN systems’ internal visualization for developing interpretability is to ask a game designer or an artist to execute a task set in a particular scenario. The evaluation would be accurate if the scenario setting and goal assigned to the user are similar to ones used in the studies. The main observation to make is if a user can interpret and understand the system’s outputs. Other important aspects to evaluate would be:

- If a user can understand these internal visualizations and their importance
- If a user can take agency to investigate an unexplored territory by a human creator before by looking into the Latent Space.
- If a user while exploring these visualizations use agency offered by the system to generate more unexpected outputs.
- If a user can recognize the reason for the problem arising by looking at activation atlas or optimization and take necessary steps to align generations with their intent.
- If a user can switch between different internal visualization options to fathom GAN system’s outputs, especially erroneous ones.

Observation Study 2

Many observation points came across in Study 2. Study 2 is meaningful for both constrained and unconstrained GAN system outputs that can serve as a source of inspiration for the user wanting to explore. These generations can push a user into exploratory directions, and provisions of latent Space explorations and interpolating between elements can expand computational creativity into developing a system’s interpretability for an AI novice user. When a user is generating chimera or landscapes, a GAN system’s capability to envision and create something beyond human imagination is fun and gives rise to new creative directions, but there are bound to be errors, and a user will want to understand the output and how they can fix it. Interpo-

lating and experiencing internals’ visualization with some aspects of the agency as developed in study 1 can help users interpret the system working and tweak results. Although Study 2 provides visual explainability to the GAN system, it lacks textual explainability. A user can understand systems output and its origins, but there is a need for a walkthrough and textual explanations. A need for textual descriptions sets the stage for the third study, where I will work on exploring means and methods to show textual explanations. Another element to observe is the many ways to design a particular visualization, and it would be advantageous to test multiple explorations of showing visualizations on the interface and the interface options that suit users’ mental models. Finally, although I have designed visualization of latent space, unit visualization, and activation atlas, they will vastly differ from the visualizations shown in the figures(4.2.1 - 4.2.13) because of the difference in functions and models used, and as a result their visual interpretability to the user may also be affected. There are multiple internal layers in a neural network, and it will be necessary to research which layers of the visualization will make the most sense to a user. Engineers and designers need to work together with users to find which layers make the most sense. For Study 2, I have assumed that the best internal visualizations that are easily interpretable will appear to the user.

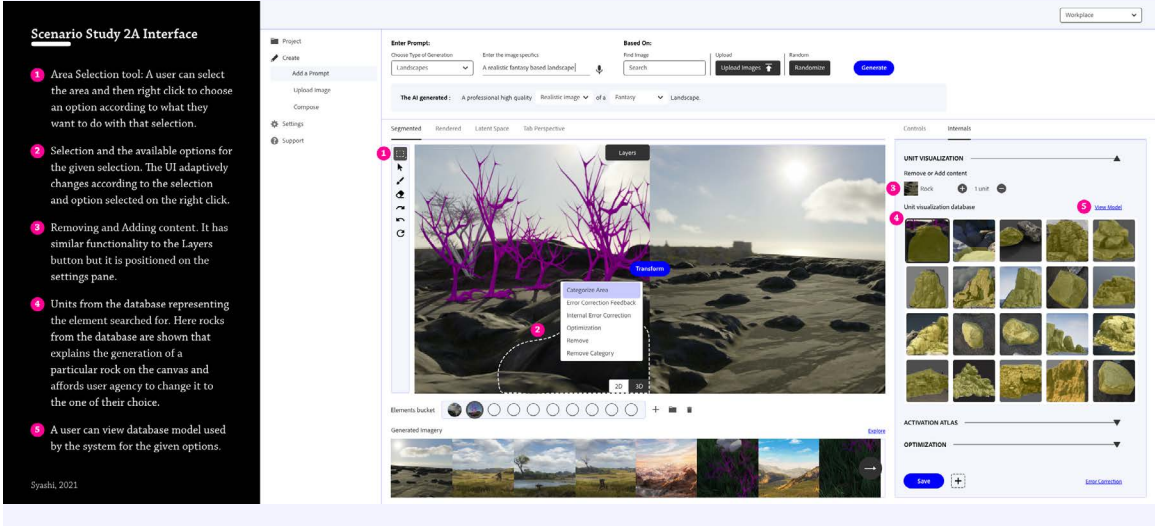
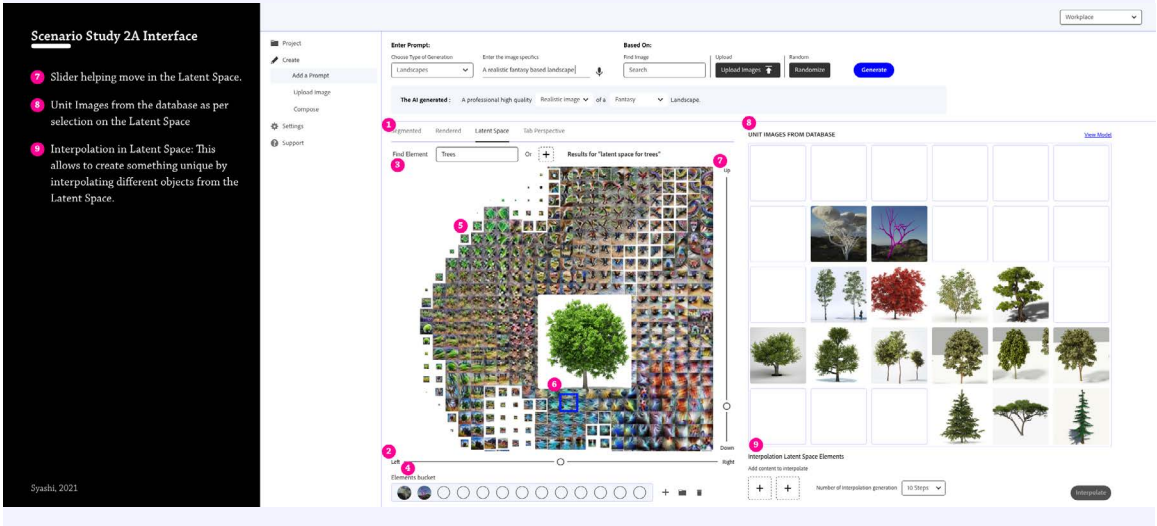
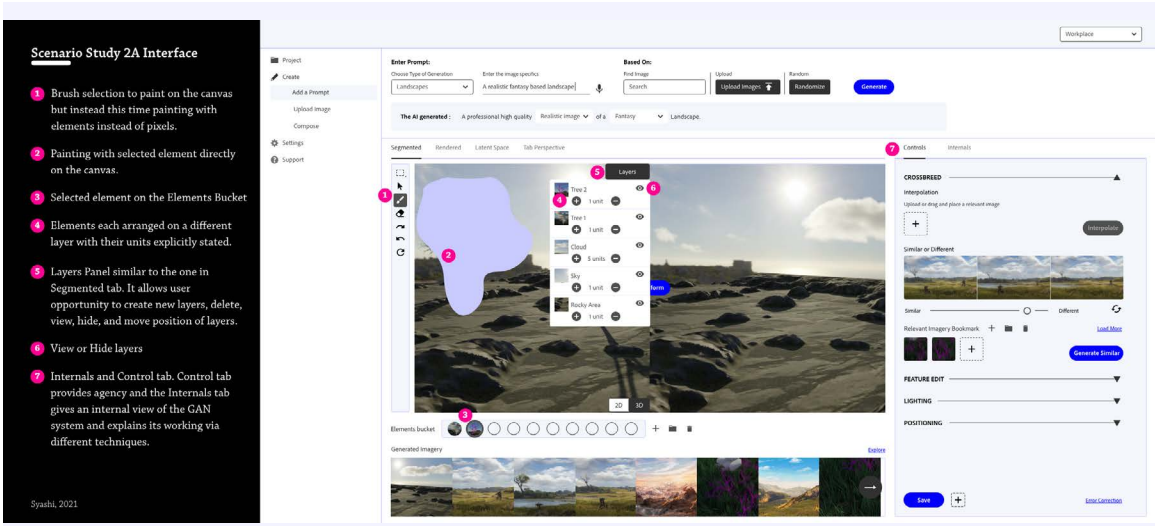
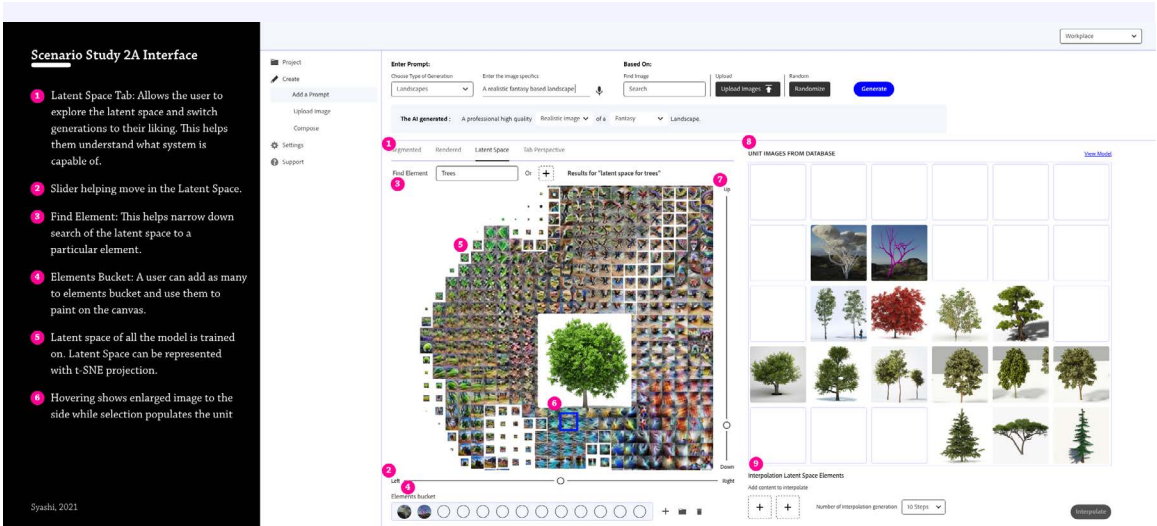


Figure 4.2.1- Latent Space Tab exploration. Latent Space visualization may differ as per the dataset and method used to visualize. The images shown do not exactly represent the Latent space.

Figure 4.2.2- Continuation of Figure 4.2.1, Latent Space Tab exploration.

Figure 4.2.3- Given affordability to the user to paint content directly on the canvas and available layer options.

Figure 4.2.4- Unit visualization and available options to the user.

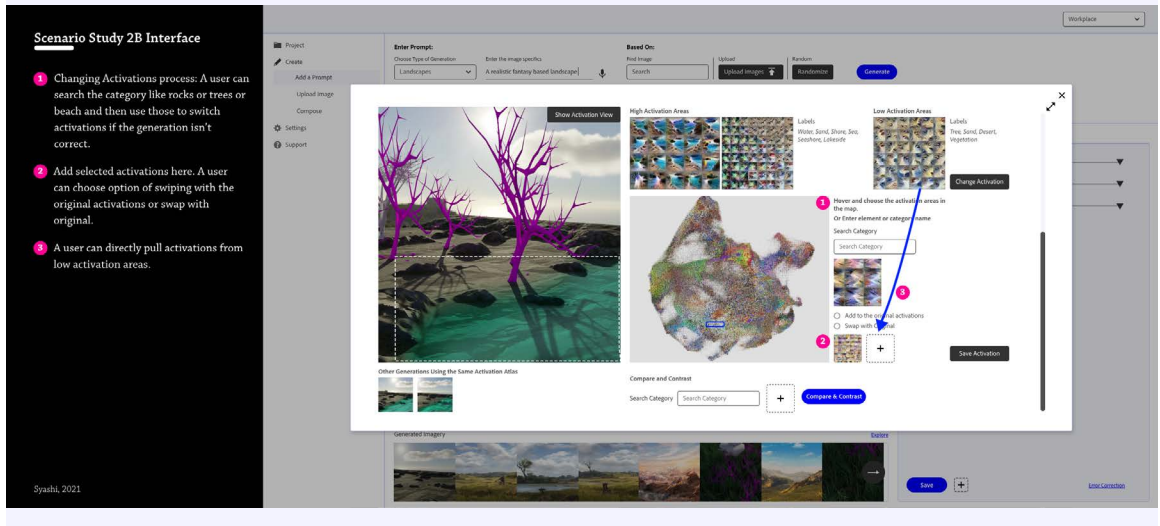
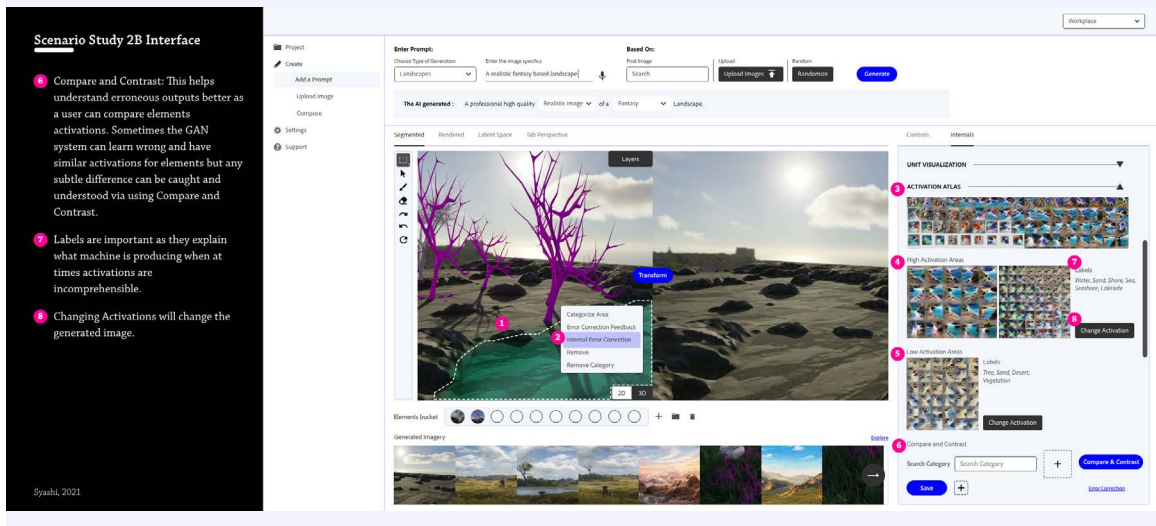
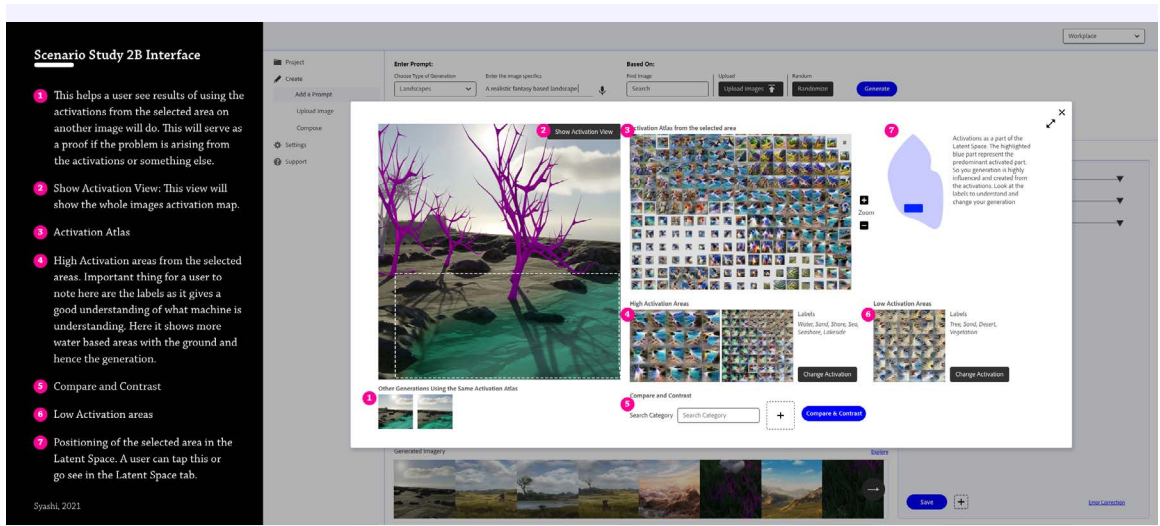
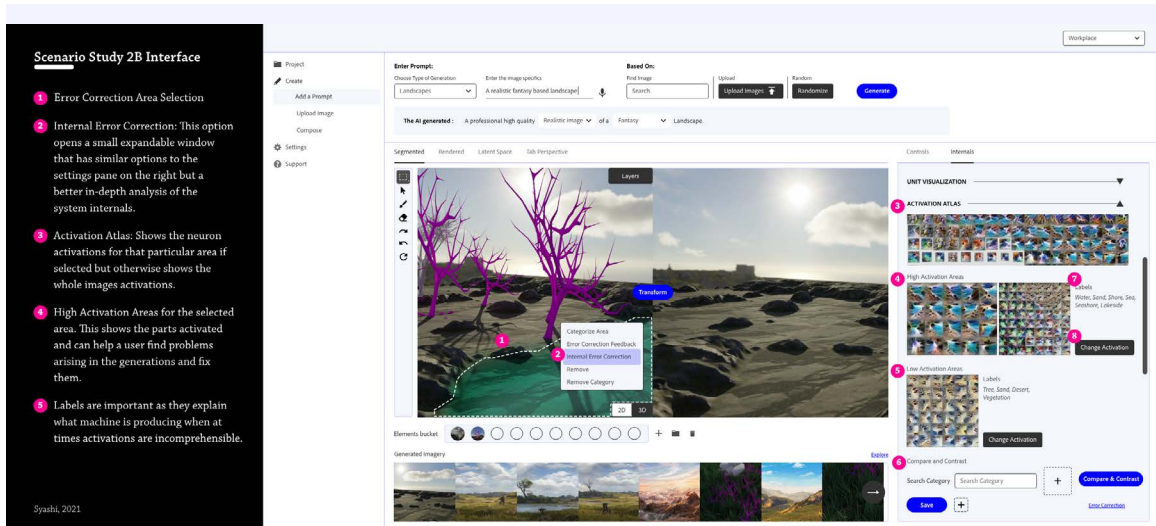


Figure 4.2.5- Activation Atlas visualization on the settings pane and available options to the user.

Figure 4.2.6- Continuation of Figure 4.2.5, Activation Atlas visualization on the settings pane and available options to the user.

Figure 4.2.7- Activation Atlas exploration pane

Figure 4.2.8- A user can change activations and, as a result, the image will change. The outputs may or may not look as per user wants or meet user expectations, but the generations help with explorations.

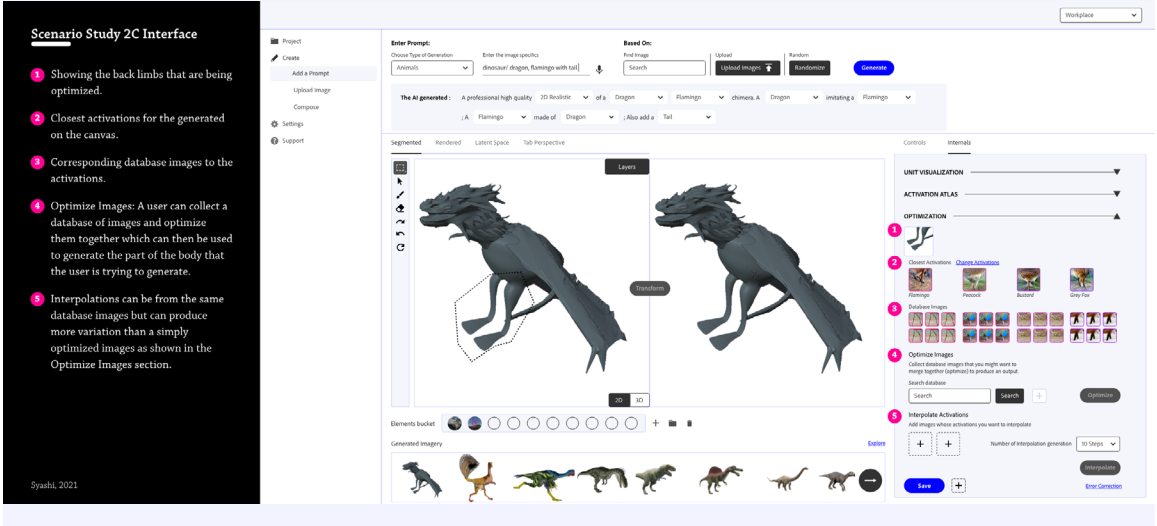
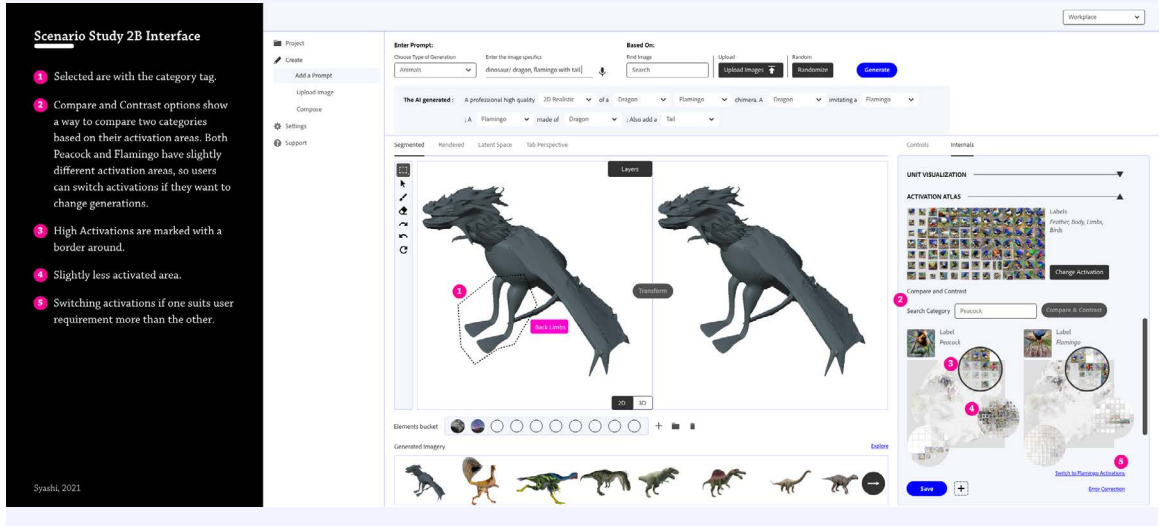
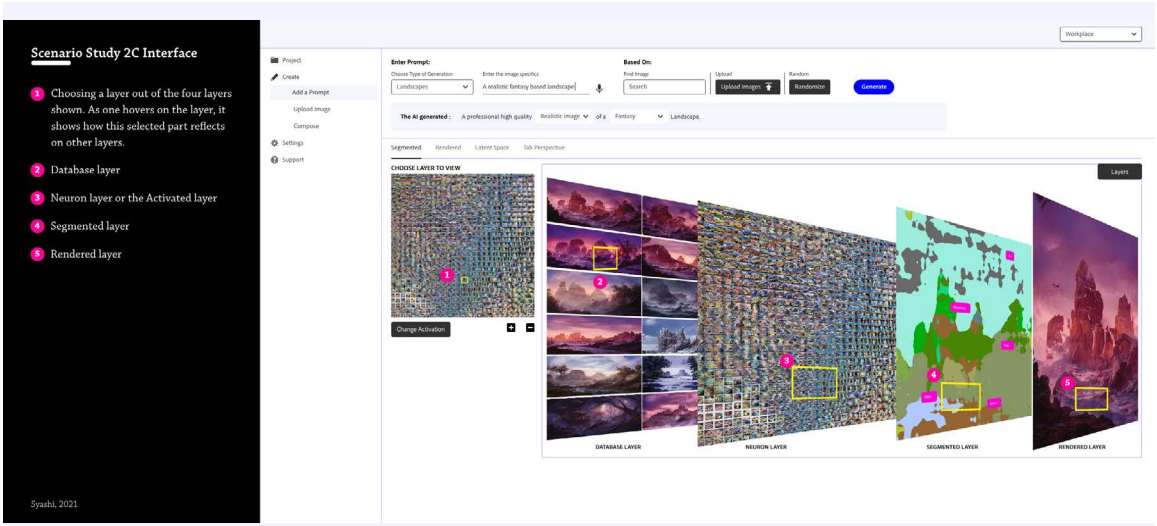
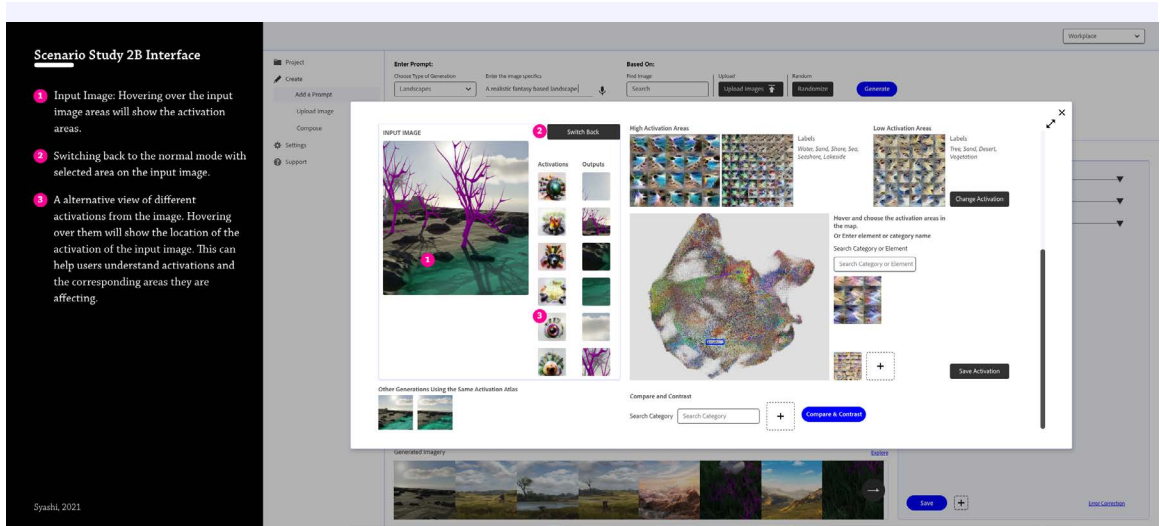


Figure 4.2.9- A alternative view of activations from the image. Hovering over them will show the location of the activation of the input image.

Figure 4.2.10- Compare and contrast options show a way to compare two categories based on their activation areas. Both Peacock and Flamingo have slightly different activation areas, so users can switch activations if they want to change generations.

Figure 4.2.11- Tab Perspective tab explains different tabs and their relations to one another to the user.

Figure 4.2.12- Optimization Visualization- A user can look at activations and corresponding database images, that are combined to create that optimized activation.

4.3. TEXTUAL EXPLANATIONS

Question

How can textual explanations communicate varying levels of information to facilitate AI novice’s trust in the GAN system?

Study 3

Within study 1 and study 2 there are features and designs that would require labels, pop-ups, descriptions, information icons, walkthrough explanations, etc. Explanation strategies that can calibrate the trust of a user to an optimum would be necessary for scenarios that require understanding the internals of the GAN system. This study supplements study 1 and 2 and enrich their content. Once a user develops an understanding of a system’s capabilities and limits, they can understand how and when to trust the system to accomplish their goals. Both the PAIR worksheet and Hoffman et al.(2019) talk about ways trust can be engendered through explainability. Study 1 provided interface features for users to develop agency, Study 2 built on the interface design and visual forms to explain the system internals, Study 3 intends to build on textual explanations and ways they can be presented to the user. Study 3 is done to instill users’ trust in the system and it is important because AI has bias and as based on models coded by humans who are inherently biased, the AI is bound to have errors. If a user like an AI novice can understand the origin of the errors and fallibility of the machine, they can then help fix the errors and improve the system. A collaboration between AI novice-and the system reduces the automation bias that, occurs when human operators ignore other senses of information including their faculties, as they overly trust the automated system (Zerilli et al., 2019), and algorithmic omniscience, which means users over-accepting system outputs (Hollis et al., 2018).

For Study 3 I started off by listing the system features that could confuse users because of technical details or user’s distrust the system, then finding explanation strategies that could mitigate both these situations, and finally drafting different explanations to be tested with the users. The forms of explanations can vary from text to audio to visuals. Some of the explanation forms that I have chosen to explore are as follows

∞ Pop ups

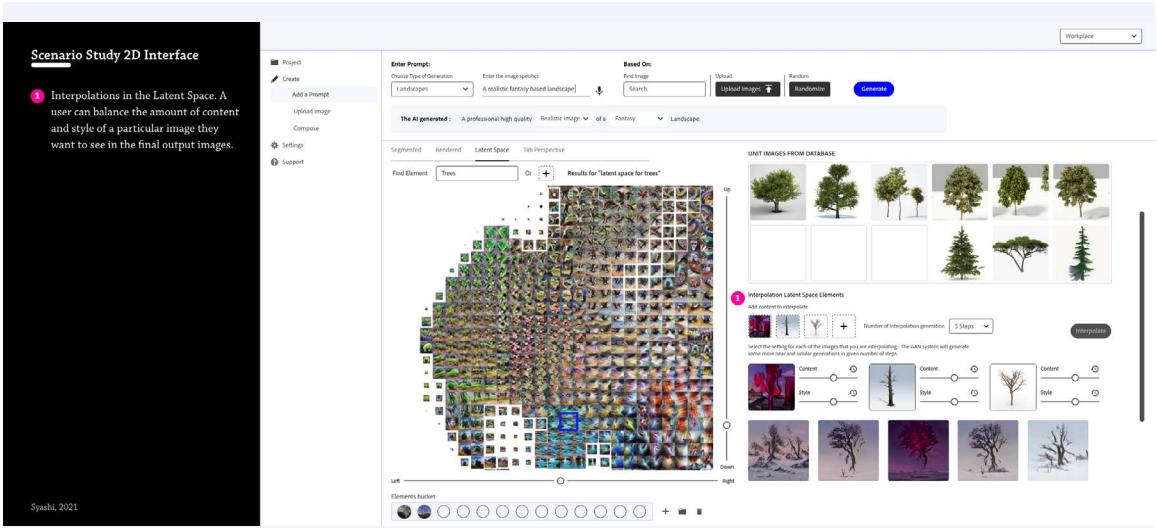


Figure 4.2.13- Interpolations in the Latent Space. A user can balance the amount of content and style of a particular image they want to see in the final output images.

- ∞ Information buttons
- ∞ Labels
- ∞ Walkthrough explanations
- ∞ N-best alternatives
- ∞ What ifs
- ∞ Data visualizations.

For evaluation measures of this Study 3 users can be tested on which explanations helped build mental models and trust (Cahour-Forzy scales) in the system (Hoffman et al., 2019). High fidelity wireframes can be accessed through [link in references section 6.1](#) and scenario video through [link in references section 6.1](#).

Ideation Study 3

Study 3A:- Building trust through explanations for interface features that help both explain and explore the system.

Textual explanations can be provided for models that are used by the GAN system. This text can explain the reason why output is such via regression and show activation atlas for an image with labels that can tell us what models had the most impact on the generation. If explained properly this improves users’ trust.

Scenario:

As a user comes to the interface there are multiple ways of providing the explanations and a system will need to adaptively learn from users’ learning and exploring patterns, as to what suits their needs.

Solution Ideation:

A user is provided with multiple kinds of explanations(figure 4.3.1, 4.3.2, 4.3.3, & 4.3.4). Especially the system’s internal option explorations, a user gets to walk-through an example to understand different

options and meanings. These walkthrough tutorials pop up in the form of tips either towards the side or the top of the interface. A user gets a choice to dismiss them or work along with them. Alongside tips, a user gets a time assessment of a particular walkthrough. A user is given the option of finding information on a particular element by hovering or selecting. The system also provides helpful tips and pointers to unexplored areas. A user gets to see how the system gets to a particular result and tips on how they can explore better with the system. Alongside working with the tips, a user can check out the system on their own using the interface options and choosing to see the walkthroughs, animations, or tooltips as and when required. The system keeps providing information in the most understandable and transparent form and keeps learning from user preferences.

Evaluation Study 3

Study 3 can be evaluated by users after the user has had considerable experience with the system. Using Cahour-Forzy (2009) Scale, Adams, et al. (2003) Scale, and PAIR worksheet I have made these following questions to be tested with the user. Instead of using bipolar answers I plan to use a seven point scale. In seven point scale answers can be noted with on one end marked affirmatively while the other end marked with negation of that affirmation.

- Is the automation tool useful?
- How reliable is it?
- How accurately does it work?
- Can you understand how it works?
- Do you like using it?
- How easy is it to use?
- On this scale, show me how trusting you are of this recommendation.

What questions do you have about how the system came to this recommendation?

What, if anything, would increase your trust in this recommendation?

How satisfied or dissatisfied are you with the explanation written here?

Observation Study 3

Developing textual explanations especially interaction based explanations will require user testing in multiple rounds because there are places where explanations might be absolutely unnecessary, or distracting and inconspicuous explanations might not be needed. Specific outputs will necessitate specific explanations of how AI reached that output. Explanations and naming can be very difficult in case of system actions explanations and so there I tried to provide partial explanations and interactions that can help users understand. Having AI generated prompts for walkthroughs at different points can be really helpful for the user. A user can use these walkthroughs when unable to understand the system and its features, or wants to explore but doesn't know how. Knowing that this system that I am designing for is complex and scientific, it can be hard for an AI novice to approach it, but having textual explanations at all points makes it more explainable and in turn trustable. Additionally, the system's reactions and textual generations are dependent on users' action, hence building collaboration and mitigating automation bias. Study 3 helps a curious user understand system outputs in depth, which further helps mitigate problems of algorithmic omniscience.

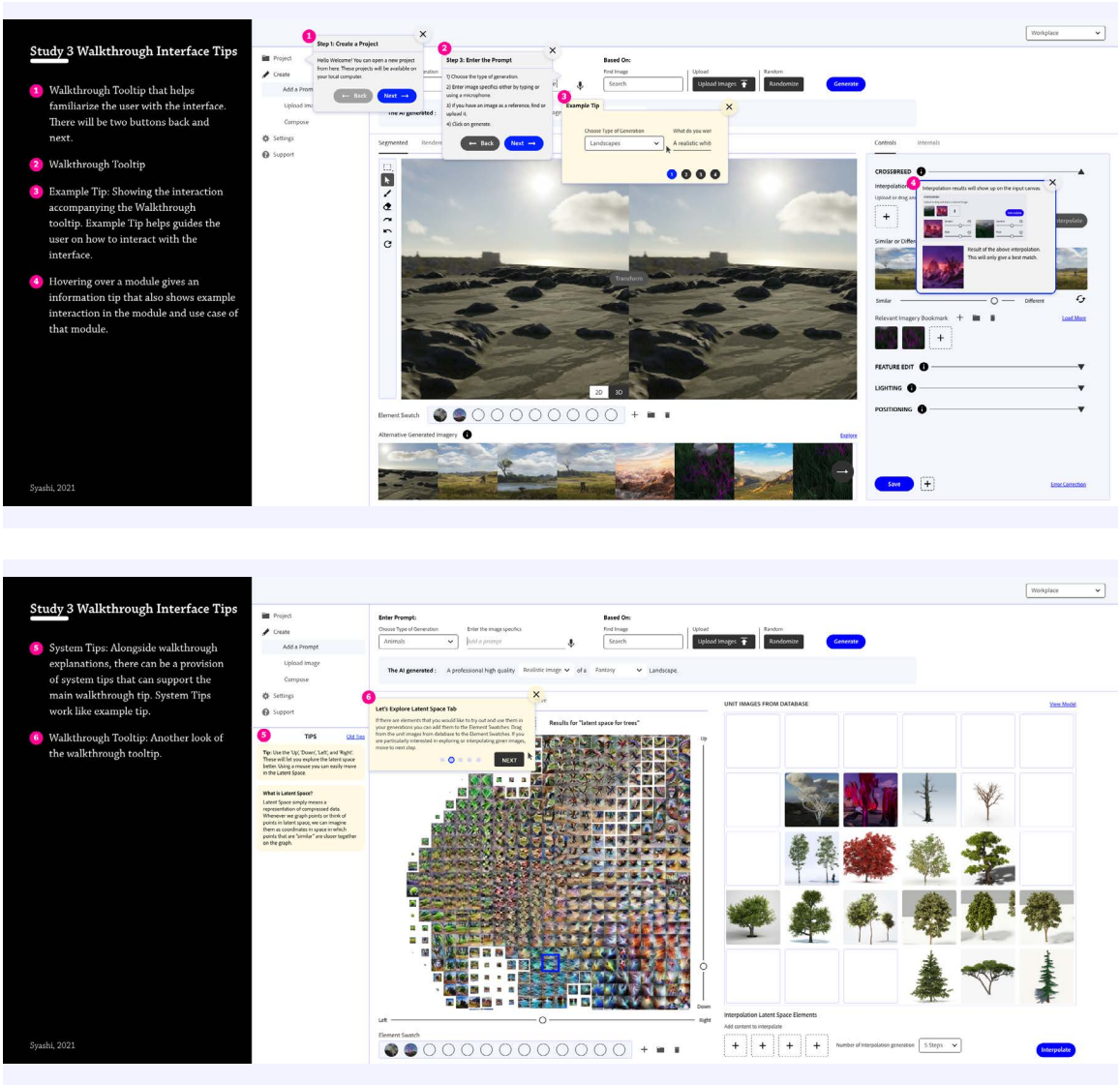


Figure 4.3.1- Walkthrough Interface Tips. An exploration of Interface tip look and information content.

Figure 4.3.2- Walkthrough Interface Tips continued. An exploration of Interface tip look and information content.

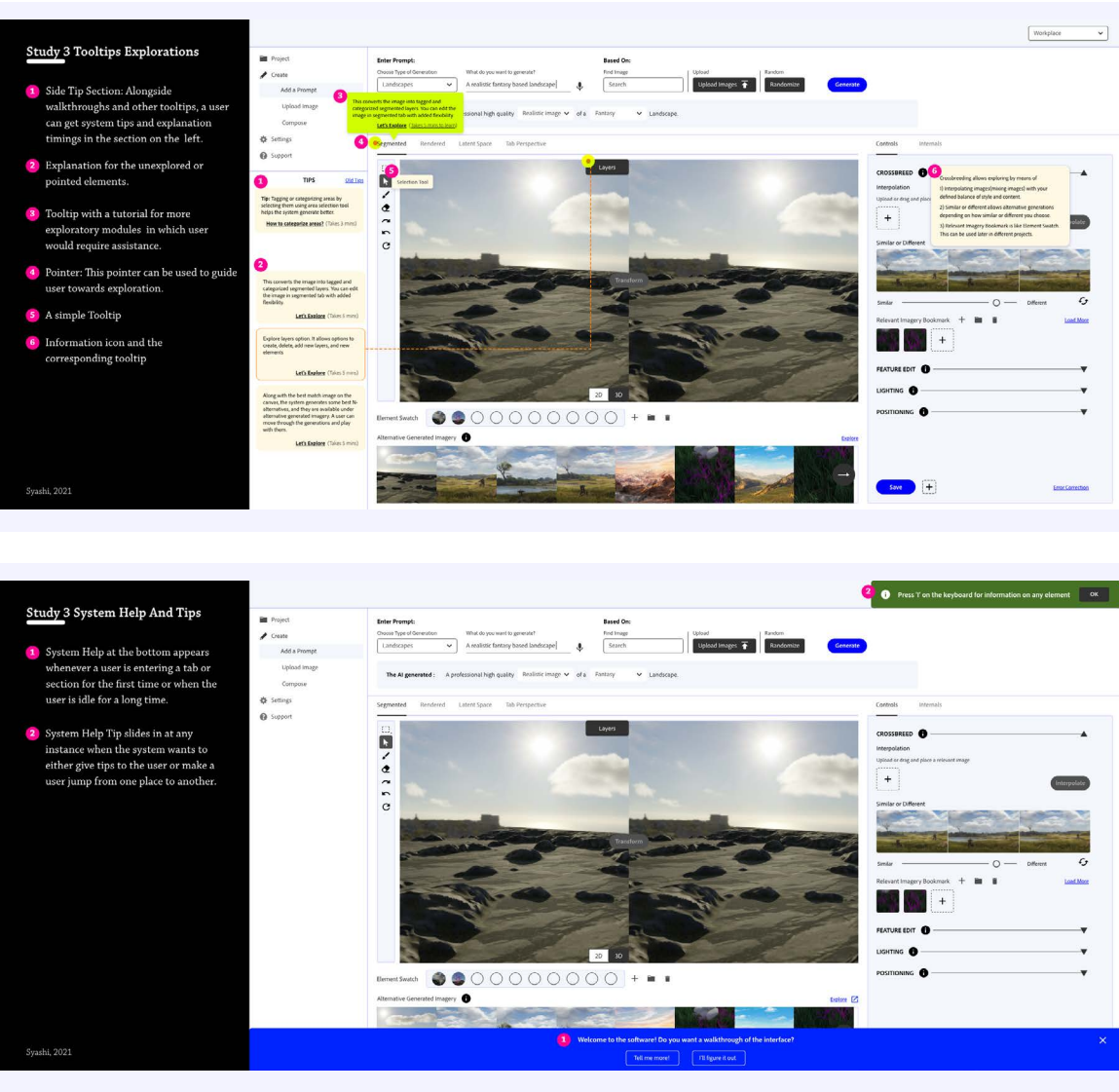


Figure 4.3.3- Tooltips exploration. Different tooltips and pointers help users navigate and learn the system with ease. Tooltips appear in the system when a user hovers over an interface element.

Figure 4.3.4- System help and tips. Along with the pointers and tooltips on the side, the other system-generated help, and tip explorations.

4.4. XAI INTERFACE DESIGN

Question

How can feature visualization, textual explanations, and interface features together build mental models in AI novices so that they might practice agency in such a way that the output matches their intention while allowing for exploration?

Study 4

The previous studies demonstrate how feature visualization, forms of explanations, and interface features can together build an AI novice's agency to change input in such a way so that the output matches novice's intent in the GAN system. This study 4 summarizes, brings together all the other studies, and learnings from the studies together into a working interface design.

Here it is important to mention that the user intent is not to generate the exact image but to constrain some parts of the image and see what system can generate based on the specified constraints. The generated output is a collaborative effort of a user with the system to explore the unconventional. A user with agency can constrain some parts that they would like to either see or not see in the generated output. The system works within those given user constraints to generate images a user otherwise couldn't think of or create within seconds. The high fidelity wireframes can be accessed through [link in references section 6.1](#) and scenario video by [link in references section 6.1](#).

Ideation Study 4

Scenario:

Rose comes to the system for the first time and she gets a walkthrough of the system. She decides to do some explorations to get to know the system. On finding some interesting bits generated by the system she delves into the details a little more, exports an image, and leaves. She decides to come back to the system again to explore generations for her game design. She observes that the system changes explanations according to her preference. She is now able to explore, build mental models of the system, and get some generations to use for her game scene prototype.

Solution Ideation:

Rose enters the system and is greeted by exploration walkthroughs. The system implicitly records Rose’s likes and her ways of using these process walkthroughs and tooltips. She enters each tab and is taken through examples. She enters the final tab where she learns about the other tabs and the role. She plays with it to understand the system’s overarching functioning. Finally, she decides to see what kind of generations the landscape generator produces to see if this tool will be useful to her. She finds fun new generations for game landscape design. She tries to dig into how the system was generating some of the images. She goes inside the latent space tab, while the system hints and tips give her an idea that she can set constraints and explore with the given sliders and other interface features. After getting a view of the system’s capabilities and satisfied with the agency and innovative explorations that the system affords she leaves the system to return later. When she comes back to the system, she has a general idea about the specific kinds of generations she might want. She starts with creating a new landscape and lets the system generate something according to her initial constraints. The guiding walkthroughs on the system change a bit and she likes them more this way. She thinks about exploring latent space for the sky and explores some interpolating options. She isn’t entirely familiar with the interface options and the naming but as she is playing with system hints she keeps developing the systems’ working’s mental model. Generations or the outputs after using interface options give her a sense of their functionality. She tries element bucket and then after some beautiful generations with the system, she gets to choose specific areas and learns about internal activations. Finally, after developing an understanding of the system she once again looks at the tab perspective and looks at the generated dissected between layers and gets a good understanding of the system working in the backend. She leaves the system with a generation for the game prototype and an understanding of the system working.

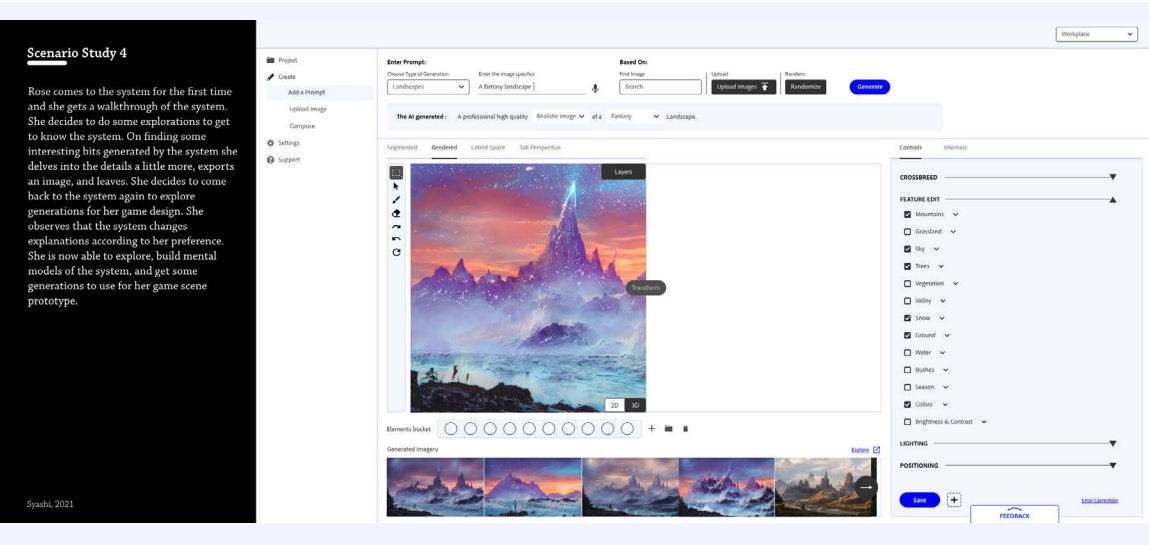


Figure 4.4.1- Click on the figure to watch the scenario video.

DISCUSSIONS

- ∞ 5.1 Design Principles
- ∞ 5.2 Future Work
- ∞ 5.3 Conclusion

5.1 DESIGN PRINCIPLES

During this investigation, I designed interface features and visual and textual explanations that will not only help build user’s mental models of the system, but also give them agency to collaborate with the system. While designing, I followed certain principles that can possibly help future designers on a GAN or AI based system.

Easing intention setting: An ease of setting intentions of the system which tells the system what the user is looking for and at the same time system explaining what it understands from the users intention. A user can then edit the parts that the system understands so that it explains the user’s position to system and systems position to the user. It creates a space for dialogue between system and the user.

Using metaphors to design and explain: I used a metaphor of layers for explaining different tabs to a user. Concept of layers is something a designer would understand and so having the system model explained through the layering metaphor eases the load of explaining complex concepts.

Tips, walkthroughs, and pointers: Tips and walkthroughs should exist to guide a user on an adaptive interface. Prompts like pointers or highlights that push a user to explore the internals and interface help the user get accustomed to the interface. A user stuck or unable to take any action should be provided with tips that enables them to explore more of the interface. Even the hidden elements or the tabs can be hard to find or access and this can be made easy with the help of tips. These tips, explanations, and walkthroughs should change and appear according to users preferences.

Labeling of elements: A balance should be there while naming interface elements, they should not be overly complex or technical to remember and should inform the user of the correct technical terms of use. A user can get scared or overwhelmed when familiarity is less and there is a lot to learn. This creates a need to divide familiarizing the interface concepts to the user into phases. This investigation tries to familiarize users with terms while trying to provide the explanations via easy text based language and animations.

Feedback: The ability to feedback can help improve the system. The human-in-the-loop system requires an ability for users to feedback the

system. It gets tricky to ask for feedback especially with the systems which can be biased and hurt user sentiments. Another facet to think about is that human beings are biased and those biases can also trickle into the system if all feedback is taken. So a moderator would be required to filter feedback. Two kinds of feedback can be asked from a user; one would be error correction feedback and the other one is system functioning feedback. Error correction feedback is important to be captured in a human-in-the-loop system. System feedback is a general system functioning and user satisfaction feedback. Also feedback can both be explicit and implicit. Both kinds of feedback should be accessible to the user.

Transparency and privacy of users data: A user should be able to see the data mined and implicitly taken from the clicks made by them on the interface. As the machine learning works on detecting patterns to personalize settings. It would be on the designer to assign a designated place on the system where all of the information about the user is available to and editable by the user. Any default settings to gather data should be avoided and users should be given an option to opt in to gather data for system improvement. Most of the models should be stored locally, if not then the users should be informed about it and its impact on their privacy.

Breakdown of system generations: The system's generated output should be systematically broken down and explained to the user. GAN generations can be hard to understand and so breaking the output down can also increase trust of the user in the system. I have tried to explain system generations using the different methods of feature visualization. Both textual and graphical explanations help users understand system outputs.

5.2 FUTURE WORK

There is a potential for the GAN system used in this research work to expand in depth and breadth in the future as new researchers get published. As of now the system is limited to just image based models and specifically landscape and character generator models. This research in the future could address other models that are not image based like audio and textual. These GAN models would necessitate discussions around the ethical aspects of models like models concerning human faces. This project only addressed a small part of the bigger question of making ethical, trustable, and understandable GAN systems. Societal biases in relation to GAN systems must be addressed thoughtfully in the future.

Research in this field often caters to experts and engineers. As newer research gets published on explainability and the increasing capabilities of AI systems, it would be important to incorporate this new knowledge into this system in a manner that would make it accessible to an AI novice. Another important aspect is that not all interactions and detailed interface options are discussed in this investigation and they should be considered in the future as that will improve the users experience of the system.

To move forward with this investigation it will be important for a designer to work with engineers, researchers, and content strategists. Engineers can inform the models used and the tags or categories that the system understands after training. Designers can use this given knowledge of the system and models to design the interface features. Researchers can steer designers' understanding of users' mental models of the interface and the system's explainability. Design research is important and I would think that evaluations as mentioned in the studies should be conducted with the user to test out the design. The visual studies that I conducted are based on theory and precedents available in the market. I would want to test it out with actual users to see how they interact and if the system is able to provide agency, trust, and interpretability. Through this design research I will be interested in discovering more about users' perspectives on interface design. Interesting observations can come from the research that could aid the system design. Content strategists can help designers in naming and working on language as that is an important part of this system.

I am using machine learning capabilities not only to generate the content but also to take user input and parse it both in audio and

textual form. Machine learning is also used to identify patterns of users’ usage of the system. This will help users collaborate with and understand the system. There are multiple ways of providing textual explanations and visualizations but if this machine learning system can identify patterns through users’ implicit feedback then the best form of explanations and visuals can be provided to the user that will help build their mental models for the system.

5.3 CONCLUSION

Numerous artists are working directly with GAN models to make interesting pieces of work but this process isn’t accessible to AI novices. An AI interface system that caters to AI novice users—in particular a system that lets them explore the system internals, that explains the system, and that grants them the agency to collaborate—does not currently exist. I designed a human-in-the-loop system that serves the dual purpose of improving the system and users’ understanding of the system. Such interpretable systems encourage people in professions like design or art who have no knowledge of AI technology to learn a system and experiment with it. I believe my own research will help provide such interpretability to the user.

GAN systems can produce unimagined or unconventional works which designers and artists can tap into to design better and explore, leading us to computational creativity! Working in collaboration with a GAN system—given the benefits of interpretability, explainability, and agency—will change how designers interact with machines and alter our design processes. This investigation works at the intersection of automation and manual work, combining human intelligence with GAN generations. This powerful ability to collaborate and generate with the GAN system needs a trustable and explainable system that allows users to understand why an AI makes error or biased decisions and to correct it. Currently the GAN system which this research explores focuses on image based models but GANs can work well with text and audio as well. In the future there will be a need for creating systems that give agency to users for audio and textual GAN models.

Finally, through this research I have explored a seamless way of integrating such a system on top of an existing software in the form of an API. Such an integration will serve as an exploratory tool and supplement working of the software as well. API integration creates a plug and play system that requires the user to switch-on setting for an API and easily explore while using an underlying software of their choice. This enhanced ability to use GAN models on top of any creative software helps increase user’s exploratory powers without compromising underlying software’s efficiency.

REFERENCES

- ∞ 6.1 Links
- ∞ 6.2 Image Credits
- ∞ 6.3 References

6.1 LINKS

These links can also be accessed through an online document available at: <https://tinyurl.com/np4aewve>

Study 1 wireframes: <https://xd.adobe.com/view/95268e8d-eded-4ce1-9852-2b502f54bc57-c075/>

Video 4.1.1: <https://youtu.be/rUQYUR6VQz8>

Study 2 wireframes: <https://xd.adobe.com/view/0fb3606d-31d3-4c99-85b6-1c5222533440-5ae0/>

Video 4.2.1: <https://youtu.be/tmd1mNZxDB8>

Study 3 wireframes: <https://xd.adobe.com/view/4a2393ae-2e03-4d42-9c98-2d3c88321d5d-1f70/>

Video 4.3.1: <https://youtu.be/4vYpIc0xZQw>

Study 4 wireframes: <https://xd.adobe.com/view/fb775064-d242-4f6b-a430-b61501e1e5d5-ae24/>

Video 4.4.1: <https://youtu.be/5txJTqm8shU>

6.2 IMAGE CREDITS

Artbreeder: <https://www.artbreeder.com/>

Evermotion: https://evermotion.org/shop/show_product/tree-24-am171-archmodels/13147

Nvidia: <http://nvidia-research-mingyuliu.com/gaugan/>

Distill: <https://distill.pub/2019/activation-atlas/>

**** Images are used only for explaining concepts in the research and in no way for monetary puporse.**

6.3 REFERENCES

Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). *Power to the people: The role of humans in interactive machine learning*. *AI Magazine* 35, 4 (2014), 105–120.

Bau, D., Zhu, J.-Y., Strobel, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2018). *GAN dissection: Visualizing and understanding generative adversarial networks*. *ArXiv:1811.10597 [Cs]*. <http://arxiv.org/abs/1811.10597>

Bau, D., Zhu, J.-Y., Strobel, H., Zhou, B., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2020). *On the units of GANs (Extended Abstract)*. *ArXiv:1901.09887 [Cs, Stat]*. <http://arxiv.org/abs/1901.09887>

Boukhelifa, N., Bezerianos, A., & Lutton, E. (2018). *Evaluation of interactive machine learning systems*. *Human and Machine Learning Human–Computer Interaction Series*, 341-360. doi:10.1007/978-3-319-90403-0_17

Blayone, T. J. B. (2019). *Theorising effective uses of digital technology with activity theory*. *Technology, Pedagogy and Education*, 28(4), 447–462. <https://doi.org/10.1080/1475939X.2019.1645728>

Browne, K., Swift, B., & Gardner, H. (2018). *Critical challenges for the visual representation of deep neural networks*. *Human and Machine Learning Human–Computer Interaction Series*, 119-136. doi:10.1007/978-3-319-90403-0_7

Bryan, N. J., G. J. Mysore, and G. Wang. 2014. “ISSE: An interactive source separation editor.” *ACM CHI Conference on Human Factors in Computing Systems*. Toronto.

Carter, S., & Nielsen, M. (2017). *Using artificial intelligence to augment human intelligence*. *Distill*, 2(12), e9. <https://doi.org/10.23915/distill.00009>

Carter, S., Armstrong, Z., Schubert, L., Johnson, I., & Olah, C. (2019). *Activation atlas*. *Distill*, 4(3). doi:10.23915/distill.00015

Chrysos, G. G., Kossai, J., Yu, Z., & Anandkumar, A. (2020). *Unsupervised controllable generation with self-training*. *ArXiv:2007.09250 [Cs, Stat]*. <http://arxiv.org/abs/2007.09250>

Colton, S. & Wiggins, Geraint. (2012). *Computational creativity: The final frontier?*. *Frontiers in Artificial Intelligence and Applications*. 242. 21-26. 10.3233/978-1-61499-098-7-21.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., & Bharath, A. A. (2018). *Generative adversarial networks: An overview*. *IEEE Signal Processing Magazine*, 35(1), 53–65. <https://doi.org/10.1109/MSP.2017.2765202>

Deltorn, J.-M. (2017). *Deep creations: Intellectual property and the automata*. *Frontiers in Digital Humanities*, 4. <https://doi.org/10.3389/fdigh.2017.00003>

Dudley, J. J., & Kristensson, P. O. (2018). *A review of user interface design for interactive machine learning*. *ACM Transactions on Interactive Intelligent Systems*, 8(2), 8:1–8:37. <https://doi.org/10.1145/3185517>

Fogarty, J., Tan, D., Kapoor, A., & Winder, S. (2008). *CueFlik: Interactive concept learning in image search*. *Proceeding of the Twenty-Sixth Annual CHI Conference on Human Factors in Computing Systems - CHI '08*, 29. <https://doi.org/10.1145/1357054.1357061>

Goff, K. L., Rey, A., Haggard, P., Oullier, O., & Berberian, B. (2018). *Agency modulates interactions with automation technologies*. *Ergonomics*, 61(9), 1282–1297. <https://doi.org/10.1080/00140139.2018.1468493>

Ghosh, A., Zhang, R., Dokania, P. K., Wang, O., Efros, A. A., Torr, P. H. S., & Shechtman, E. (2019). *Interactive sketch & fill: multiclass sketch-to-image translation*. *ArXiv:1909.11081 [Cs, Eess]*. <http://arxiv.org/abs/1909.11081>

He, Z., Zuo, W., Kan, M., Shan, S., & Chen, X. (2018). *AttGAN: Facial attribute editing by only changing what you want*. *ArXiv:1711.10678 [Cs, Stat]*. <http://arxiv.org/abs/1711.10678>

Heim, E. (2019). *Constrained generative adversarial networks for interactive image generation*. *ArXiv:1904.02526 [Cs, Stat]*. <http://arxiv.org/abs/1904.02526>

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2019). *Metrics for explainable AI: Challenges and prospects*. *ArXiv:1812.04608 [Cs]*. <http://arxiv.org/abs/1812.04608>

Hollis, V., Pekurovsky, A., Wu, E., & Whittaker, S. (2018). *On being told how we feel: How algorithmic sensor feedback influences emotion perception*. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(3), 114:1–114:31. <https://doi.org/10.1145/3264924>

Holzinger, A., Plass, M., Kickmeier-Rust, M., Holzinger, K., Crişan, G. C., Camelia-M Pinteă, & Palade, V. (2019). *Interactive machine learning: Experimental evidence for the human in the algorithmic loop*. *Applied Intelligence*, 49(7), 2401-2414. <http://doi.org/10.1007/s10489-018-1361-5>

Jasper, R. J., & Blaha, L. M. (2017). *Interface metaphors for interactive machine learning*. *Neurocognition and machine learning* (Vol. 10284, pp. 521–534). Springer International Publishing. https://doi.org/10.1007/978-3-319-58628-1_39

Kahng, M., Thorat, N., Chau, D. H., Viégas, F. B., & Wattenberg, M. (2019). *GAN lab: Understanding complex deep generative models using interactive visual experimentation*. *IEEE Transactions on Visualization and Computer Graphics*, 25(1), 310–320. <https://doi.org/10.1109/TVCG.2018.2864500>

Lieberman, H. (2009). *User interface goals, AI opportunities*. *AI Magazine*, 30(4), 16. <https://doi.org/10.1609/aimag.v30i4.2266>

Magassouba, A., Sugiura, K., Quoc, A. T., & Kawai, H. (2019). *Understanding natural language instructions for fetching daily objects using GAN-based multimodal target-source classification*. *ArXiv:1906.06830 [Cs]*. <http://arxiv.org/abs/1906.06830>

Mohseni, S., Zarei, N., & Ragan, E. D. (2020). *A Multidisciplinary survey and framework for design and evaluation of explainable AI systems*. *ArXiv:1811.11839 [Cs]*. <http://arxiv.org/abs/1811.11839>

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). *The building blocks of interpretability*. *Distill*, 3(3), e10. <https://doi.org/10.23915/distill.00010>

Olah, C., Mordvintsev, A., & Schubert, L. (2019, August 28). *Feature visualization*. Retrieved September 27, 2020, from <https://distill.pub/2017/feature-visualization/>

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2020, January 10). *The building blocks of interpretability*. Retrieved September 27, 2020, from <https://distill.pub/2018/building-blocks/>

Putzu, L., Piras, L., & Giacinto, G. (2020). *Convolutional neural networks for relevance feedback in content based image retrieval*. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-020-09292-9>

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. *ArXiv:1708.08296 [Cs, Stat]*. <http://arxiv.org/abs/1708.08296>

Sbai, O., Elhoseiny, M., Bordes, A., LeCun, Y., & Couprie, C. (2019). *DesIGN: Design inspiration from generative networks*. *Computer Vision – ECCV 2018 Workshops* (Vol. 11131, pp. 37–44). Springer International Publishing. https://doi.org/10.1007/978-3-030-11015-4_5

Schnabel, T., Bennett, P. N., & Joachims, T. (2019). *Shaping feedback data in recommender systems with interventions based on information foraging theory*. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 546–554. <https://doi.org/10.1145/3289600.3290974>

Shama, F., Mechrez, R., Shoshan, A., & Zelnik-Manor, L. (2018). *Adversarial feedback loop*. *ArXiv:1811.08126 [Cs]*. <http://arxiv.org/abs/1811.08126>

Shih, P. C. (2018). *Beyond human-in-the-loop: Empowering end-users with transparent machine learning*. *Human and Machine Learning* (pp. 37–54). Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_3

Shilman, M., Tan, D.S. and Simard, P. (2006). *CueTIP: A mixed-initiative interface for correcting handwriting errors*. *Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2006)*, 323-332.

Springer, A., & Whittaker, S. (2019). *Progressive disclosure: Empirically motivated approaches to designing effective transparency*. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 107–120. <https://doi.org/10.1145/3301275.3302322>

Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., & Herlocker, J. (2007). *Toward harnessing user feedback for machine learning*. *Proceedings of the 12th International Conference on Intelligent User Interfaces - IUI '07*, 82. <https://doi.org/10.1145/1216295.1216316>

Tripathi, D., Medlar, A., & Glowacka, D. (2019). *How relevance feedback is framed affects user experience, but not behaviour*. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 307–311. <https://doi.org/10.1145/3295750.3298957>

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). *Designing theory-driven user-centric explainable AI*. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3290605.3300831>

Zerilli, J., Knott, A., Maclaurin, J., & Gavaghan, C. (2019). *Algorithmic decision-making and the control problem*. *Minds and Machines*, 29(4), 555–578. <https://doi.org/10.1007/s11023-019-09513-7>

Zhou, J., & Chen, F. (2018). *2D Transparency space—Bring domain users and machine learning experts together*. In J. Zhou & F. Chen (Eds.), *Human and Machine Learning* (pp. 3–19). Springer International Publishing. https://doi.org/10.1007/978-3-319-90403-0_1

