Title	Labeling : reflecting gender bias in human-machine relationship through an critical design concept
Sub Title	
Author	陳, 佳欣(Chen, Chia-Hsin) 南澤, 孝太(Minamizawa, Kōta)
Publisher	慶應義塾大学大学院メディアデザイン研究科
Publication year	2020
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2020年度メディアデザイン学 第805号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40001001-00002020-0805

Master's Thesis
Academic Year 2020

Labeling: Reflecting Gender Bias in Human-Machine Relationship through an Critical Design Concept

Keio University
Graduate School of Media Design

Chia-Hsin Chen

A Master's Thesis
submitted to Keio University Graduate School of Media Design
in partial fulfillment of the requirements for the degree of
Master of Media Design

Chia-Hsin Chen

Master's Thesis Advisory Committee:

Professor Kouta Minamizawa	(Main Research Supervisor)
Professor Keiko Okawa	(Sub Research Supervisor)
Professor Hideki Sunahara	(Sub Research Supervisor)

Master's Thesis Review Committee:

Professor Kouta Minamizawa	(Chair)
Professor Keiko Okawa	(Co-Reviewer)
Professor Hideki Sunahara	(Co-Reviewer)

Abstract of Master's Thesis of Academic Year 2020

Labeling: Reflecting Gender Bias in Human-Machine Relationship through an Critical Design Concept

Category: Design

Summary

No matter how generation innovated, human being, a creature with individual diversity, lives with unconsciously biased in constant. With the advance of technology, biased human being created Artificial intelligence that related to complicated ethics. Nevertheless, when algorithmic bias made a false counter-charge, human being chooses to filter out errors without facing up the problem that is human nature. In this vicious cycle of human-machine relationship, if started exploring this issues by critical design methodology, choosing to apply gender bias in algorithms as a reflective medium, how would biased machine reflect to human-machine relationship nowadays? How might it provoke an empathic understanding of individual diversity to human being?

Starting from questioning above problems with critical design mindset, this research explored the discussion with participants by the creation named "Labeling", three reflective installations that algorithmic bias was applied, to reflect gender bias that hidden in human-machine relationship. This research aims to provoke participant's introspective contemplation on both gender bias and human-machine relationship nowadays by designing provocative installations.

Keywords:

Critical Design, Gender Labeling, Gender Bias, Scientific Philosophy, Machine Ethics, Human-Machine Relationship

Keio University Graduate School of Media Design

Chia-Hsin Chen

Contents

List of Figures

List of Tables

Acknowledgements

I would first like to thank my main research supervisor Prof. Kouta MINAMIZAWA of the Graduate School of Media Design at Keio University for his supportive feedback. Many people around me have questioned even this research, and he never says no and let me freely finish it. When I lost the direction, he always guiding me in the right direction. On the other hand, same as the first people that need to be mentioned, I would also like to thank Marcelo Padovani firstly, the mate also from Embodied Media, this research will never be done without his support. Marcelo knows what I want to express, the tacit understanding between Marcelo and me is beyond words, same as my appreciation to Marcelo is beyond words. I would also like to thank people who were involved in the research project: Keitaro TSUCHIYA, Yurike CHANDRA, Yi-Fan ZHUANG.

I would like to thank my co-advisor, Prof. Keiko OKAWA, for the understanding and supportive feedback for this research. The door to Prof. KUNZE office was always open whenever I ran into a trouble spot or had a question about my research or writing. He consistently allowed this paper to be my work but steered me in the right direction whenever he thought I needed it. On the other hand, I appreciate Prof. WALDMAN for the supportive understanding when there was nobody to understand this research.

I would also like to thank people involved in the survey for this research project: Hera Wen Cheng, Bessy Shih-Hua Lin, Grace Shou-En Wang, Sheena Liman, Cong Gao, other mates from Global Education. The supportive backing from Embodied Media: Shuang Hao, Harry Krekoukiotis. It is my pleasure to be the same year's intake with these talented people.

Without their passionate participation and input, the validation survey could not have been successfully conducted.

Chapter 1
Introduction

With the rapid progress of science and technology, human beings live with various types of human-centered technology to achieve innovative futures. Taking our daily life as an example, by making sure a facial recognition system could scan the face, users could easily unlock our phone screen. Moreover, some industries have started to give it a try on applying Artificial Intelligence into both human resource information and management systems. Even for government relations, some technologists have started to support public sectors to apply both Artificial Intelligence and computer vision into Law Enforcement, from fining violate a traffic regulation more precisely to addressing crime detection.

However, as for issues that exposed in the examples mentioned above, both applications of facial recognition and human resource relations have been pointed out that there are serious gender racist issues needed to be facing up. When human-centered technology has set into involving the discussion of deciding a human being is potentially a good employee or not, even helping people to define another person's gender based on their appearance, what premises have set beyond these decisions?

Although human beings, as an innovative creator, created Artificial intelligence and other human-centered technologies related to complicated ethics, human beings live with unconsciously biased in constant. In this premise, things would be more complicated than an innovative generation, that human being has kept chasing for. Nevertheless, when algorithmic bias made a false counter-charge, human beings choosing to filter out errors without facing the origin of the problem is precisely human nature.

If started to face up the root of this issue instead of filtering out errors happened in technology, and viewing the origin of the gender bias from the angle of gender labeling, human learned to define gender roles by recognizing the appearance fea-

ture of male or female, and labeled gender with a socialized and binary definition subconsciously [1]. In this premise, gender bias is unconsciously manifested, and complexity is the fundamental reason behind gender equality's indistinct goals.

Similarly, third-wave HCI has focused on contributing to continuously complex representations of users [2]. Emerging technologies that are versatile in machine learning, such as gender classification, have been surrounding our daily lives [3], and interestingly, the process of gender classification is similar to gender labeling happened in human society. Moreover, the accuracy of classification systems depends on the data inputted by human, algorithms could be biased if trained on biased data [4].

Humans could see what influence algorithmic biases have brought on racial issues [5] and gender issues [6], start to solve the issue with filtering algorithms, even though humans could not precisely define both gender boundary and unconscious gender biases in our daily lives yet [7]. When algorithmic bias made a false counter-charge, human beings choose to filter out errors without facing up the problem is precisely human nature.

However, what if there are exploratory possibilities that biased machines could be applied as a reflective medium, becoming more than just the inferior design needed to be erased? How might it provoke human's introspection gender bias in both algorithms and socialized gender boundaries, becoming more than just defaulting the unpleasant truth?

In order to explore potentially hidden and alternative design values, this research start from questioning the present state with critical design [8], a design research methodology foregrounds the ethics of design practice, to reveal those values in the context of gender labeling, both by humans and by algorithms. In order to reflect the issues, three installations have been designed as the provocative mediums to bring participants into the context it built, triggering an empathic understanding on individual diversity, and questioning the present stage of human-machine relationship with participants.

Chapter 2
Literature Review

To explore both similarities and potential application between gender labeling from human context and gender classification from algorithmic context, the literature review of this research started the discussion from labeling effect, diving into to machine ethics as the way of exploring human-machine relationship, and introducing critical design mindset that applied as the methodology of this research, including the introduction of related works.

2.1. Gender Labeling as the Origin of Gender Bias

Labeling theory, a branch of symbolic interaction theory, explains how the identity and behavior of human beings are influenced by how society has classified them [9]. Gender labeling, a branch of labeling theory, has mainly focused on labeling theory in gender issues, indicates how human beings learned to classify biological gender based on social and institutional values consciously [10].

As stated above, if discussed the origin of gender boundary and gender bias from the angle of gender labeling, human learned to define gender roles by recognizing the appearance feature of male or female, and subconsciously labeled gender with the socialized and binary definition. In this premise, unconscious gender bias is manifested invisibly(Figure 2.1).

Moreover, it is hard to precisely define the baseline of gender bias due to cultural diversity and complicated individual factors in human society. Even gender bias can be conscious bias or unconscious bias and may manifest in various ways around our lives [11]. Therefore, the complexity of gender bias has generally been regarded as the fundamental reason behind gender equality's indistinct goals, and

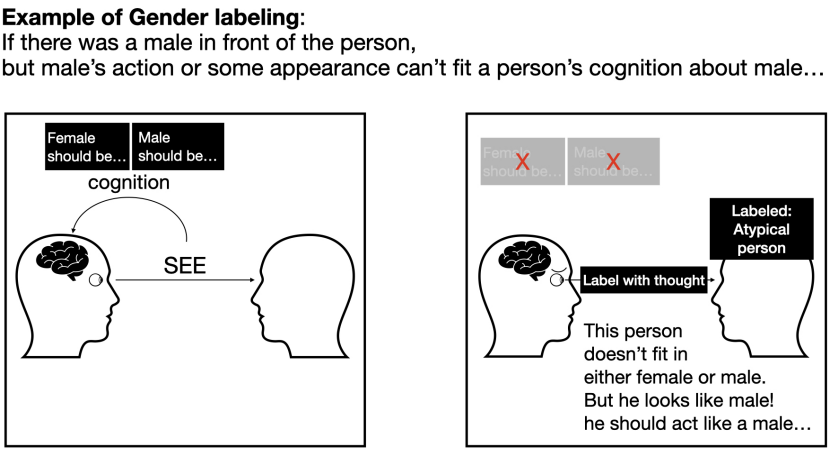


Figure 2.1 An example of gender labeling

the unconscious prejudices of humanity are a vicious cycle of gender issues.

2.2. Human-Machine Relationship

2.2.1 Machine Ethics

The beginning of modern Artificial intelligence (hereinafter referred to as "AI") was assumed to describe human thinking as a symbolic system by classical philosophers, and the field of AI research wasn't formally founded until 1956 at Dartmouth conference, a workshop held on the campus of Dartmouth College during the summer of 1956, where the term "Artificial intelligence" was coined. After passing the golden years and the first AI winter from 1956 to 1980, boom and bust from 1980 to 1993, AI research has finally come to the present stage, which is the stage of diving into deep learning, big data, and artificial general intelligence [12].

In the present stage, AI could be applied to biological image-recognition, and the well-known application of biological image-recognition is evolved facial recognition, which has evolved the accuracy with neural network technology. The present wave of these kinds of applications has been indicated to be could recognise and "understand" human beings' needs in an innovative way, as similar as

the smartphone could be unlocked once facial recognition system assumed that we look like human beings after comparing with its database.

Nevertheless, once machine and AI technology have implicated in human-centered issues that ethics, morals, and human diversity are all included, things would become more complicated than we could ever imagine.

2.2.2 Similarities Between Gender Labeling and Gender Classification

Since gender labeling, a branch of labeling theory in symbolic interaction theory as stated in Chapter 2.1, is applied as both prime theory for problematization and the main subject in this research, how similar would Gender labeling happened in human society, and gender classification would be? To explore the similarities between gender labeling and gender classification, an exploring of both gender labeling processes in a human context and supervised gender classification in algorithmic context is conducted by sketching both processes and making the comparative approach(Figure 2.2).

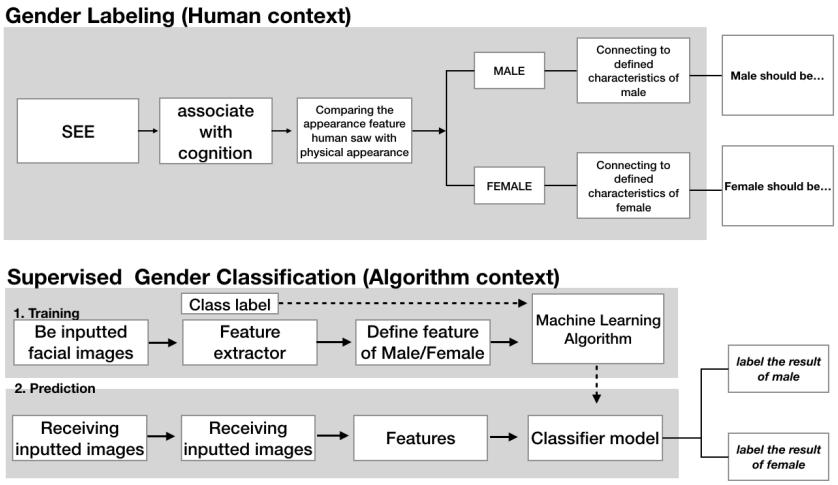


Figure 2.2 Process sketch of both gender labeling and gender classification

The cognition process of both human context and algorithmic context is generally similar to each other: Receiving the appearance, comparing it with feature extractor, defining the result, and labeling the target is male or female. Besides, apart from the different cultural contexts and personal aspects, the property of gender labeling is vaguely defined by the majority, which means the consensus of most people. In contrast, algorithm context will be firstly influenced by the face database, which could also regard as the majority in machine learning [13].

Plus, as stated previously, discussing from the angle of machine ethics in gender issues, gender classification is also regarded as the central issue in machine ethics, and it is still hard to trace a clear goal due to complicated human diversity. Moreover, by drawing inferences from stated above, the indifference attitude might not only addressed the constancy of unconscious gender bias but also influence on nonhuman context undoubtedly.

To trigger people to rethink the gender bias hidden in our daily lives, could algorithmic bias in gender classification be applied as provocative works? How might bias happened in machine ethics become provocative works, instead of being regarded as error that needed to be filtered out?

How could human beings, the creator of technology, avoid discussing humanity's complexity, but keep creating human-centered technology?

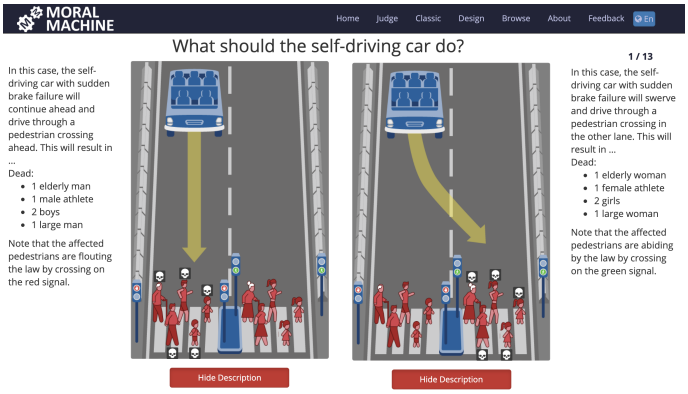
2.2.3 Provocative Works Related to Machine Ethics

To reflect the status quo of human-machine relationships, more and more critical design researchers, creators, and artists express their messages by creating provocative works related to human-machine relationship and machine ethics. In this section, related researches challenging humanity by questioning the human-machine relationship, AI experiments that got the unexpected failure, and related works that were also inspired by the similarities between humans and machines, are detailed below.

The Moral Machine experiment designed by MIT Media Lab [14][1], questions whether human beings could teach robots right from wrong in moral dilemmas by

1 The Moral Machine
`https://www.moralmachine.net/`

showing moral dilemmas to human beings and letting them choose the lesser of two evils(Figure 2.3). There are no correct answers to each question due to the various social values in our society. However, The Moral Machine has pointed out that human beings, the creator of technology, expect future technologies without thinking deeper.



(Source: The Moral Machine experiment [14])

Figure 2.3 The Moral Machine experiment

If without thinking deeper, what would happen? Tay [15], an AI chatbot made by Microsoft, ultimately showcased how human behavior could influence AI chatbot to be a racist [16]. In 2016, Microsoft had released Tay, a girly AI Chatbot, to experiment on conversational understanding with humans on social networks. She was made for testing the intersection of machine learning, natural language processing, and social networks. Trough tweets or direct messages to Tay, people could not only chat but teach Tay new vocabulary. Nevertheless, after only a few hours since Tay had been released on Twitter, she started tweeting extremely offensive words(Figure 2.4). Twitter users started registering their outrage, and the final decision that Microsoft made is to lock Tay's Twitter account. Based on the reflection that Tay had reflected on human behavior, several researches related to the human-machine relationship have also indicated a boundary called "ethic" between human-machine symbiosis [17].

In the field of media art, there are also many media art artist has expressed

(Source: Microsoft's Artificial Intelligence Tay Became a "Racist Nazi" in less than 24 Hours [15])

Figure 2.4 Microsoft's Artificial Intelligence Tay Became a 'Racist Nazi' in less than 24 Hours.

their thought on the human-machine relationship. Project "Zoom Pavilion" [18][2] is an interactive installation that inviting viewers to enter the space and being involuntarily surveyed, with the CCTV immediately played back at exhibited space, the work keep monitoring viewers' physicality (Figure 2.5). Viewers felt unnerving otherworldly, and also receive uncertain feelings during the experience. Through this work, the artist has evoked a form of surveillance that is inherent in innovative urban life; we are constantly filmed, tracked, subjected to algorithms and archived, the boundary of human-machine relationship has subtly is becoming blurred in this generation.

On the other hand, "Us and Them"(Figure 2.6) [19][3] is a multi-modal installation, applied the data set that trained on two hundred thousand tweets from accounts identified as bots, after the 2016 United State presidential election and consequently evicted from Twitter. Inspired by the phenomenon of social media use, "Us and Them" features 20 machine-learning-driven printers which endlessly spew AI-generated political tweets by imaginary and generated people. It invites

2 Zoom pavilion, RAFAEL LOZANO-HEMMER
`http://www.lozano-hemmer.com/zoom_pavilion.php/`

3 Us and Them, Mike Tyka, AI Art Gallery
`http://www.aiartonline.com/art/mike-tyka/`

the viewer to re-think their relationship with the machine we live inside, even discuss the fake news issues nowadays.

(Source: Zoom Pavilion, RAFAEL LOZANO-HEMMER [18])

Figure 2.5 Zoom Pavilion (2015)

(Source: Us and Them, Kinetic Installation (2018) Commissioned by Seoul Museum of Art [19])

Figure 2.6 Us and Them (2018)

2.2.4 When AI Bias Happens, Whose Fault?

When the indifference attitude of human beings not only addressed the constancy of unconscious gender bias but also influence on nonhuman context undoubtedly, what would happen? Which groups would be influenced?

Consider AI in human-resourced relations as the first example; the use of AI and algorithms in recruitment is expected to grow [20]. Nevertheless, there are well-known AI bias issues happens in this application: Amazon's machine-learning specialists exposed a big problem that their new recruiting engine prefers to choose male job seeker. There are two main indicated reasons which caused this failed AI application.

Firstly, because Amazon's computer models were trained to select applicants. By observing resumes submitted to the company over ten years, resumes were most came from male. Therefore, it reflected male dominance across the tech industry.

On the other hand, human beings, as a creature with complicated humanity, have unconscious biases inevitably. This inevitable truth has been unconsciously

transferred into actual action, such as human beings hadn't considered the situation when creating an algorithm.

As for the solution made by Amazon, the related members of AI in human-resourced relation has focused using much watered-down version of the recruiting algorithm, to help with dealing basic tasks, such as screening out same candidate profiles from databases.

Here is the second example that happened in the application of facial recognition. The use of facial recognition nowadays is widely provided by Amazon, iBM, and Microsoft. Plus, these facial databases are provided to police and government relations. Nevertheless, iBM, Amazon, and Microsoft have continuously announced the decision to stop providing facial recognition to police since June 2020 [21]. These companies' chose to stop facial recognition because of the criticism of technology's inaccuracy, and several studies have indicated that facial recognition has been indicated that its algorithms fail to detect black and brown faces accurately [6]. Moreover, it might cause a massive issue if kept applying immature human-centered technologies into our social structure.

However, if made an overall conclusion of the examples stated above, we could found that solutions of all failed AI applications and biased algorithms are being stopped or filtered the algorithms. There is no solution for unconscious biases from human beings, the creator of AI applications and algorithms.

If did not choose to face up what humanity has influenced on both AI application and algorithms, how could human being, the creator of AI application and further human-centered technologies, prevent these problems from being a vicious cycle?

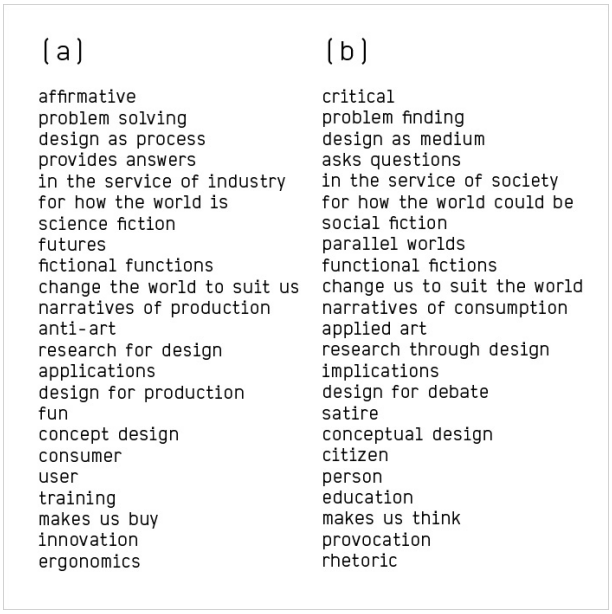
2.3. Critical Design

2.3.1 What is Critical Design?

In order to explore the problem stated in the previous section, critical design, a conceptual design methodology that foregrounds the ethics of design practice to explore both potentially hidden and alternative design values, is applying to this research as both design methodology and design mindset.

Affirmative design includes proposals, however humble, modest or straightfor-

ward design, and the goal is to solve the problem. In contrast, critical design starts with problem finding, questioning how the world could be, and designing a reflection to make people think. Refer to the comparison of affirmative design and critical design(Figure 2.7)[4], affirmative design is the general design methodology that researchers use to achieve with, and its context is generally in the service of industry. On the contrary, the context of critical design is in the service of society.



(a)	(b)
affirmative	critical
problem solving	problem finding
design as process	design as medium
provides answers	asks questions
in the service of industry	in the service of society
for how the world is	for how the world could be
science fiction	social fiction
futures	parallel worlds
fictional functions	functional fictions
change the world to suit us	change us to suit the world
narratives of production	narratives of consumption
anti-art	applied art
research for design	research through design
applications	implications
design for production	design for debate
fun	satire
concept design	conceptual design
consumer	citizen
user	person
training	education
makes us buy	makes us think
innovation	provocation
ergonomics	rhetoric

(Source: BEYOND DESIGN THINKING: AN INCOMPLETE DESIGN TAXONOMY [22])

Figure 2.7 The comparison of Affirmative Design and Critical Design

Plus, the goal of critical design is to push design research beyond an agenda on reinforcing values of consumer culture and to embody cultural critique in designed artifacts instead, and it mainly aims to design the reflection [23].

Moreover, the core value of critical design is challenging facts by both speculative design proposals and designing reflective mediums. The outcome of critical design is usually abstract and endless, but provoking introspection on those facts, and bringing the new perspective to participants by reflective design works. There-

4 BEYOND DESIGN THINKING: AN INCOMPLETE DESIGN TAXONOMY
http://www.cd-cf.org/articles/beyond-design-thinking//

fore, different from affirmative design, there are no specific target users in critical design methodology.

As for the output of critical design in this research, in order to provoke people's thinking on status quo of both gender bias and human-machine relationship, the designed output of critical design is a reflective medium, which is a provocative work to guide people dive into the context of design concept.

2.3.2 The Design of The Reflection

As stated in the previous section, reflective mediums are the output of critical design methodology, and it could also be regarded as artwork with a clear definition. Nevertheless, the message that reflective mediums aim to express is generally more durable and clearer than general artworks [24]. Moreover, The reflective design is generally regarded as the project aiming to examine the interrelationship between people and technology [25].

The format or output of reflective design could be designed in various forms, just like the various output of affirmative design could be. Generally, the reflective design might be designed as an applied art installation, an irony work, or anything that could be provocative. However, to let participants rethink the issues that reflective design has stated, the message of reflective design should be as direct as possible.

In this research, the definition of reflective design would be defined as critical reflection, it aims to trigger participants' unconscious or hardened aspect into conscious awareness or empathic understanding, in order to explore the discussion of established facts, such as complicated ethics, cultural diversity or various social issues.

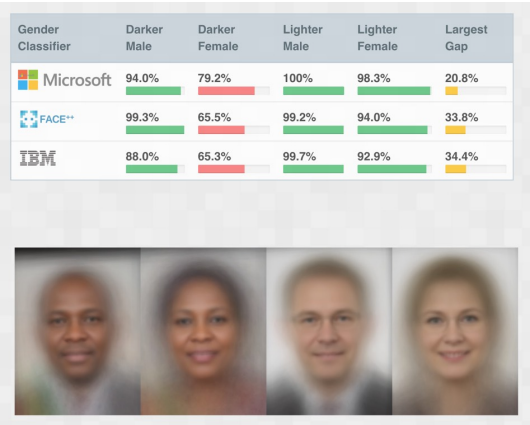
Therefore, in this research, the reflective design would be designed as a provocative art installation, which including the reflection of gender bias in the machine, to let participants not only re-think how our gender labeling is influencing the machine but influencing the innovative future that surrounded by human-centered technology.

2.3.3 Algorithmic Bias as a Reflection of Human Behavior

Inspired by critical design methodology, to reveal those values in the context of gender labeling, both by humans and by algorithms, how might we apply algorithmic bias into reflective design? How might we provoke human introspection on both gender bias and gender boundaries?

Based on the hypothesis that inspired by the status quo of both gender issues and human-machine relationships, two related works applied algorithmic bias as a reflection of human behaviors.

Gender shades [6][5] lead by Joy Buolamwini from Center for Civic Media of MIT Media Lab, is challenging how well do famous technical-related services such as IBM, Microsoft, and Face++ AI services could guess users' gender of a faces(Figure 2.8). This research has exposed that the facial recognition system nowadays, the female is hard to be recognized as a "human" than male, moreover and people within dark skin is hard to be recognized as a "human" than white-skin people. Based on this defect of facial recognition, the class consciousness behind the machine has been exposed relatively.



(Source: Gender Shades [6])

Figure 2.8 Gender Shades (2018)

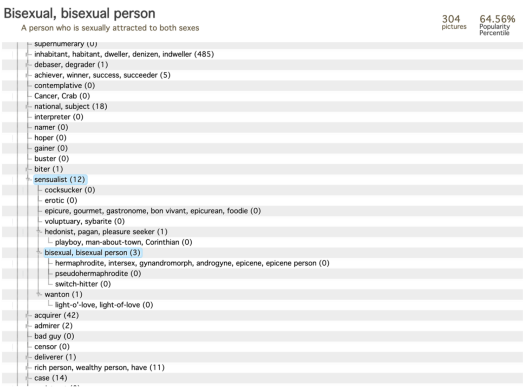
(Source: ImageNet Roulette [26])

Figure 2.9 ImageNet Roulette (2019)

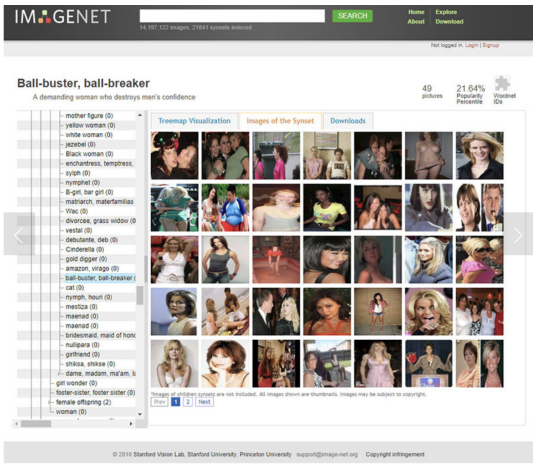
5 Gender Shades
http://gendershades.org//

On the other hand, ImageNet Roulette [26][6] is a provocation designed to help us see into the ways humans are classified in machine learning systems, co-designed by social data researchers and artists. ImageNet Roulette allowed participants to upload their facial image on the website and be classified by categories set on ImageNet, and results showed obvious human bias on how humans think other peoples' appearance(Figure 2.9).

In the beginning of this research, researchers found that there were some strange categories that arranged by ImageNet (Figure 2.10), and the reason why this categories be arranged in this way remains a mystery nowadays. If entered a specific category in detailed, people could see that some random facial images were labeled with a strange definition without reasons (Figure 2.11). Since ImageNet has been a well-known faical dataset provider, if people tried to trained these facial dataset within machine learning, the consequences could be disastrous.



(Source: ImageNet Roulette [26])



(Source: ImageNet Roulette [26])

Figure 2.10 Strange categories of facial dataset in ImageNet

Figure 2.11 Random facial images were labeled with a strange definition

After ImageNet Roulette released their experiment online, ImageNet has removed 600,000 images without any statement of why and who set up those categories in ImageNet. Moreover, what ImageNet has done did not solve the problem.

6 Excavating AI: The Politics of Images in Machine Learning Training Sets, `http://https://www.excavating.ai//`

After all, the problem is always coming from humanity. This medicine only treats the symptoms but not effect a permanent cure.

2.4. Summary

This research is stated as critical design research, which the design methodology, design process, and even the definition of design are different from those of affirmative design. To exploring and questioning the status quo accordingly, it is essential to state both theoretical basis and related works of this research at the same time simultaneously. Therefore, this chapter has included a mixed description of both theoretical basis and related works.

In order to prove that machine ethics could not only be regarded as issues need to be focusing on, but also could be transferred into the provocative works, the theoretical basis of this research to state the origin of gender bias is defined by elaborating the similarities between gender labeling and gender classification. Moreover, researches and provocative works that are questioning human-machine relationships is also mentioned.

After the elaboration and the introduction, those are stated above, the exploring discussion of whose fault when AI bias happened is stated as being a precursor of critical design methodology, which is applied as both design mindset and design methodology of this research. Followed by the detailed introduction of critical design methodology, reflective medium, and the output of critical design methodology are described in detail. On the other hand, the decision made by IBM, Amazon, and Microsoft directly hint that human beings are not ready for the facial recognition system.

To sum up, as for these kinds of human-centered technology, which would indirectly or directly influence how human beings would be objectifying as various labels, they are becoming serious issues. After all, as living in human society, gender diversity, cultural diversity, and individual diversity are included. These complicated factors are more profound than the reasons for, and methods of human beings build human-centered technology. Plus, it is already not big news that creators of these kinds of human-centered technology subconsciously instilled their unconscious bias on gender issues or racial issues. Based on this premise, could

we expect design as a provocative work, reflecting the severe status quo to human beings to trigger human beings to re-think what kinds of innovative future we are chasing for?

By creating provocative works with the combination of gender labeling and human-machine relationship, which are stated previously, “Gender shades” and “ImageNet Roulette” could be regarded as related works of this research. However, both “Gender shades” and “ImageNet Roulette” exposed bias data in specific systems, such as Amazon, IBM or ImageNet, and more on facial classification.

To explore introspective contemplation on gender bias in human-machine relationships with participants, provoking participants to re-think the human-centered technology we are chasing for, how might we apply gender bias happened of the machine into a reflective medium, by critical design methodology? How would it provoke human introspection on both gender bias and human-machine relationship issues?

As the hypothesis stated above, this research focuses more on designing the reflective medium that socialized gender classification is applied to, and exploring how it would provoke human introspection on both gender boundary and gender bias.

Chapter 3
Concept

As the critical design is applied as the core methodology, this research would focus on three parts: research concept, concepts of reflective mediums, and the process of proving the concept. Therefore, technical implementation of the reflective design or clear problem-solving would not be the main discussion of this research.

3.1. Design Concept

Referring to the literature review, the similarities between gender labeling in human-machine relationships, especially for the cognition process of both human context and algorithmic context, are generally similar. Take labeling theory as an example. In the human context, human beings saw other peoples' appearances, comparing it with the characteristics, cultural background, and social value of both males and females in the cognition process, defining the result and labeling other people as male or female. On the other hand, in algorithmic context, especially facial recognition system and gender classification system, also be designed as working in the similar process: Receiving the appearance, comparing it with feature extractor and data set, defining the result depended on the accuracy, framing people's face as a human or not, and labeling the target is male or female with an actual label.

Moreover, human could see what influence algorithmic biases have brought on racial issues and gender issues, and start to solve the issue with filtering algorithm, even though human could not precisely define both gender boundary and unconscious gender biases in our daily life yet.

Nevertheless, what if there are exploratory possibilities that biased machines

could be applied as a reflective medium, becoming more than just the inferior design needed to be erased? How might it provoke human's introspection gender bias in both algorithms and socialized gender boundaries, becoming more than just defaulting the unpleasant truth?

Based on the exploration and the hypothesis explored from the literature review, this research is applying biased machines as a reflective medium, to provoke participant's introspection on both gender bias and the meaning of gender boundary. Reflective medium would be designed as installations that allowed participants to interact with biased machines that have been socialized. (Figure 3.1).

3.2. Pilot Study

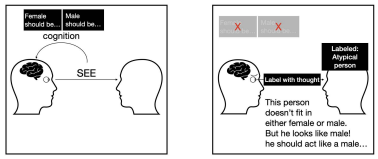
To first prove the feasibility of this research, we conducted a pilot study to design a medium to expose gender bias in the human-machine relationship in KMD Forum, 2019.

3.2.1 The Concept of Pilot Study and First Prototype

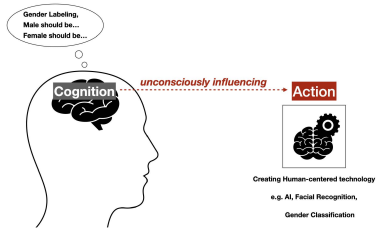
We conducted a pilot study as designing a medium for introspective reflection, that exposed gender bias in the human-machine relationship. Since there are always humans behind the technology, the results are influenced by human behaviors. Therefore, the first prototype was made for showing how human manipulation can bias classification systems in extreme consequences.

As human beings, we are always being judged and judging multifaceted, and the result is not always one-hundred percent delightful or positive. If faced up the contrary truth, what would happen?

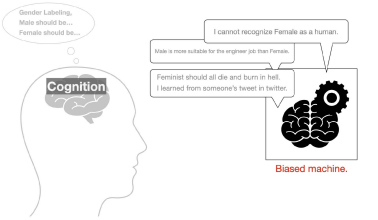
During this pilot study, the prototype was conducted to highlight the negative impact from humans by deliberately choosing the negative result, to do this as a way as sensitively discuss social stereotypes. This pilot study discussed stereotypes and how they are transmitted from human to machine through machine learning classification.



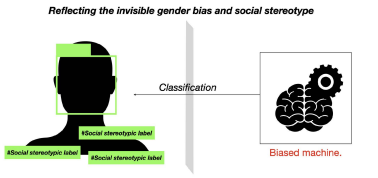
1.
Referring to **gender labeling**,
human unconsciously "label" male or female with various social stereotypic labels.
Therefore, when a human saw an atypical person, and human cannot fit this person either in male category or female category, then unconscious bias might happened.



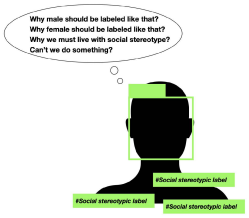
2.
On the other hand, **human being, creature that live in a complicated context,**
creates human-centered technology, such as AI, Facial Recognition, Gender Classification, in order to help human "manage", "classificate" human, for creating the "innovative life".



3.
However, cases of biased machine has been exposed in various field.
Although the reason why the machine become biased cannot be regarded as 100% from human's fault, but various cases indicated that the collection of dataset could only made by human.
And the destiny of biased machine is be banned, be filtered, or be abandoned. **Even the origin of the biased machine is from human bias.**



4.
We don't think that delete, ban or even filter the bias could solve the problem.
To face up the humanity is the only solution.
If applying biased machine as a "reflective medium", reflect social stereotypic label to human being, what will happen?



5.
How might this reflective medium provoke human's introspection gender bias in both algorithms and socialized gender boundary, becoming more than just defaulting the unpleasant truth?

Figure 3.1 Concept sketch

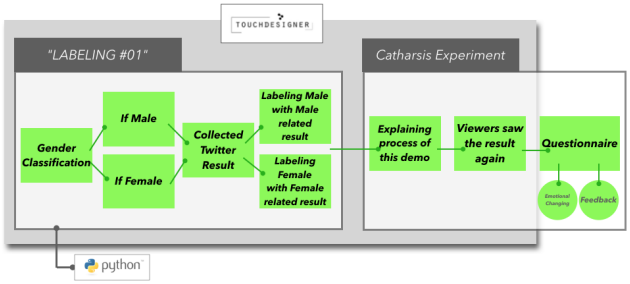


Figure 3.2 The Structure of First Prototype

3.2.2 The Design of First Prototype

During the pilot study, participants were being labeled by the pilot prototype that is biased gender classification, in which the system was developed using OpenCV and Touchdesigner(Figure 3.2).

Following the idea of letting participants be labeled by biased gender classification, participants' valuable feedback showed the feasibility of our concept triggering introspection on gender labeling by experiencing biased gender classification during the pilot study(Figure 3.4).

3.2.3 The Experience Design of First Prototype

As the first prototype of a reflective medium, we considered designing a provocative but straightforward experience as the first direction(Figure 3.3). The process of interaction is detailed below.

- First, when the participant stands in front of the first prototype, it will automatically change into ready to take picture mode.
- Second, participant has their gender labeled by the first prototype.
- Third, the output result of gender description is collected from the tweeter, an appropriate ego network for experimental analysis on ethics issues [27], that are filtered to display only the ones that are hostile or aggressive towards that gender.

- Forth, after being classified, a statement of our concept is automatically shown on the screen. All participants are asked to respond to the emotional status in the questionnaire after reading the statement.

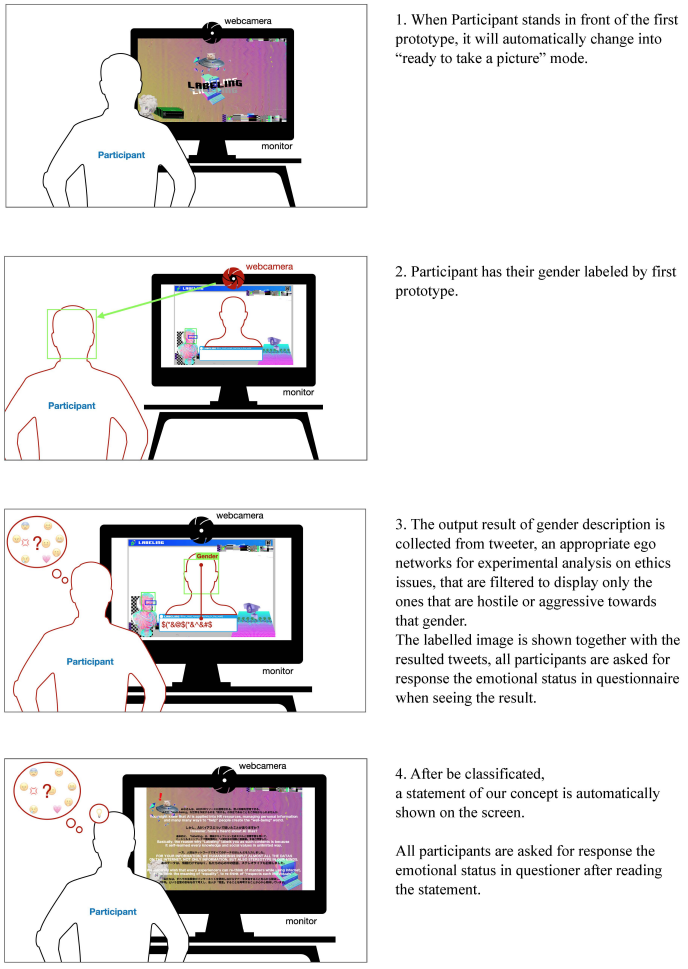


Figure 3.3 Scenario sketch of first prototype

Figure 3.4 Participant experienced the first prototype

3.2.4 The Design of Experiments During Pilot Study

Following the idea of letting participants be labeled by biased gender classification, we applied catharsis theory into the experiment.

Catharsis theory is a concept in psychoanalytic theory where the emotions are associated with traumatic events [28]. A catharsis is an emotional release that could trigger participants' empathic understanding of an issue by experiencing the traumatic works. Take *Macbeth* as an example; readers of *Macbeth* usually feel sad about the tragic central figure of the story, because his destructive preoccupation with ambition blinded the character. Moreover, a catharsis could even bring a positive impact to readers, such as learning to cherish what they have.

Since the experience that biased gender classification offers to participants might be an unpleasant experience to participants, even though it would be regarded as a traumatic experience, Catharsis theory was applied as a method to validate people's empathy and introspection. After letting participants experience our pilot study, we hope participants could trigger an empathic understanding of these issues, be aware of the seriousness of how human's gender labeling could influence

not only individuals but also the machine. To sum up, the purpose of the pilot study was not to be mean to participants. It should be an emotional charging experience to let participants face up the unpleasant truth of gender labeling.

The experience also involved an analysis of the emotional charge caused by the experience, ranging from negative feelings to relief, as the first validation. In order to observe the emotional charging, the questionnaire was conducted[1].

By applying catharsis theory in this experiment, we stated the success result as the participant's emotional state is triggered from relatively negative to feel relief feeling after receiving the message from the first prototype.

3.3. Experiment Results and Feedback

3.3.1 Experiment Results

Through two questionnaires[2] that were applied to participants: one before and one after explaining the purpose of the study, we found that emotional statements of participants was changed from negative feeling (Figure 3.5) to relief (Figure 3.6).

3.3.2 Feedback

Regarding the feedback(Table 3.1), participants received the message by experiencing the first prototype, and be reflected by the context, which also proved the feasibility of the hypothesis. On the other hand, we also received some valuable improvement points from participants, mainly related to expression technique.

To reflect a stronger message and context than the first prototype, we considered the next step as improving the first prototype and letting the prototype become a more immersive experience.

1 The detail of questionnaire during pilot study
`https://docs.google.com/forms/d/e/1FAIpQLSfZvC3q1gqAulXOAQyetHQLAIfTk9FbZ9cSpb_hfxnXxm8lzg/viewform`

2 Responses of questionnaire during pilot study
`https://docs.google.com/forms/d/1lV-q4wbfC1MLSmC2chnFiLspN0vERHB_UfxH3d8_wVo/edit#responses`

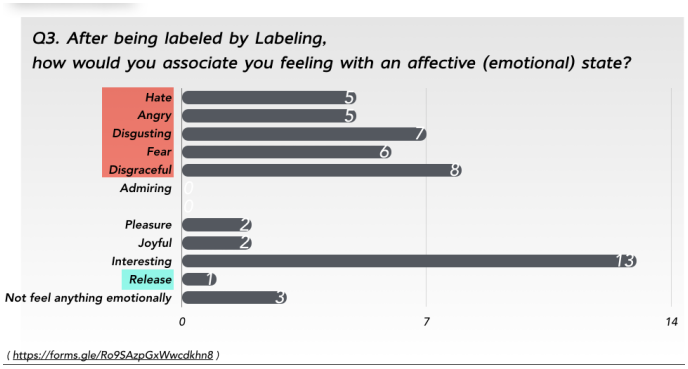


Figure 3.5 Participants' emotional statements after being labeled

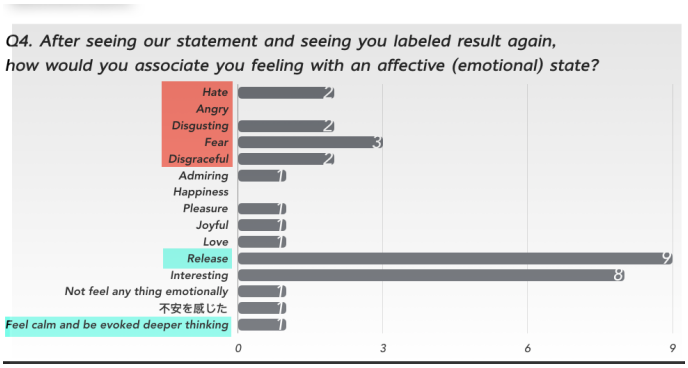


Figure 3.6 Participants' emotional statements after receiving the message

3.3.3 Insight

During the pilot study, an improving point is considered the abstract expression of the next reflective medium.

Generally, it is still hard to precisely apply a gender classification database that has high accuracy. During the pilot study, participants, especially Asian female participants, were labeled as male by the gender classification of the first prototype. Regardless of the accuracy database problem, some participants thought the incorrect classified results might be regarded as biased results.

However, the accuracy problem of gender classification also inspired our insight into this research. Not only the case of what happened to Asian female participants during our pilot study, living in the context of cultural diversity such as transgender and gender identity issues are included, how could gender classification precisely classify human being's gender based on appearance? In general, if thinking from a different perspective, since the content of the database are facial images that have been regarded as male or female by human beings, based on this subconscious social desirability, what appearance should a female or male suppose to be? Isn't that mean the majority has decided the specific gender-stereotypic physical appearance?

In this premise, how could we respect cultural diversity and gender diversity with the gender classification system, since we live in the context of cultural diversity such as transgender and gender identity issues are also included?

Based on the exploring discussion inspired by this improving point, this questioning would be applied to the next prototype to reflect how majorities thought on gender-stereotypic physical appearance.

3.4. Summary

During the pilot study, we explored the feasibility of the hypothesis, and a reflective medium did reflect a thinking context to participants by experiencing the first prototype. Moreover, a catharsis experiment was conducted as an evaluation to prove the concept.

Nevertheless, the design of the first prototype needed to be considered as an essential improving point. From the perspective of fairness, the first prototype was

made by manually filtering tweets from Tweeter. Although the hypothesis's feasibility has been proved, this technical implementation would cause severe problems on fairness and credibility.

On the other hand, to express the message more directly and comprehensively, the intensity of expression should be more immersive and discuss human-machine relationships and gender bias from more angles. After all, there are many applications in the field of human-centered technology, and gender bias could be regarded as a vast and complicated context at the same time.

Therefore, the next step of Labeling is to improve points from both technical and conceptual nature. The next improving plan of this research is to augment the immersion by designing a virtual worldview, which consisting of three installations, discuss gender labeling and gender boundary from a broader perspective.

Table 3.1 Feedback from participants in pilot study

Participants	Feedback
A	A little bit feared after knowing the fact.
B	this let me think of a philosophical problem: Before people discussing about innovation and the future, have they thought about the origin of problems? Maybe the negativity of people and the society are the first thing we need to face.
D	I think it is an interesting way to observe and critique the vast amount of negative and/or strange criticisms found on the internet.
E	I could not feel emotional "charge" of the result since it's a mere algorithm outputting stuff. It may be related to me having a degree in CS.
F	Very impressive. We should think more about the meaning of respect each other. Because this is the first time I knew that human stereotype can influence future technology. Maybe the problem is difficult to solve, because the reason is we human and the society. its very dangerous if we don't think deeper.
G	It might be more impressive if this concept can be displayed like a contemporary art installation.
H	How about show the interaction in real time?

Chapter 4
"Labeling"

After the pilot study, defining the limits of this research, exploring the possibility of improving the reflective design, and the design of the first prototype was determined the feasibility of the hypothesis, "Labeling" is updated to be as a more immersive experience, which consists of three installations that discussed both gender labeling and gender boundary from different angles: the reflection of gender-stereotypic physical appearance, gender stereotype in social networks, and human-machine symbiosis. Plus, to prove the concept more persuasively, the method applied to observe the first hand result is indispensable. Therefore, during the "Labeling" demonstration, participants were observed unwittingly since a human monitoring experiment is conducted secretly.

4.1. The Critical Design of Gender Bias in Human-Machine Relationship

"Labeling" consists of three installations as a reflective medium, which are: "Gender Shell", "Uncover Whispering" and "(Statement) in Processing" (Figure 4.1)[1]. Each of them discusses the gender-stereotypic physical appearance from the angle of gender classification, exposing human verbal behavior when discussing gender issues on the social network, and questioning how human-machine symbiosis could achieve empathic understanding, by challenging vocal emotion recognition system from different angles.

By discussing, exposing, and questioning three topics related to gender bias

1 The demo video of Labeling
https://www.youtube.com/watch?v=iPq1vsghiAo

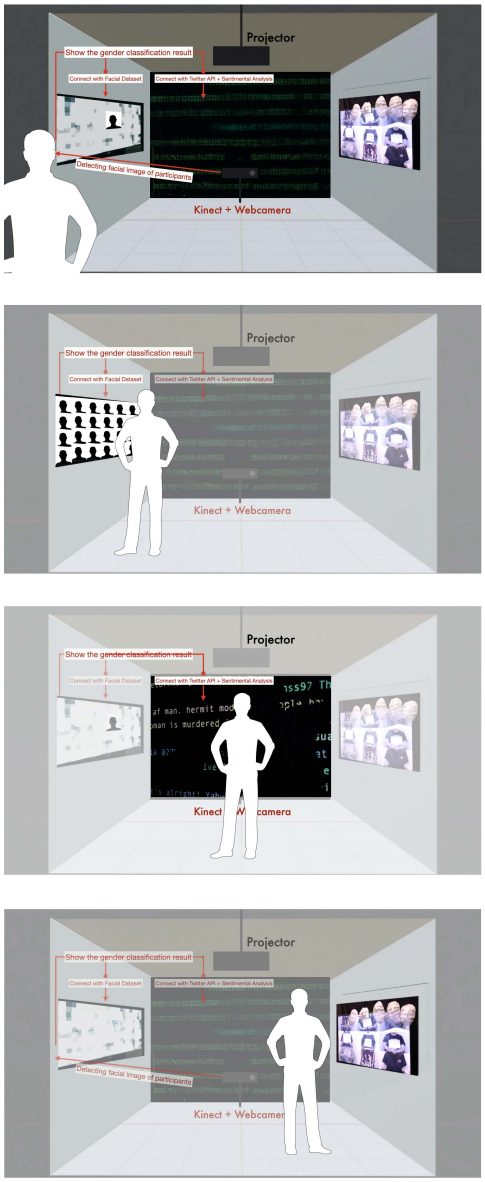
and human-machine relationships, this work aims to explore introspective contemplation on gender bias with participants and question the existential meaning of gender boundary.

Figure 4.1 The representative image of Labeling

4.2. Experience Design of “Labeling”

As stated previously, to reflect both gender bias and gender boundary in the human-machine relationship, “Labeling” is consists of three works to provoke participants’ to re-think how sincere human being could bringing their gender bias into human-centered technology. The experience flow of “Labeling”, which is experiencing three installations in an actual space coherently, is detailed below(Figure 4.2).

- Firstly, when the participant entered the room, the participants’ facial image will be automatically detected, and inputting into Work 01: “Gender Shell” and Work 02: “Uncover Whispering”. When the whole experience started



1.
When the participant entered the room,
the facial image of partipants will be automatically detected, and inputting into Work 01: Gender Shell and Work 02: Uncover Whispering.
When the whole experience started to output classificated result at the same time, the participant would be brought into the context of Labeling by this immersive experience.

2.
When viewing Work 01: Gender Shell, participant's gender will be classified by an unsupervised machine learning algorithm,
that shows the participant's picture together with a mosaic of faces it thinks it's similar,
showing how the machine would divide these faces into groups without using human made labels of faces from a dataset.

3.
work 02: Uncovering whispering is shown as projection mapping on the wall. the detected gender is used to look for tweets related to that gender, that we called whispers. Participant's body outline would be filled with real-time tweets that are discussing their own perspective on participant's gender, and all tweets are collected by twitter API in the real time.

4.
After experiencing work 01 and work 02, work 03: (Statement) in Processing is a close-loop installation. By viewing how vocal emotion recognition system analyzed human being's voice when they expressing their own experience on gender issues, participants could be triggered an introspection on the meaning of human-technology symbiosis.

Figure 4.2 The whole experience flow of labeling

to output classified results simultaneously, the participant would be brought into the context of “Labeling”.

- Secondly, when viewing Work 01 “Gender Shell”, an unsupervised machine learning algorithm will classify participant’s gender, shows the participant’s picture together with a mosaic of similar faces in the background. This work shows how the machine would divide these faces into groups without using human-made labels of faces from a dataset.

- Third, work 02: “Uncover whispering” is expressed as projection mapping on the wall. The detected gender uses to look for tweets related to that classified gender. The participant’s body outline is filled with real-time tweets, that discussing their perspective on the participant’s gender, and all tweets are collected by twitter API in real-time.

- Forth, after experiencing work 01 and work 02, work 03: “(Statement) in Processing” is a closed-loop installation. By viewing how vocal emotion recognition systems analyzed human being’s voice when expressing their own experience on gender issues, participants could be triggered an introspection on human-machine symbiosis, re-thinking the human-machine relationship nowadays.

4.3. Work 01: Gender Shell

4.3.1 Concept

Work 01: “Gender Shell” (Figure 4.3) is inspired by the theory of gender-stereotypic physical appearance [29], which described that human being has an unconscious standard when classifying a person’s biological sex, such as female is generally labeled with a delicate appearance, and males’ is with an overwhelming strength appearance, this gender-stereotypic physical appearance is not only influencing the growth of gender stereotype, also indirectly effecting gender classification, as the pre-collection of dataset still rely on human manipulation. Once this human manipulation is transferred into a psychological manipulation, it might not only

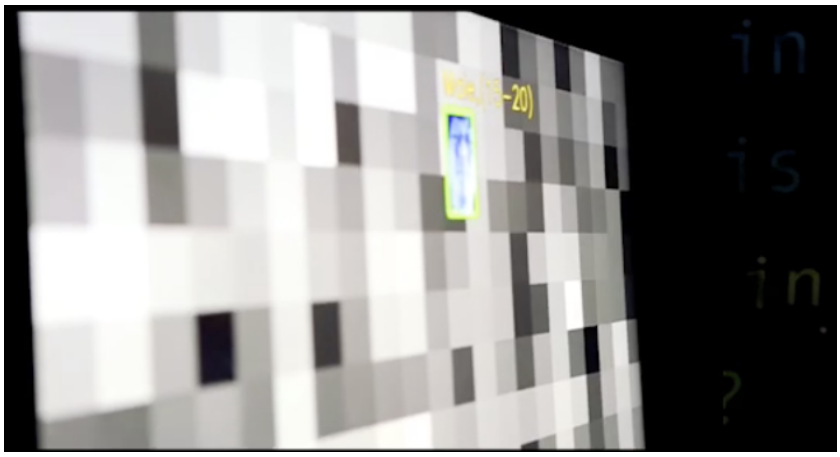


Figure 4.3 Work 01: Gender Shell

influence the classification result of human beings but deepen the gender boundary gap. Therefore, through this work, we would like to reflect the discussion below with participants:

Technology comes from humanity; what it builds is a reflection of the majority view. In this premise, what “body shell” should we be? What will the majority of databases think about our natural appearance? When facing the reflection of social expectation, Which “gender shell” would our biological sex be linked?

4.3.2 Technical Implementation

When the participant entered the exhibition space, the facial image of the participants automatically took by a web camera, and input into the unsupervised machine learning algorithm, which is “Gender Shell”.

“Gender Shell” would compare participant’s facial image with its dataset, showing both gender and similar facial images results, which it thoughts could label the participant.

4.3.3 Experience Flow

Based on this concept, the participant’s gender would be classified by an unsupervised machine learning algorithm that shows the participant’s picture and a

mosaic of similar faces.

By the experience flow of “Gender Shell”, this works to show how the machine would divide these faces into groups, without using human-made labels of faces from a dataset(Figure 4.4).

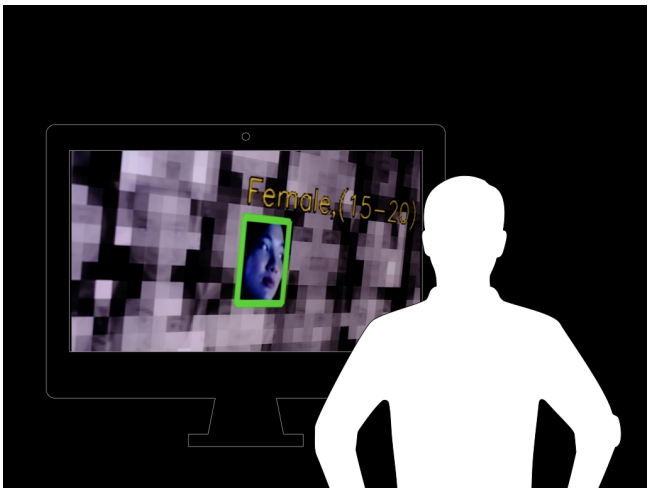


Figure 4.4 Work 01: The experience Flow of Gender Shell

4.4. Work 02: Uncover whispering

4.4.1 Concept

Work 02: “Uncover whispering”(Figure 4.5) could be regarded as an evolved version of the first prototype. It shows humans’ thoughts about gender on twitter by collecting manually during the design of the first prototype. Nevertheless, to avoid unnecessary disputes, in this term, every tweet is collected by applying twitter API, in which the system structure is fairer than the first prototype when designing this work. Moreover, to enhance the immersive experience, we designed a real-time interaction with real-time collecting twitter online, and the expression is conducted to be a real-time projection mapping.

Through the design of work 02: “Uncover whispering”, we would like to reflect the discussion below with participants: The social network is a collection of human

behavior; what is exposed is a reflection of real feelings. Every word people have whispered in social networks will be recorded as data, staying between virtuality and reality. If our natural appearances transferred into the social network, how would residents on social networks discuss them? What whispers should be linked to our “Gender Shell”?

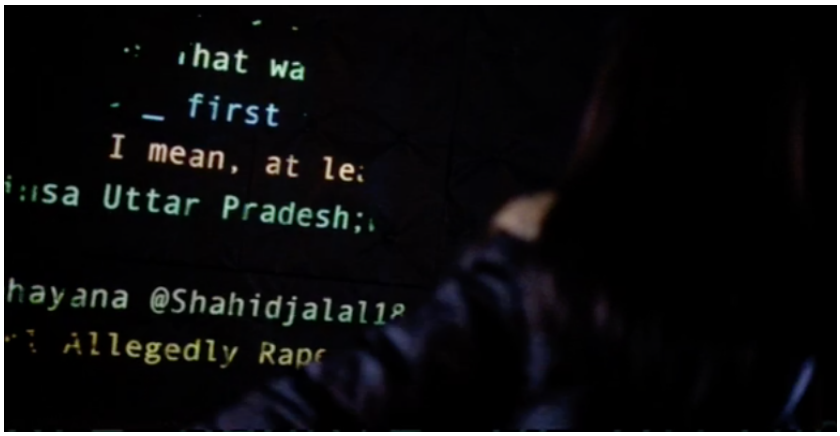


Figure 4.5 Work 02: Uncover Whispering

4.4.2 Technical Implementation

Based on the message of this work above, the detected gender is used to look for tweets related to that gender, that we called whispers. The filtering was done only to remove web addresses and retweets for this application of twitter API.

After analyzing the sentiment of whispers that displayed in projection colored, that comes from a gradient from -1 (negative sentiment) in red, 0 (neutral) in green, and 1 (positive) in blue, the participant can interact with the text using his body silhouette, captured by a Kinect sensor. Plus, the coloring gives a psychological effect of positive or negative feelings based on tweets.

Keywords used for searching the tweets were: “male is”, “man is”, “boy is”, “male”, “man”, “boy” if the participant is labeled as “male“. And “female is”, “woman is”, “girl is”, “female”, “woman”, “girl” if the person is labelled as female.

4.4.3 Experience Flow

The experience flow of “Uncover whispering” is shown below(Figure 4.6).

When the participant entered the room and be classified by “Gender Shell”, the classified result would also be sent to “Under Whispering”. The participants would see their body filled with tweets related to the classified gender, to glance at what and how people hidden behind the interest discussed based on gender.

In order to show the positive tweets and negative tweets, sentimental analysis is applied to this work with color expression. Every tweet would be shown as a color gradient from blue to red, which means if the tweet is regarded as a more positive content, its color would be close to blue; if contrary, it would be close to red. Plus, the projection mapping would show in two ways: filling tweets in or out of the participants’ body outline.

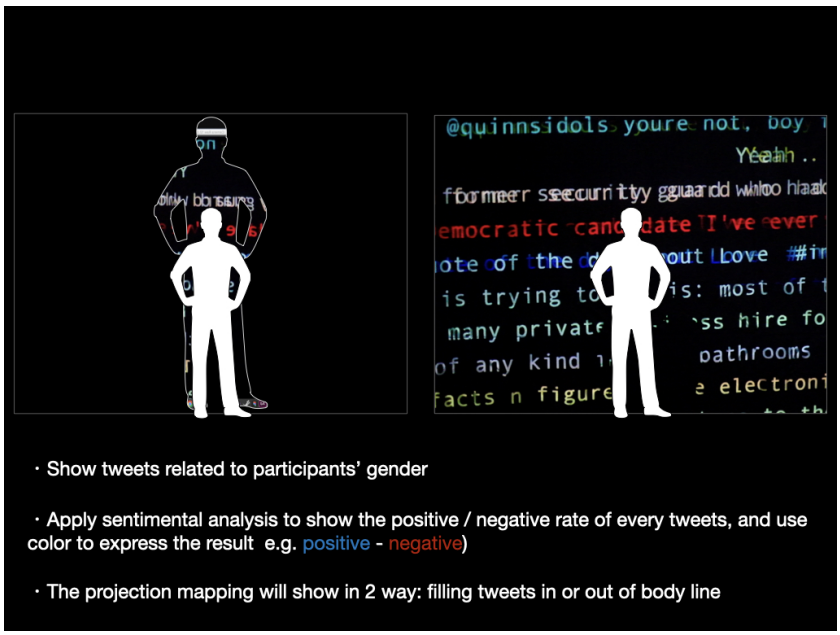


Figure 4.6 Work 02: The experience Flow of Uncover Whispering

4.5. Work 03: (Statement) in Processing...

4.5.1 Concept

“(Statement) in Processing”(Figure 4.7) is an installation that questioning the boundary of human-machine symbiosis.

Human society is not only a starting point of technology but a destination. No matter how advanced technology will come to human society, people are still receiving, sharing, and experiencing each other's perception with vivid souls. People are individually diversified and living with complicated thoughts and feelings.

In this premise, could we expect an empathic understanding of human-machine symbiosis?

Figure 4.7 Work 03: (Statement) in Processing...

4.5.2 Technical Implementation

Based on the questioning of human-machine symbiosis above, we conducted interviews with people from Taiwan, China, Indonesia, Colombia, and Greece, where they shared their personal experience and perspective on facing gender issues. Although their cultural background is different, their statements were surpris-

ingly similar, all of the people interviewed thank that this world has much room for improvement on gender equality, and expressed this with sadness and severe emotion.

Their statements were inputted into Empath [30], a well-known AI application that classifies people's emotions based on sound, and the result controlled LED lights placed on top of the screens, which represent five emotions: Anger, Sorrow, Energy, Joy, Calm. For the actual expression, each analyzed emotion will emerge as color expression below:

- Anger: Show color Red with angry facial mask
- Sorrow: Show color blue with sorrow facial mask
- Energy: Show color green with energetic facial mask
- Joy: Show color yellow with joyful facial mask
- Calm: Show color white with expressionless facial mask

By presenting results analyzed by the machine, this work would like to question: As a human being with a vivid soul and human complexity, what do we expect from human-centered technology? What kinds of human-machine symbiosis are we chasing for?

4.5.3 Experience Flow

The experience design is shown as below(Figure 4.8). The participant could hear interviewers' sound in the space. The machine also could be inputted the sound. However, as a human being, the participant could understand the statements and feel their emotional status when they talked about their statements. When hearing statements, the participants could also observe analyzed results from the machine, which was expressed by an LED color with emotional facial masks installed on the monitor. By comparing how interviewer's emotional status is the participant received and how results of emotional status the machine has analyzed, the participant might receive the sense of conflict and sense of irony, re-thinking about the baseline of human-machine symbiosis, and what kinds of innovative future are human being would like to live with.

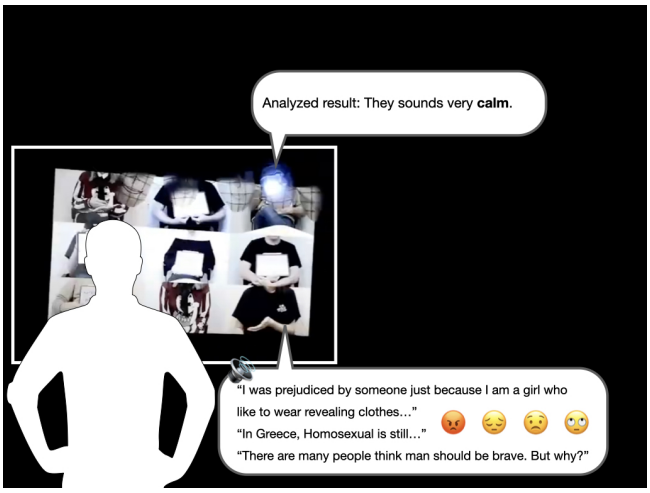


Figure 4.8 Experience design of (Statement) in Processing.

4.6. The Design of Experiment

4.6.1 Human-monitoring: Dummy Memo Test

During this term, the situation-producing theory [31] and participant observation [32] are applied as observation methods into the experiment.

Before participants experiencing the second reflective design, we set an Ipad with a previously written negative comment collected from Twitter (Figure4.9), which the participants know nothing about where it came.

After the experience, they could leave comments on Ipad freely. Through this experiment, we would like to observe whether participants would reply the negative comment or not. The comments left on the Ipad were analyzed, and the majority of participants replied to the comment already present. Furthermore, the comments showed a willingness to participate in gender labeling and rethink their stereotype on the subject(Figure4.10).

During this experiment, to keep the fairness and the credibility of the Human-monitoring experiment, two principles are highlighted below:

- The experience is without any verbal description. Participants could freely be affected by “Labeling”.

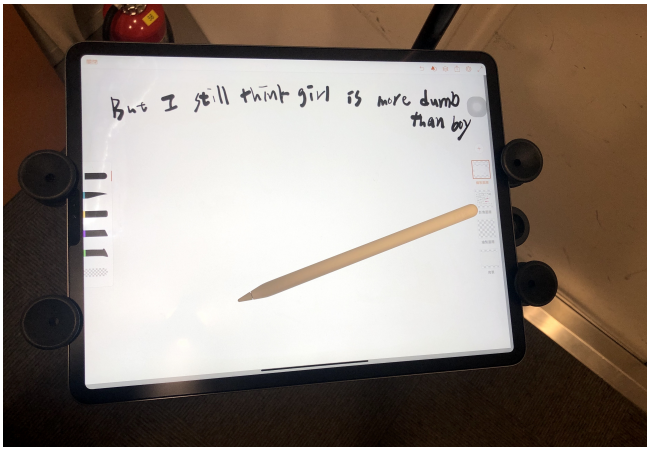


Figure 4.9 An Ipad with a previously written negative comment collected from Twitter.

Figure 4.10 The concept design of dummy memo test.

- Although the dummy memo test is secretly in progress, participants were all free to fill their feedback without any instruction.

4.6.2 Demonstration

“Labeling” was exhibited in Media Studio two times, which was the first time on March 2, and the second time was on March 9, 7 participants joined the first term on March 2, 5 participants joined the second term on March 9.

In order to objectively observe how participants' could be provoked by experiencing “Labeling” more objectively, all participants freely interacted with three installations of “Labeling” without any verbal description. As stated previously, the dummy memo test was conducted at both terms of the demonstration. The negative content we set up in two terms of the dummy memo test is detailed below (Figure4.11)(Figure4.12).

Moreover, before participants leave the exhibited space, an video interview to let participants freely give oral feedback to “Labeling” was also conducted.

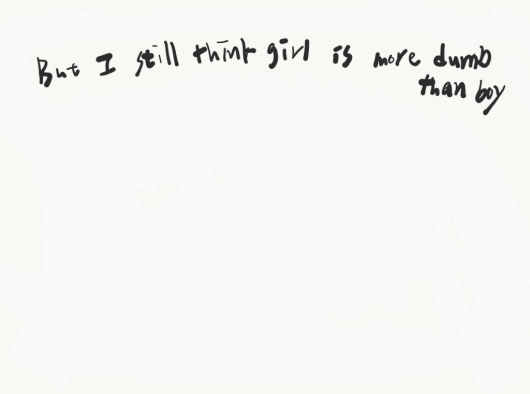


Figure 4.11 The negative comment in first term of dummy memo test

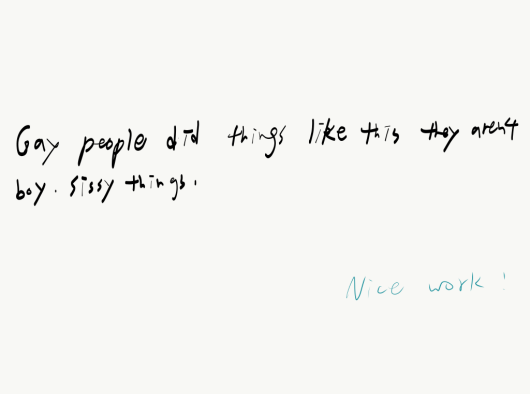


Figure 4.12 The negative comment in second term of dummy memo test

The dummy memo test was set up in front of Media Studio; participants were free to fill their feedback without any instruction (Figure4.13).

Figure 4.13 Participants left their feedback initiatively.

4.7. Proof of Concept

4.7.1 The Result of Human-monitoring Test

After the first term of exhibiting “Labeling” and dummy memo tests, the feedback broad had been left comments by many participants automatically. Surprisingly, some participants replied to the negative comment in the dummy memo test without any instruction(Figure4.14).

Comments left by participants of first term is detailed below.

- WHAT?
- There are tons of girls smarter than you out there.

Moreover, the results in the second term are worth mentioning. During the second term, 5 participants lived their comments on the Ipad, and four comments could be regarded as the reply to the negative comment (Figure4.15). Comments left by participants of the second term is detailed below.

- DO NOT LABEL PEOPLE
- mutual respect plz
- After watching this and still say something biased like this?
- Maybe u can keep this thought in your mind, BUT, DON’T TRANSFER IT INTO WORDS AND ACTIONS

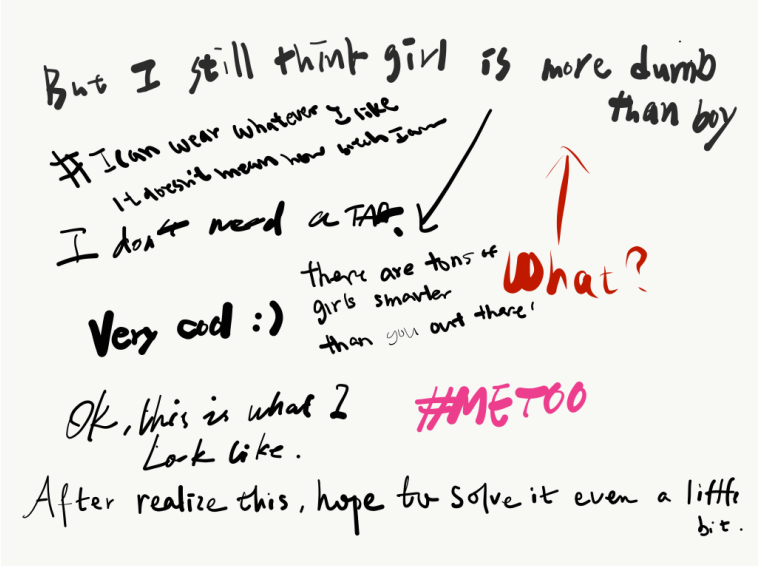


Figure 4.14 Participants in the first term commented the negative comment without any instruction.

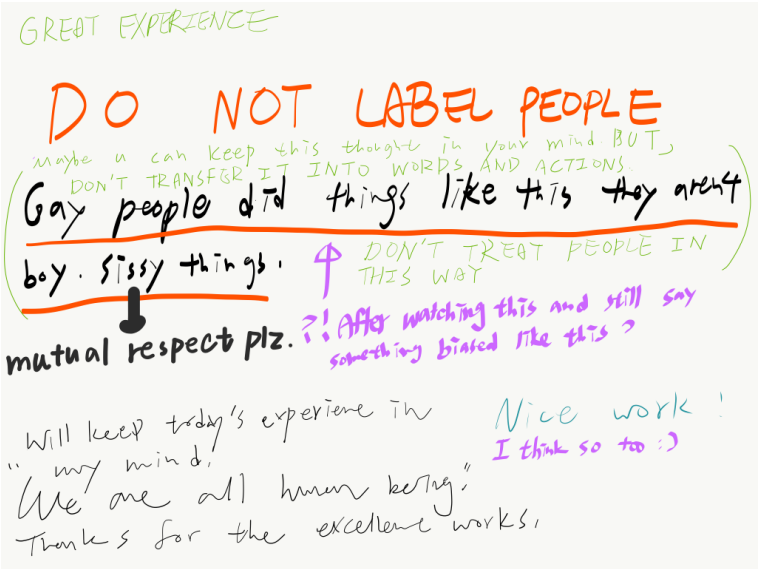


Figure 4.15 Participants in the second term commented the negative comment without any instruction.

4.7.2 Oral Feedback of video interview

On the other hand, a video interview was be conducted before participants leave the exhibited space[2]. Five participants experienced “Labeling” and willing to give oral feedback. The oral feedback is detailed below.

- Participant A: I think this work is very meaningful and very interesting, because Labeling right now is happening everywhere, especially for women. Women right now living in this world is gradually become equal, even in some part still have something to work on. For me, especially living in Asia, as a woman, still have to be act in a certain way. Especially for me, whatever I want to wear, will reflecting on how people look at me. If I wear revealing clothes not like a typical woman, they will think that I am too bitchy or something. I think whatever you look like or whatever you do, doesn't effect anyone to look at you, you just do whatever you want. Because you are being yourself, every individual is independent. So don't label people based on what they look like, don't judge people from what they look like. This is what I thought after I experienced this work.

- Participant B: I think this project is very good for raise people's awareness about gender equality. Also now, everyone is discriminating each other, even though you are conscious about gender discrimination, but we still like discriminating others unconsciously in some way. But after experiencing this project, I think it can make us be aware of what we say and what we label each other every day, and how we should act to other people. So it is a really good project to help raising the awareness.

- Participant C: This immersive experience is different from VR or any immersive experience we generally think about. This work offered me a oppressive feeling, because I suddenly receive how the machine classified me as a very simple label, like I am just a micro data without any value, and even in Work 02, my body is filled with people's tweets based on that gender label. When experiencing, I suddenly understand the reason why this work called

2 The oral feedback of video interview in “Labeling”
`https://drive.google.com/drive/u/0/folders/1plqqzK259jCUgAk0yoDHuoE8kd7vDrB2/`

“Labeling”, because for gender classification, it label people with a gender label; for human, human label each other with stereotype, such as women should do something girly, man should be brave. This work really let me think about the definition of innovation, and what kinds of future we are desiring for. For me, it is not only discuss gender issues, it also discuss how we influence machine. This work is amazing. The immersive experience of this work is much more stronger than any other technology thing, and the story-telling of this work is very outstanding. I really admire this work.

- Participant D: I really like the overall atmosphere of this exhibition. And the first one are the most impressive one for me, I begin to realize that how bias is around us everyday in our life. For the second one, the way this installation shows tweets is a little bit hard to me to read the information. if the subtitle was from left to right, will be much more easier for me to read.

4.7.3 The Discussion During ACM DIS2020 Student Design Competition

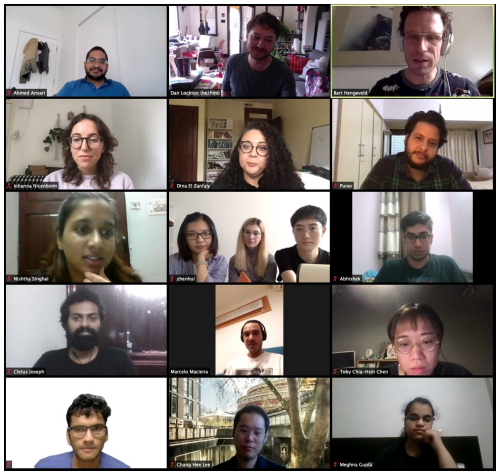
Besides the demonstration, “Labeling” had the honor of been accepted by the ACM DIS2020 Student Design Competition. Despite the conference could only be held virtually, this research had the chance to discuss the insight with the jury on July 8th (Figure4.16).

“Labeling” received the recognition by the winning entry for the inaugural student design competition. The jury mentioned that they particularly appreciated this research is positioned to garner critical reflection, with the high maturity of the theoretical framework, design process, and goals(Figure4.17).

Since the theme for DIS 2020 is “More than Human-Centred Design”, the conference expected to rethink the research and contributions humans make in design and HCI [33]. During the discussion with the jury, the valuable discussion about the ideal future of the human-machine relationship was conducted.

When conducting the previous demonstration held in Japan, “Labeling” has received hesitation from participants, and an explanation of both theoretical framework and critical design are needed. In the contrary, “Labeling” was more easily

be understood by the jury and other participants of the Student Design Competition. The discussion was explored deeper than the demonstration; we discussed the ideal future of the human-machine relationship, such as how a gender classification could respect a transgender person, the necessity of human-centered technology, and how careful a HCI researcher should be aware of chasing the innovation.



(Source: Prompt report by SDC Chair, Dan Lockton [34])

Figure 4.16 DIS2020 Student Design Competition was held online

(Source: Prompt report by SDC Chair, Dan Lockton [34])

Figure 4.17 The Prompt report of Student Design Competition

4.7.4 Insight

During two terms of demonstration, the present stage of “Labeling”, three installations that are reflecting gender bias in the human-machine relationship, had been exhibited in an actual space. In order to keep both the fairness and credibility of this Human-monitoring experiment, the demonstration was all without previous explanation and instruction. Without any verbal description, participants could freely be affected by “Labeling”.

The result of the dummy memo test is totally beyond expectation. In the beginning, this experiment was designed every carefully with a theoretical basis because this kind of human-monitoring test is more natural to be offensive to

participants. On the other hand, since the dummy memo test's design was low level, the effectiveness before presenting it as our experiment was worried. However, during two terms of the dummy memo test, a discussion had been explored without any instruction, and participants focused on fighting back to negative comments separately but corporately.

On the other hand, when collecting oral feedback by the video interview, what is worth mentioning is that participants could generally receive the statement we would like to express by experiencing “Labeling” and gave feedback that indicated their empathic understanding of both gender bias and human-machine relationship. The reflection that “Labeling” has brought to participants was not only gender bias but the boundary between human and machine.

After these two terms of demonstration, The method to validate and evaluate the effects of the experience on the participants is regarded as the future plan of this research.

Take the catharsis theory that was applied in the pilot study as an example, to measure the emotion charge of users with sensing technology, the sensing data such as Electrodermal Activity (EDA) and other biological signals [35], can provide further information to allow a better evaluation. Therefore, if catharsis theory could be proved with sensing technology, there might be some exciting found or even a validation to be explored.

On the other hand, considering the cultural background of participants might be an improvement for the next step. Since the subject of “Labeling” is gender issues, this subject highly depends on the context, different countries and cultural backgrounds would influence the received feedback. Nevertheless, the present stage of this research didn't precisely including the discussion based on cultural diversity.

However, since Coronavirus Disease 2019 (COVID-19) [36] has been influencing the whole world, the declaration of a State of Emergency had also been issued in Kanto Area due to COVID-19[3]. On the other hand, “Labeling” could only be exhibited in an actual space to provide the experience to participants. To keep the

3 [COVID-19] Declaration of a State of Emergency in response to the Novel Coronavirus Disease (April 16) April 16, 2020
https://japan.kantei.go.jp/ongoingtopics/_00020.html

social distancing has became the first priority. Therefore, not only the following studies and future plan, but for the next demonstration of “Labeling”, are hard to be conducted.

Although the follow studies and the future plan are hard to be conducted, through this critical design experience, the insight that reflects to participants is beyond the statement itself. Plus, it aims to let participants keep this reflection in their mind, to re-think of what kinds of social context and innovative future we are chasing for. Participants in the dummy memo test have explored a spontaneous discussion, the oral feedback by video interviews has also proved that “Labeling” has already provoked a deep reflection of participants.

Chapter 5
Conclusion: A Critical Reflection

Algorithmic bias has been a well-known issue in related research fields, and solutions aiming to erase, filter, or ban bias are presented. Nevertheless, as human beings created algorithmic machine and human-centered technology, if algorithmic bias happened, the origin of this problem is human bias from a thing called humanity. In this premise, since present solutions of algorithmic bias all conducted the fault to an algorithm, how could human beings, the origin of this problem, still keep silent without re-thinking this present state?

By presenting "Labeling", we applied algorithmic bias as a key to provoking human's introspective contemplation on the subject, discussed the ethical issues involved. "Labeling" is similar to a mirror. It aims to reflect gender bias in human-machine relationship to participants, exploring the broader discussion with participants by presenting this experience.

Besides, we explored both gender labeling and gender boundary from a critical design perspective to not only implementing it as the concept of reflective medium, but revealing hidden values on gender issues. This research described the design process of different iterations of this medium by the proof of concept and results of each phase.

To sum up, the exploration of gender bias is an example. The part which this research tried to challenge is disrupting established facts through critical design. By applying negativity into creating the reflective experience, this research proposed the works of nature could be implemented as a way to explore the discussion around the issue, more than being framed by existed definition. In this research, algorithmic bias were originally regarded as an "error", however, this "error" create a critical reflection to participants. After all, to face up the origin of the problem is the key to the correct direction.

Similar as the fight of social issues, there is no clear goal of this critical design

research. Nevertheless, once human being is provoked by an experience, start discussing unpleasant truth that might be forgotten, it could be regarded as a starting point, which could let us carefully think about how an ideal innovation should be, or how we should respect each individual's diversity instead of labeling each other.

References

[1] Beverly I Fagot, Mary D Leinbach, and Cherie O'boyle. Gender labeling, gender stereotyping, and parenting behaviors. *Developmental Psychology*, 28(2):225, 1992.

[2] Susanne Bødker. When second wave hci meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, pages 1–8, 2006. URL: `https://doi.org/10.1145/1182475.1182476`, `doi:10.1145/1182475.1182476`.

[3] Gopinaath Kannabiran, Jeffrey Bardzell, and Shaowen Bardzell. How hci talks about sexuality: discursive strategies, blind spots, and opportunities for future research. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 695–704, 2011. URL: `https://doi.org/10.1145/1978942.1979043`, `doi:10.1145/1978942.1979043`.

[4] Ayanna Howard and Jason Borenstein. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

[5] Simon Ings. Impaired visions. *New Scientist*, 244(3251):30–31, 2019.

[6] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018.

[7] Emily Pronin, Daniel Y Lin, and Lee Ross. The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, 28(3):369–381, 2002.

[8] Jeffrey Bardzell and Shaowen Bardzell. What is" critical" about critical design? In *Proceedings of the SIGCHI conference on human factors in com-*

puting systems, pages 3297–3306, 2013. URL: https://doi.org/10.1145/2470654.2466451, doi:10.1145/2470654.2466451.

[9] Terrell A Hayes. Stigmatizing indebtedness: Implications for labeling theory. *Symbolic Interaction*, 23(1):29–46, 2000.

[10] Marsha Weinraub, Lynda Pritchard Clemens, Alan Sockloff, Teresa Ethridge, Edward Gracely, and Barbara Myers. The development of sex role stereotypes in the third year: Relationships to gender labeling, gender identity, sex-types toy preference, and family characteristics. *Child development*, pages 1493–1503, 1984.

[11] Hari Krishna Behera, Sunil Kanta Behera, and Indira Behera. Gender bias in indian news media: A study of indian women.

[12] Andreas Kaplan and Michael Haenlein. Siri, siri, in my hand: Who's the fairest in the land? on the interpretations, illustrations, and implications of artificial intelligence. *Business Horizons*, 62(1):15–25, 2019.

[13] David Andre, Forrest H Bennett III, and John R Koza. Discovery by genetic programming of a cellular automata rule that is better than any known rule for the majority classification problem. *Genetic programming*, 96:3–11, 1996.

[14] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[15] Mohit Kumar. Microsoft's artificial intelligence tay became a 'racist nazi' in less than 24 hours. https://thehackernews.com/2016/03/tay-artificial-intelligence.html.

[16] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. Let's talk about race: Identity, chatbots, and ai. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[17] Gina Neff and Peter Nagy. Automation, algorithms, and politics— talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 10:17, 2016.

[18] Rafael Lozano-Hemmer and Krzysztof Wodiczko. Zoom pavilion. http://www.lozano-hemmer.com/zoom_pavilion.php.

[19] Mike Tyka. Us and them. (kinetic installation, 2018, commissioned by seoul museum of art). http://www.miketyka.com/?p=usandthem.

[20] Annika Richterich. *The Big Data Agenda. Data Ethics and Critical Data Studies*. University of Westminster Press, 2018.

[21] KATE KAYE. Ibm, microsoft, and amazon's face recognition bans don't go far enough. https://www.fastcompany.com/90516450/ibm-microsoft-and-amazons-face-recognition-bans-dont-go-far-enough.

[22] Ian Gonsher. Beyond design thinking: An incomplete design taxonomy. http://www.cd-cf.org/articles/beyond-design-thinking/.

[23] Lars Hallnäs and Johan Redström. Slow technology–designing for reflection. *Personal and ubiquitous computing*, 5(3):201–212, 2001.

[24] Phoebe Sengers, Kirsten Boehner, Shay David, and Joseph'Jofish' Kaye. Reflective design. In *Proceedings of the 4th decennial conference on Critical computing: between sense and sensibility*, pages 49–58, 2005.

[25] Phoebe Sengers. What is reflective design. http://www.cs.cornell.edu/people/sengers/Projects/reflectivedesign/.

[26] Trevor Paglen Kate Crawford. Excavating ai :imagenetroulette, the politics of images in machine learning training sets. https://www.excavating.ai/.

[27] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Fabio Pezzoni. Ego networks in twitter: an experimental analysis. In *2013 Proceedings IEEE INFOCOM*, pages 3459–3464. IEEE, 2013.

[28] Thomas J Scheff and Don D Bushnell. A theory of catharsis. *Journal of Research in Personality*, 18(2):238–264, 1984.

[29] Thomas F Cash and Timothy A Brown. Gender and body images: Stereotypes and realities. *Sex roles*, 21(5-6):361–373, 1989.

[30] Empath. inc. vocal emotion recognition by empath. https://webempath.net/lp-eng/.

[31] Kim Ann Tolley. Theory from practice for practice: is this a reality? *Journal of Advanced Nursing*, 21(1):184–190, 1995.

[32] Ann Bonner and Gerda Tolhurst. Insider-outsider perspectives of participant observation. *Nurse Researcher (through 2013)*, 9(4):7, 2002.

[33] ACM DIS2020. Dis 2020 : Conference on designing interactive systems =https://dis.acm.org/2020/.

[34] Dan Lockton the Imaginaries Lab. The prompt report of dis 2020 student design competition. https://twitter.com/i/status/1280828242874728448.

[35] Filipe Canento, Ana Fred, Hugo Silva, Hugo Gamboa, and André Lourenço. Multimodal biosignal sensor data handling for emotion recognition. In *SENSORS, 2011 IEEE*, pages 647–650. IEEE, 2011.

[36] Catrin Sohrabi, Zaid Alsafi, Niamh O'Neill, Mehdi Khan, Ahmed Kerwan, Ahmed Al-Jabir, Christos Iosifidis, and Riaz Agha. World health organization declares global emergency: A review of the 2019 novel coronavirus (covid-19). *International Journal of Surgery*, 2020.

Appendices

A. The Questionnaire during pilot study

"Labeling" Feedback and Survey

Q1. Your sexuality

Male
Female
Prefer not to say

Q2. Have you heard about "AI Bias" before experiencing "Labeling"?

Yes
No

Q3. After being labeled by Labeling, now would you associate you feeling with an affective (emotional) state?

Hate
Angry
Disgusting
Fear
Disgraceful
Admiring
Happiness
Pleasure
Joyful
Interesting
Release
not feel anything emotionally

Q4. After seeing our statement and seeing you labeled result again, how would you associate you feeling with an affective (emotional) state?

Hate
Angry
Disgusting
Fear
Disgraceful
Admiring
Happiness
Pleasure
Joyful
Interesting
Release
not feel anything emotionally

Q5. Without "Labeling", had you been criticized with related results?

Frequently does (over 90%)
Common (70%)
Once in a while (50%)
Rarely does (30%)
Never (0%)

Q6. Do you have feeling or feedback for "Labeling" ?