Civil War Twin: Exploring Ethical Challenges in Designing an Educational Face Recognition Application

Manisha Kusuma

Dissertation submitted to the Faculty of the

Virginia Polytechnic Institute and State University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Science and Applications

Kurt Luther, Chair

Deborah Tatar

Sylvester A. Johnson

December 7, 2021

Blacksburg, Virginia

Keywords: Human-AI Interaction, Facial Recognition, Digital History, Ethical Design

Civil War Twin: Exploring Ethical Challenges in Designing an Educational Face Recognition Application

Manisha Kusuma

(ABSTRACT)

Facial recognition systems pose numerous ethical challenges around privacy, racial and gender bias, and accuracy, yet little guidance is available for designers and developers. We explore solutions to these challenges in a four-phase design process to create Civil War Twin (CWT), an educational web-based application where users can discover their lookalikes from the American Civil War era (1861-65) while learning more about facial recognition and history. Through this design process, we synthesize industry guidelines, consult with scholars of history, gender, and race, evaluate CWT in feedback sessions with diverse prospective users, and conduct a usability study with crowd workers. We iteratively formulate design goals to incorporate transparency, inclusivity, speculative design, and empathy into our application. We found that users' perceived learning about the strengths and limitations of facial recognition and Civil War history improved after using CWT, and that our design successfully met users' ethical standards. We also discuss how our ethical design process can be applied to future facial recognition applications.

Civil War Twin: Exploring Ethical Challenges in Designing an Educational Face Recognition Application

Manisha Kusuma

(GENERAL AUDIENCE ABSTRACT)

Facial recognition systems, such as those used in cities, smartphone application and airports, pose numerous ethical challenges around privacy, racial and gender bias, and accuracy. Little guidance is available for designers and developers to create ethical facial recognition systems. We explore solutions to these ethical challenges of creating facial recognition systems in a four-phase design process to create Civil War Twin (CWT), an educational web-based application where users can discover their lookalikes from the American Civil War era (1861-65) while learning more about facial recognition and history. CWT allows users to upload a selfie, select search preferences (e.g., military service, gender, ethnicity), and use facial recognition to discover their "Civil War twins" (i.e., photographs of people from the American Civil War era who look like them). Through this design process, we synthesize industry guidelines, consult with scholars of history, gender, and race, evaluate CWT in feedback sessions with diverse prospective users, and conduct a usability study. We iteratively formulate design goals to incorporate transparency, inclusivity, critical thinking, and empathy into our application. We found that users' perceived learning about the strengths and limitations of facial recognition and Civil War history improved after using CWT, and that our design successfully met users' ethical standards. We also discuss how our ethical design process can be applied to future facial recognition applications.

Acknowledgments

I would first like to extend my deepest gratitude to my advisor, Dr. Kurt Luther. Thank you for introducing me to the Crowd Intelligence Lab, during my final year of undergraduate studies, and believing in me as a Human-Computer Interaction researcher, before I even knew what that entailed. Your mentorship, patience, and unparalleled support have shaped me as a researcher and my work. I am forever grateful you entrusted me with this project.

I would like to express my deepest appreciation to my committee, Dr. Deborah Tatar and Dr. Sylvester A. Johnson, your support and valuable advice has helped shape me as a researcher and has guided me throughout my thesis.

Next, I would like to thank my coauthor, Vikram Mohanty, for his dedication and guidance throughout my graduate studies. I also wish to thank the members of the Crowd Intelligence Lab and my friends for their continuous encouragement and sympathetic ear.

Most importantly, I am grateful for my parents, Sreedhar and Ashwini Kusuma, and my sister, Tanya Kusuma, for their unconditional support and belief in me.

Lastly, I'm forever grateful for my 4.5 years at Virginia Tech and the opportunity to explore both the Blacksburg and Arlington campuses. To all the connections and memories I have made, I will always hold them dear. To the Computer Science department, thank you for providing me the opportunity to grow as an engineer and for fostering an inclusive environment.

A special thanks to all the users (present and future) of Civil War Twin.

Contents

List of Figures **ix**

List of Tables **xi**

1 Introduction **1**

2 Review of Literature **4**

2.1 Ethical Design Approaches for AI Applications 4

2.2 Applications of Facial Recognition 6

2.3 Digital History and Storytelling in HCI 8

3 Conceptualization of Civil War Twin **10**

3.1 Original Prototype 10

3.2 Limitations 11

3.2.1 The prototype ignores the limitations of facial recognition algorithms. 12

3.2.2 The prototype does not consider how historical bias impacts user experience. 13

3.2.3 The prototype failed to prioritize user's privacy and security. 14

4 Ethical Design Approach **15**

4.1 Phase I: Synthesizing Industry Guidelines 15
4.1.1 Design Goals 16
4.1.2 System Description 18
4.2 Phase II: Consulting Academic Experts 25
4.2.1 Expert Feedback 25
4.2.2 Iterative System Description 29
4.3 Phase III: Collecting Prospective User Feedback 33
4.3.1 Methods 34
4.3.2 Findings 36
4.4 Phase IV: Conducting Usability Testing 42
4.4.1 Mobile Development 43
4.4.2 Crowd Feedback 44
4.4.3 Release 47

5 Discussion **48**
5.1 Ethically Framing Facial Recognition Applications 48
5.2 Fostering General Awareness of AI 50
5.3 Interactive Digital Humanities 51

6 Limitations **53**

7 Conclusion **54**

Bibliography **55**

Appendices **64**

List of Figures

3.1 Prototype of the Civil War Twin system, 2018. 11

4.1 System workflow for the Civil War Twin website. The AI Text provides information about facial recognition throughout the system. **(A)** Upload a Photo: The user uploads a selfie of themselves to the website. **(B)** Search Preferences: The user can specify search preferences (Military Service, Gender, Ethnicity) for the twins they would like to see. **(C)** Finding Matches: The user waits while our matching algorithm determines the top-four similar looking twins. **(D)** Discovering Twins: The user can see the four twins along with a baseball card graphic. 19

4.2 Textual content on the Civil War Twin website when the user is prompted to select their search preferences for gender. The left-hand panel is referred to as the AI text. **(A)** Behind the Scenes: Part of the AI text that explains in layperson's terms how the technology works. **(B)** What Could Go Wrong?: Part of the AI text that describes potential shortcomings of the technology (e.g., gender and racial bias) and/or historical records (e.g., historical bias). Includes links to further resources to learn more about AI. **(C)** Speculative Question (added in Phase II): Part of the AI text that prompts users to think and answer the posed question. Followed by a link to provide more context for the question. **(D)** Historical Text: Part of every search preference, provides historical information and links to learn more about each identity during the 1860s. 20

4.3 System workflow for the Civil War Twin website. The green sections were added in Phase II after receiving expert feedback. **(A)** Demographics Overview: The user can view graphical visualizations of the CWT database and its relation to the US population during the 1860s. **(B)** Uploading a Photo: The user can then upload a selfie of themselves to the website or select a set of stock photos to use. **(C)** Selecting Search Preferences: The user can specify search preferences (Military, Gender, Ethnicity) for the twins they would like to see. **(D)** Finding Matches: The user waits while our matching algorithm determines the top-four similar looking twins. **(E)** Discovering Twins: The user can see the four twins along with a baseball card graphic. Biographical text is also added to provide information about each twin. The user can also view four additional twins outside their selected search preferences to learn more about the people in our database. 31

4.4 Mobile version of the Civil War Twin website. 43

4.5 Civil War Twin usability survey overview. 46

1 Graphical visualization of the database when the user selects the search preferences: Union, Civilian, Women and White. 65

2 Screenshot of the "Discovering Twins" page with the four twin matches. . . . 65

3 The baseball card created for each twin result. 66

4 Screenshot of the "Demographics Overview" page with the Census visualizations of the five different ethnicities. 66

5 Screenshot of the "Demographics Overview" page with the Database visualizations of the five different ethnicities. 67

List of Tables

4.1 Sample questions for expert feedback. 26

4.2 List of Participants for Phase III 34

List of Abbreviations

ABT	American Battlefield Trust
AI	Artificial Intelligence
CWPS	Civil War Photo Sleuth
CWT	Civil War Twin
HCI	Human Computer Interaction
VSD	Value Sensitive Design
XAI	Explainable AI

Chapter 1

Introduction

Facial recognition technology has established its presence in the daily lives of Americans, from its conceptualization in the 1960s as a military system, to its transformation into a highly commercialized product used in cities, smartphones, and airports [3, 21, 61]. Awareness of facial recognition's use in business and the criminal justice system has been documented in the media as both the government and private sectors have continued to adopt this technology. Despite public awareness about facial recognition technology, there is constant misinformation about its accuracy and security [58]. There has been well-documented research about its problems and limitations, including algorithmic bias [12, 48] and privacy concerns [68]. When looking more broadly at human-AI interaction, the complexity of AI systems, specifically facial recognition algorithms, poses a unique challenge to inform public understanding of these technologies and how it can impact a person's experiences.

One way to better inform users is to create more transparent and explainable AI models, however, the challenge is around making these explanations interpretable for end-users to help inform interactions and decisions with AI systems [19]. Prior facial recognition applications, for example, have used a user's photo to help educate the user about how an algorithm interprets their face, what information can be extracted, and how it can be used, implementing some of the methodologies of Explainable AI [54, 68]. However, the challenges of designing an AI-infused interface are reflected and amplified in facial recognition systems as a user's photo and faceprint are being analyzed by the AI algorithm. Several existing fairness toolkits [10] and Responsible AI guidelines [38, 44] focus on general AI systems, with little guidelines in how to design around the unique challenges of facial recognition platforms.

In this paper, we report on our design process and ethical issues we encountered while creating an educational web application that teaches members of the public about face recognition, as well as history. We partnered with a non-profit historic preservation organization, the American Battlefield Trust (ABT). The ABT approached us with the idea to create a fun and engaging app where people could discover their lookalike for the Civil War era and raise awareness of ABT's mission. Based on their pitch, we created an application, Civil War Twin (CWT), that would allow users to upload a selfie, select search preferences, and use facial recognition to discover their "Civil War twins" (i.e., photographs of people from the American Civil War era 1861–65 who look like them). CWT utilizes the thousands of historical portraits publicly identified on Civil War Photo Sleuth (CWPS), a free website that combines crowdsourcing and facial recognition to identify unknown Civil War photos, which has attracted over 15,000 registered users who have identified hundreds of previously unknown photos [40].

In designing CWT, we quickly encountered a number of underlying ethical issues that required careful attention, including data privacy and transparency, gender and racial bias, and limitations of our historical archive. To address these issues, we created a four-phase, iterative design process of consulting industry guidelines, academic experts, potential users, and crowd workers. Through this design process, we addressed the limitations of our original prototype by creating new design goals focused on maintaining an ethical facial recognition platform for CWT. For example, to ensure that the application was inclusive and transparent, we diversified our photo database to be more representative and created graphical visualizations to highlight the contributions of minority ethnic groups.

We conducted a qualitative evaluation of CWT with nine demographically representative users to understand their feedback and experiences using the platform. We found that the users perceived to learn more about Civil War history and the strengths and limitations of facial recognition after using CWT, and that our design successfully met users' ethical standards. We then conducted a usability study with crowd workers to test the mobile and desktop version of CWT and publicly launched the CWT website. Our primary contributions are:

- proposing a four-phase ethical design process for facial recognition systems,
- developing a novel human-AI system, Civil War Twin, focused on educating users about facial recognition and history, and
- evaluating the Civil War Twin system with demographically representative users

We also discuss how our ethical design process can be applied to future facial recognition systems.

Chapter 2

Review of Literature

2.1 Ethical Design Approaches for AI Applications

Researchers and practitioners have developed several methodologies for designing AI applications based on ethical and fairness considerations, expert values, user evaluation, and crowd feedback. We drew inspiration from existing, more general AI design approaches to create our design process for specific facial recognition-based applications.

To address ethical and fairness challenges, HCI researchers have proposed guidelines [2, 44] for designing human-AI interaction, AI fairness checklists [17, 37], and toolkits [10]. Complementing these efforts, researchers have also been studying the role of race and gender in technology and outlining inclusive practices for human-AI interfaces [16, 26]. Lee et al. created "Design Justice Network Principles" to address current design processes that frequently marginalize certain communities [42]. Many of the existing fairness checklists and toolkits look at how models and algorithms containing bias can be adapted to ethical standards [10, 37]. These guidelines and principles have similar themes around privacy, inclusivity, transparency, and accountability that can be applied to facial recognition-based applications, as documented in the first phase of our ethical design process for CWT.

To determine the values and limitations of an AI system, HCI researchers collaborate with subject-matter experts during the design process. One approach is value-sensitive design

(VSD) which focuses on identifying stakeholders in a given system and their values to determine potential limitations of an application [7, 65, 71]. Subramonyam et al. look at how AI engineers can collaborate with direct stakeholders, UX designers, to achieve a desirable AI-infused user experience when prototyping [63]. Inspired by these efforts, we consulted with indirect stakeholders (i.e., experts in the field of Civil War history, gender studies, and race and technology) in the second phase of our ethical design process, to surface the values of the communities and disciplines they study as they relate to our software prototype.

To understand public perception of interactive AI systems, end-users play a vital role in surfacing socio-technical concerns and interpretability issues. The Explainable AI initiative, commonly referred to as XAI, focuses on making AI models understandable for users through documenting the functionality, explaining the impact of AI algorithms, and shifting away from a "black box" design approach of AI algorithms [22]. Recent shifts in the XAI community focus on "user-centered XAI" approaches for bringing transparency to end-users by designing interfaces focused on addressing ethical questions about AI systems [19, 35]. When considering XAI applications in facial recognition, the models' limitations and biases directly affect the user's experience during the facial detection or facial identification process. Our third design phase looks at evaluating the CWT system based on representative end-users to see if they can interpret the facial recognition algorithm, and offers an opportunity for users to collaborate with developers about design features.

To understand the usability concerns, HCI researchers have used crowdsourcing to discover areas of confusion and platform errors. Kittur et al. looked at how crowdsourcing tasks can be used to determine the quality of Wikipedia pages by surveying workers with Likert scale and free form responses [30]. Building on prior work which uses crowd workers to evaluate the usability of websites [36], the final phase of our design process incorporates crowd workers to provide feedback on CWT. We additionally address the challenges of designing a facial recognition crowdsourcing task by taking specific consideration in protecting the anonymity and privacy of crowd workers.

2.2 Applications of Facial Recognition

Given the prominence of facial recognition applications, such as the Apple Face ID to unlock phone accessories [3] and biometric scans at US airport terminals [61], the public interacts regularly with this technology. Despite general awareness of facial recognition technology, people are still misinformed about its accuracy and the biases it imposes. In a 2019 survey, Pew found that "majorities of Americans think facial recognition can effectively identify individual people, as well as classify them by gender and race" [58]. Yet, multiple studies have shown that facial classification has lower accuracy for females and darker-skinned individuals [12, 48]. These racial and gender biases are a circumstance of the benchmarks used to create these models. For example, the LFW benchmark [28] widely used by facial recognition models is primarily composed of men (77.5%) and white individuals (83.5%) [24]. Many facial recognition systems also fail to address the limitations of the photo databases which leads to racial inequality, racial stereotyping, and gender bias within these systems [9, 26, 45, 57]. For CWT, we selected the Microsoft Azure Face API [39] because among commercial facial recognition services, it has shown some of the most substantial gains in reducing race and gender bias [27], though challenges remain [66]. Additionally, CWT recognizes the biases of its recognition model and leverages its limitations as a tool to inform users about how algorithmic bias can affect a user's twin results.

One approach to foster awareness for facial recognition technology is by creating systems that implement "learning by doing", which moves away from the fact-based knowledge and into learning through direct experiences [53]. Cloud-based face recognition services such as

Microsoft Azure Cognitive Services and Amazon Rekognition [1, 39] have enabled a variety of face recognition applications to emerge in the commercial market. Some, like the Google Arts & Culture app, help users discover matching faces in digitized artworks from museums around the world [18]. Another app called StarByFace became popular in 2020 due to the "Who Are Your Celebrity Parents?" challenge on Twitter, which helps users match themselves to their celebrity lookalikes [60]. These playful applications primarily allow users to interact with facial recognition technology through an engaging platform and inform their understanding of the technology. Although the Google Arts & Culture app went viral in 2017, it drew criticism from members of the Asian community as the search results contained a disproportionately high number of Japanese geishas [31]. Raji and Fried [47] analyzed the development of facial recognition technologies over time and cited the importance of transparency and accountability when deploying such AI systems with ethical and technical limitations. Like these other apps, CWT tries to engage the audience with a "hook" of finding your historical lookalike. But unlike them, CWT employs transparency by educating users of the various steps of a facial recognition model such as the detection and matching process.

Recent research provides more targeted efforts towards explaining how facial recognition technology functions and its applications [20, 59]. Wouters et al. created an interactive display called the Biometric Mirror focused on provoking public reactions to facial recognition technology to raise awareness about how a user's photo could be used [68]. Other applications aim to inform the public about the type of data that can be inferred by a user's facial photo [32, 54, 55] and how it can be potentially misused. Unlike these applications, CWT does not use a facial recognition algorithm to infer users' demographic characteristics (age, gender, ethnicity, etc.), but employs textual summaries to raise awareness about the technology and how it is being used.

2.3 Digital History and Storytelling in HCI

Microhistories is the process of looking at history through a narrow lens focusing on an individual or community to understand the everyday experiences within well-known historical events [50]. There have been several digital history projects focused on teaching users about uncovered microhistories. Sandwell and Lutz designed Great Unsolved Mysteries in Canadian History, a digital game-based learning platform that focuses on historical mysteries [52]. The site uses a combination of readings and links to primary and secondary sources to allow users to form their own opinion on the mystery and the real people involved.

Most similar to our work, other applications focus on primary sources such as photographs to share stories. Lee et al. [33] created a tool called Newspaper Navigator to search digitized newspapers from the US Library of Congress dating back to the 1900s, creating a visual archive of photographs and portraits based on a given search filter. Bagnall and Sherratt created a digital wall of portraits that documents the people affected by the "White Australia Policy" [6]. This "never-ending" wall of portraits aims to give users an opportunity to learn and share the stories about the people affected and their history. Within American history, the Daughters of the American Revolution created an online quiz that matched its users to a historical portrait based on shared characteristics with the individual photographed [43]. Although these applications do not use facial recognition, it does give biographical information about the user's match as an opportunity to learn more about these historical figures. CWT similarly builds on the existing work of storytelling through the use of portraits to talk about the microhistories of individuals during the Civil War era. By sharing the stories of the individuals photographed during the American Civil War, CWT aims to educate users in an engaging, non-traditional approach, by focusing on microhistories of individual experiences of the time, and encouraging empathy with those whose lived experiences differed from the user's.

In prior work, Mohanty et al. developed Civil War Photo Sleuth (CWPS) [40], a free, public website that combines crowdsourcing and AI-based face recognition to identify unknown soldiers in photos from the American Civil War era. CWPS allows users to upload an unidentified soldier photo, filters results by military service, and uses Microsoft's Face API to return a shortlist of potential matches with high facial similarity from a database of reference photos. Through this community, members have been able to develop a historical network of photographs taken during the Civil War and archive digitized portraits of the people who served in different regiments. The process of transcribing inscriptions and identifying sources of unknown portraits are all microhistory methods of gaining more information about the person photographed to eventually result in an identification. CWT's photo database uses the portraits identified using CWPS to introduce users to the soldiers and civilians affected by the war.

Chapter 3

Conceptualization of Civil War Twin

Drawing inspiration from CWPS's face recognition-based identification workflow and the Google Arts & Culture app [18, 40], the American Battlefield Trust (ABT) proposed a fun application for users to find and learn about their historical look-alike(s) from the Civil War era. This application, Civil War Twin, would help promote the visibility of their non-profit organization and contribute towards their mission of educating people about the American Civil War.

3.1 Original Prototype

Based on the proposal by ABT, we originally conceptualized Civil War Twin as a simple, web-based application that would allow users to upload a selfie, and then use Microsoft Azure Face API [39] to find similar-looking matches from a dataset of Civil War portraits. With permission from the CWPS team, we scraped photos from CWPS, along with the associated biographical information of the person in the photo, to prepare this dataset, or *twin search pool*. This biographical information included the person's name and their service details (i.e., ranks, units, branches, and regiment), along with their gender, which was collected from primary and secondary sources such as military records, medical records, and scholarly books. The original prototype (see Figure 3.1) allowed users to select preferences for their potential twins' gender (male and/or female), which would then filter the twin search pool.

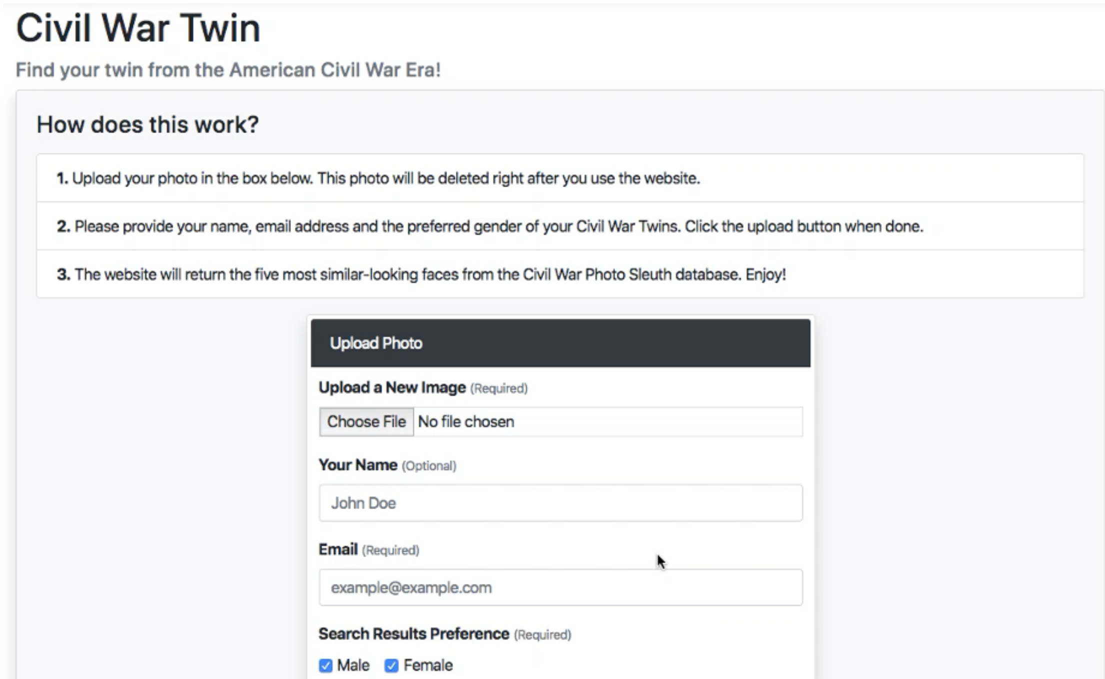

Figure 3.1: Prototype of the Civil War Twin system, 2018.

The facial recognition algorithm would then retrieve the top four similar-looking match(es) from the filtered pool, and display them as downloadable artwork ("baseball cards") for the user.

3.2 Limitations

However, this prototype had several limitations, outlined below, surrounding facial recognition, bias in the historical database, and privacy concerns, that raised ethical concerns requiring additional considerations and redesigns. To address these limitations, in this paper, we consider: How can we design an *ethical* face recognition based system?

3.2.1 The prototype ignores the limitations of facial recognition algorithms.

Black-box facial recognition systems provide little interpretability on how the input and output of a model are correlated [20]. In high-stakes scenarios like law enforcement, the "confidence score" determined by facial recognition algorithms has a direct effect on the safety and civil liberties of individuals in a society. Multiple studies have shown facial recognition has accuracy and bias problems, such as lower performance for dark-skinned or transgender faces [12, 56]. The black-box nature of such algorithmic models allows for racial and gender bias in real-time applications to go undetected. Raji and Buolamwini audited several corporate AI systems and found that despite recent improvements in classification systems, algorithmic bias is prevalent and continues to affect marginalized groups [46]

In a low-stakes scenario, like CWT, the black-box design of the prototype requires an input of a selfie, email, and gender preference to curate an output of twins. With no explanations as to how a user's twins are determined, the user has little awareness as to how their results could be directly affected by algorithmic bias. The issue of automatically guessing the user's ethnicity or gender in the prototype was a concern for several reasons. First, by using automated techniques, the chances of inferring the wrong race or gender are unacceptably high, and a wrong guess could offend a user's sense of identity or dignity. Second, it is hard to generalize what types of twins users will want to be matched with. For example, some users might prefer to see twins of a different gender, while others might find these same matches offensive.

Sensitive to the systemic nature of racial and gender bias, we shift the focus of our application to provide users with more context and awareness of the inner workings of face recognition. By positioning CWT as an educational system, users can learn about the strengths and limitations of face recognition in a low-stakes scenario to better conceptualize the implications of this technology in high-stakes applications such as surveillance and law enforcement.

3.2.2 The prototype does not consider how historical bias impacts user experience.

Due to historical circumstances, ranging from a US Navy blockade to discrimination [13], some groups from the American Civil War era, especially Confederate soldiers, women, and people of color, have fewer surviving photos [15]. CWPS's database was seeded by public collections like the US Army's MOLLUS-Mass collection, which primarily contains portraits of white Union officers from northeastern states [40]. These historical biases were echoed in the composition of the original CWPS's database and subsequently, the CWT's database. This archival bias would lead to CWT users most likely getting matched to a white male Union soldier (88.3% of photos in CWT's database). One negative consequence is that the photos in other demographic groups, having a smaller reference pool, lack diversity. Users who choose to be matched to female twins, for example, will have lower-similarity matches, and many users will receive the same matches within these groups. Additionally, due to recordkeeping practices of the 1860s, gender on CWT is limited to the male-female binary, and soldiers' race classification is based on outdated historical legal frameworks such as the "one drop rule" or inferred from membership in racially segregated military units.

Given the limitations in our photo database, we consider how a user's positionality will affect their experience on the website. Despite the historical archiving bias limiting the diversity of the reference database, the CWT database does have limited photos of civilians, women, and people of color. Focusing the CWT system on microhistories around the individuals in our database could subsequently highlight the contribution various demographic groups made to the war effort. Providing a closer look into the American Civil War from the perspective of individuals and cultural groups offers an opportunity for users to reflect on their positionality as it related to these historical figures.

3.2.3 The prototype failed to prioritize user's privacy and security.

As a rising majority of Americans feel there is more risk than benefits involved with sharing personal information, we consider the implication of data collection on the prototype [5]. Garvie et al. found that over 117 million American adults are present in law enforcement-based facial recognition databases, many without direct consent or probable cause [20]. Due the structure of the prototype, very little information is provided to the user about how their data is being used and who has access to their uploaded photo. With the rising use of facial recognition for biometrics, the subsequent faceprint created by the prototype during the facial detection step could potentially be misused and risk the security of users. By prioritizing making the data collected anonymous and limiting the entities user information is being shared with will protect user personal information and help maintain trust in the system.

Chapter 4

Ethical Design Approach

To address the above limitations, we employed a four-phase ethical design process: 1) address known issues with facial recognition based on industry guidelines; 2) consult experts in the fields of race, gender, and history; 3) iterate on feedback from representative user groups; and 4) conduct usability testing with crowd workers. In the first phase, we synthesized existing guidelines on AI ethics and fairness and applied them specifically to facial recognition systems. In the second phase, we consulted experts to get their perspective on our system and identify any oversights in our design decisions. In, the third phase of the design process engaged with end-users to identify how our system met the values and expectations of users. Finally, in fourth phase of the design process we created a mobile friendly version of CWT and utilized crowd workers to identify any usability issues on the platform. Each phase of the process was used to inform and iterate on the design decisions taken on the CWT system.

4.1 Phase I: Synthesizing Industry Guidelines

Recent advancements in AI and machine learning have opened up exciting possibilities for enabling novel, beneficial forms of human-AI interaction. However, AI's complexity, unpredictability, and over-reliance on data pose numerous challenges for designing ethical and effective AI-infused applications [21]. To address these challenges and the original limitations of the prototype, we begin our design process by reviewing the literature of industry and academic studies around ethical concerns with AI. Researchers have proposed multiple guidelines [2, 38, 44, 69] and AI fairness checklists [37] for addressing known issues. Most of these guidelines focus on general AI/ML capabilities rather than the unique challenges of face recognition. Thus, we use the existing literature to synthesize themes that can be applied to facial recognition applications, in particular, to establish new design goals for the CWT system.

4.1.1 Design Goals

Design Goal 1: System mechanisms should be transparent and explainable.

Multiple Responsible AI guidelines propose transparency and explainability to help maintain user trust and accountability in a model's output [38, 69]. Explainable practices provide users with information about the prominent limitations of the model, how AI models make decisions, and how a user's action can impact a model's output. Google's People + AI Handbook recommends developing interfaces with "partial explanations" in colloquial terms, providing context for a system's behavior [44]. If an error does occur (i.e., an unexpected result), these explanations provide an opportunity to be transparent about why the error occurred and what can be done to address the problem.

With respect to facial recognition, there is a lot of information that can be documented as to how a user's image is detected, how photo datasets are labeled, and how similarity is determined. Moving towards a "white-box" model will help users better interpret the results they see and the steps taken by facial recognition systems. In the context of CWT, a facial recognition-based system adopting these principles can help users understand how faces are detected in an image, how image search pools are constructed, and how similar-looking images are retrieved. Framing CWT as an educational platform will allow us to focus on explaining the inner workings of the underlying facial recognition algorithm as the user goes through the process of finding their twin.

Design Goal 2: Systems should work towards being fair and inclusive.

Responsible AI guidelines emphasize the idea that AI can introduce and reinforce unfair biases [2]. Madaio et al. outlined an AI fairness checklist recommending that systems should consider fairness-related harms from diverse perspectives, including different demographic groups [37]. As many have noted, facial recognition algorithms are trained with biased datasets, which leads the technology to be less accurate for some ethnic and gender groups [24]. For example, African American women are the most inaccurately identified group, while white men have the highest accuracy rates [12]. Such studies have led companies like Microsoft to invest specifically in reducing racial and gender bias in facial recognition, making substantial gains, though gaps remain. Given that these biases cannot be completely minimized [46], CWT provides users more control over the algorithm by allowing for the customization of twin results through the selection of military service, gender, and ethnicity. We also choose not to use algorithmic-based detection techniques to determine the ethnicity and gender of the user or the historical figures in our dataset.

In addition to algorithmic bias, there is also historical bias in the dataset of Civil War portraits used. As noted in the Limitations, many demographics lack representation in our database. For the CWT system, we can focus on educating users about these historical inaccuracies and focus on microhistories about the soldiers and civilians. Providing users a look into the individuals in our database and their stories can help highlight the contributions made by marginalized groups in the American Civil War. The use of graphical visualizations provide an opportunity to educating users about the numerical distribution of photos and emphasize the smaller ethnic groups represented in the database. We also educate users about how race was determined in the Civil War era and the gender binary present in our dataset.

Design Goal 3: Systems should collect and handle data responsibly.

Microsoft's Responsible AI guidelines underline the importance of privacy when designing AI products [38]. In AI applications, privacy is based on providing notice and consent for how data is being used, and security focuses on mitigating user risk. With facial recognition systems, minimizing a user's security risks can be achieved through protecting the photos being uploaded and the subsequent faceprint created by the algorithm from being used in the public domain. Given the rise in public concern over surveillance, data, and misuse [5], providing the user with information over how their data is used, will help them make informed decisions about what data they are comfortable sharing. In the CWT system, we prioritize limiting the amount of personal information collected to avoid any security or privacy concerns. We also emphasize giving users control over how their data is being shared and used.

4.1.2 System Description

Based on these design goals, we designed and developed a new, web-based version of Civil War Twin (see Figure 4.1). The web application was built using Javascript, HTML, CSS for front-end interfaces and Python and Django for back-end processes. PostgreSQL was used to store the dataset of historical photos and Amazon S3 was used to store user data.

The website employs the educational technique of *learning by doing* so users can directly interact with AI to learn about facial recognition technology [53]. Through the process of discovering their look-alikes, users learn about the various microhistories from the perspective

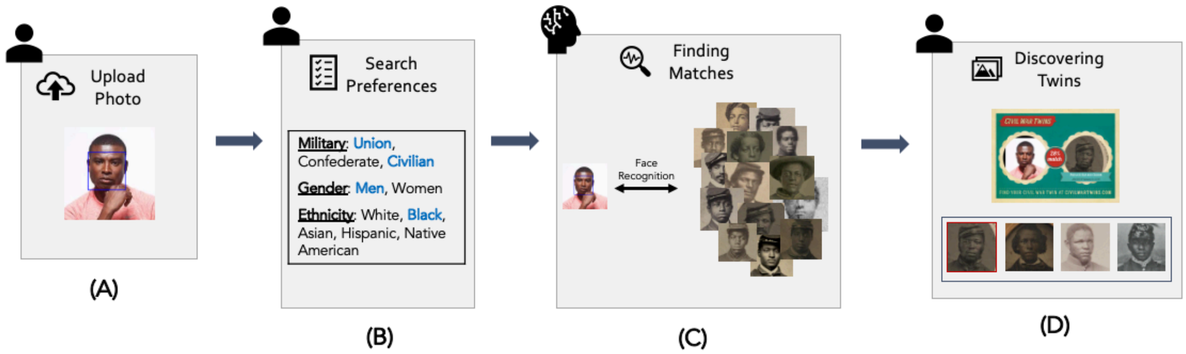

Figure 4.1: System workflow for the Civil War Twin website. The AI Text provides information about facial recognition throughout the system.
(A) Upload a Photo: The user uploads a selfie of themselves to the website.
(B) Search Preferences: The user can specify search preferences (Military Service, Gender, Ethnicity) for the twins they would like to see.
(C) Finding Matches: The user waits while our matching algorithm determines the top-four similar looking twins.
(D) Discovering Twins: The user can see the four twins along with a baseball card graphic.

of individuals during the American Civil War.

Throughout the website, the *AI Text*, a side panel, provides more information about facial recognition technology (Design Goal 1). The first part of the side panel has a "Behind the Scenes" section which explains in layperson's terms how the technology works. The second part of the side panel has a "What Could Go Wrong?" section which describes potential shortcomings of the technology and/or historical records (see A, B in Figure 4.2). This panel allow for a more nuanced look at how facial recognition works and its limitations.

The website also provides historical text and links where users can learn more about the various demographic groups and individuals who lived during the 1860s (see D in Figure 4.2).

Users discovers their look-alikes ("twins") from the American Civil War era by following a simple four step process: i) *Uploading a Photo*, ii) *Selecting Search Preferences*, iii) *Finding Matches*, iv) *Discovering Twins*.

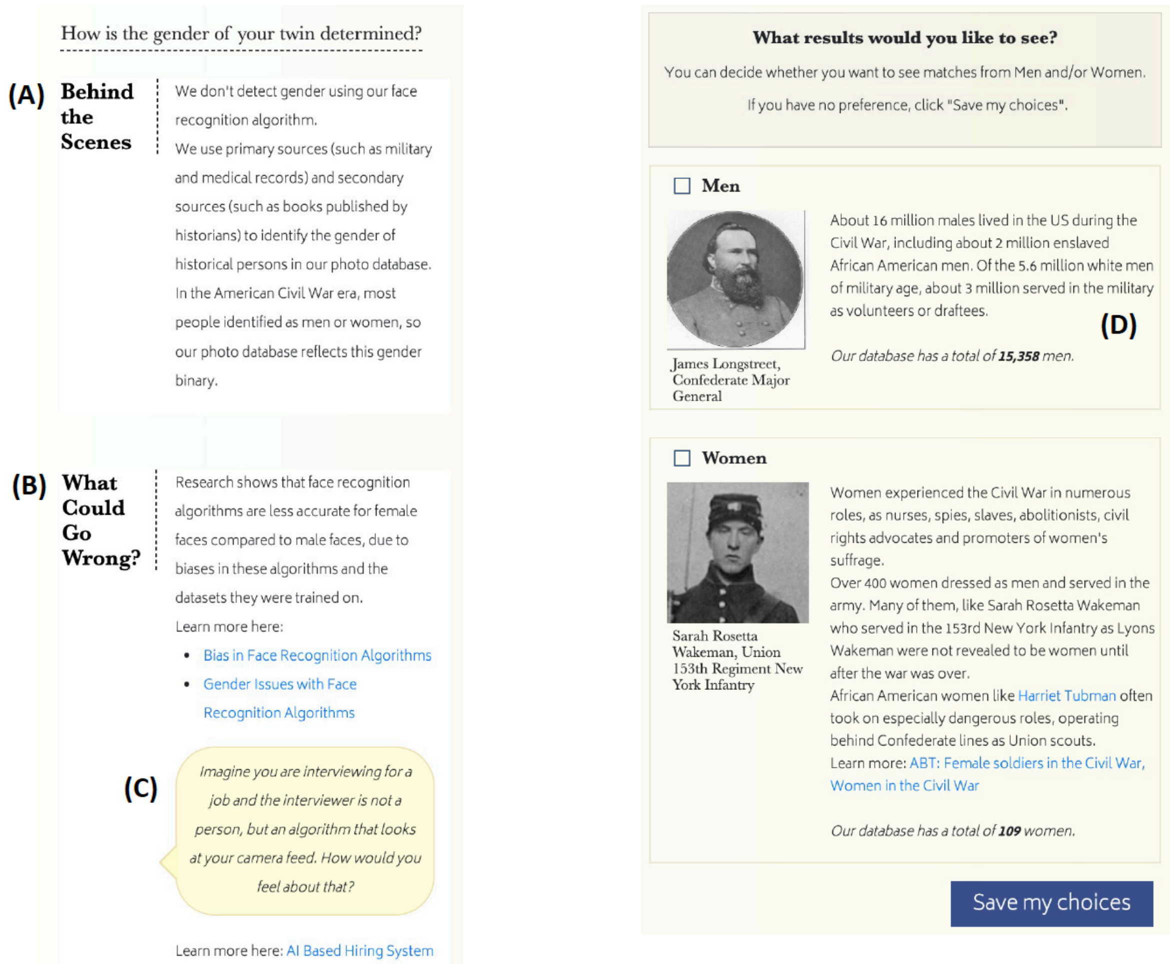

Figure 4.2: Textual content on the Civil War Twin website when the user is prompted to select their search preferences for gender. The left-hand panel is referred to as the AI text.
(A) Behind the Scenes: Part of the AI text that explains in layperson's terms how the technology works.
(B) What Could Go Wrong?: Part of the AI text that describes potential shortcomings of the technology (e.g., gender and racial bias) and/or historical records (e.g., historical bias). Includes links to further resources to learn more about AI.
(C) Speculative Question (added in Phase II): Part of the AI text that prompts users to think and answer the posed question. Followed by a link to provide more context for the question.
(D) Historical Text: Part of every search preference, provides historical information and links to learn more about each identity during the 1860s.

Uploading a Photo

This page provides the user with context on how facial recognition detects facial features and explains how certain issues in photo quality can affect the detection process.

The user begins the process of discovering their look-alike by uploading a selfie to the website (see A in Figure 4.1). The system instructs the user to not upload photos of other people (friends, celebrities, etc.) without their permission to avoid potential privacy issues or misuse. The user can then provide their name (optional) and email address. The user is then asked to consent to the IRB form outlining our study. Then, CWT uses the Microsoft Face API to verify if a face can be detected in the uploaded photo. After a successful detection, the user can see a bounding box that appears on their face. If a face is not detected, the user is prompted to try again with another photo.

On this page, the user can also view CWT's Privacy Policy which is accessible throughout the website. The policy page presents a straightforward description of what information is required, how it is being shared, and when their data is deleted (Design Goal 3). We ensured that the system does not save any personal information, giving the user control over their data. The user is only required to provide a valid email and photo, both of which will be deleted upon exiting the website. The user's photo and faceprint are not permanently stored on any web server (ours or Microsoft's) and are not used to train any facial recognition models.

Selecting Search Preferences

After uploading a photo, the user can specify search preferences (military service, gender, ethnicity) for the twins they would like to see (see B in Figure 4.1). The process is split across three pages as follows:

1. *Military Service:* This page provides the user with information on how military service is determined for their potential twins. The user can then choose from the Union, Confederate, and/or Civilian categories to filter their pool of potential twins. We use primary and secondary sources to identify and label the military affliction of the persons in our photo database.

2. *Gender:* This page provides the user with information on how gender is determined for their potential twins and explains how gender bias affects facial recognition algorithms. The user can choose from the Man and/or Woman categories to filter their pool of potential twins. The gender of a twin is determined from historical and medical records. The categories provided for the user reflect the historical dataset and thus, the practices of the era. For example, gender is limited to the male-female binary. Instead of trying to hide these categories by mapping them onto modern-day labels, we employ seamful design [14] by presenting the historical context behind these categories and inviting the user to consider their own relationship to them.

3. *Ethnicity:* This page provides the user with information on how ethnicity is determined for the potential twins and explains how racial bias affects facial recognition algorithms. The user can choose from the White, Black, Native American, Hispanic, and/or Asian categories to filter their pool of potential twins. Race in the context of the American Civil War era was determined based on a person's physical features and ancestry [41]. The five ethnic groups are reflective of the people represented in our photo database as soldiers' races were classified based on archaic legal systems and segregated military units.

These search preferences help to mitigate the effects of algorithmic bias by providing an alternative for algorithmic-based detection techniques to determine the ethnicity and gender of the user or the historical people in our database. We attempted to mitigate historical biases and increase inclusively (Design Goal 2) in several ways. First, we completed two targeted database enrichment projects to increase the number of photos of several underrepresented categories: African Americans, Asian, Native Americans, Hispanics, and women. Consequently, our collection of African American Union soldier portraits (128 photos), although small, is believed to be the largest digital collection of its kind in existence. Our database now has 13,861 Union soldiers, 1,475 Confederate soldiers, 132 civilians, 15,357 men, 110 women, 15,272 White individuals, 144 Black individuals, 16 Native Americans, 11 Asians, and 25 Hispanics represented in our database. Although our database is primarily white Union men, this is a reflection of the historical biases echoed in the original CWPS database [41]. Second, we created a real-time interactive visualization that shows how the user's preferences affect the search pool (see Appendix 1). This visualization helps educate the user of our database construction as they are choosing their search preferences.

Finding Matches

Based on the military, gender, and ethnicity preferences the user selected, our algorithm determines the top four similar-looking twins from the database of reference photos. This interstitial "matching" page provides the user with information on how their face is being compared to their possible twins and how the confidence threshold affects twin results.

The user waits for a few seconds for the matching algorithm to identify their twins. We implemented a technical solution for the matching algorithm to ensure that all possible twins are within the specified search preferences the user selected. First, we curate the CWT database to be a subset of photos with only identified portraits from reliable sources (Library of Congress, National Archives, National Park Service, etc.). Given that the CWT database is based on photos scraped from CWPS, which contains thousands of user-contributed photos, this curation process removes several types of unwanted results (non-historical portraits, unidentified twins, etc.) [40]. Second, the CWT database is divided into 10 *identity sets*, photos of historical persons from each of the available demographic categories (Union, Confederate, Civilian, Man, Woman, White, African American, Native American, Hispanic, and Asian). Third, the system builds a *twin search pool* of potential candidate photos based on the preferences selected by the user. This twin search pool is created by compiling together the identity sets from the user's selected identities. For example, if the user selected Union, Civilian, Men, African American we would build a twin search pool of historical people within the Union, Civilian, Men, African American identity sets. Fourth, the twin search pool is sent to the Microsoft Face API along with the user's uploaded photo to find the top four most similar-looking faces based on Microsoft's confidence score. These top four results become the user's Civil War twins.

Discovering Twins

Once the twins are found, the user can view their four twin matches. This page explains how the similarity score is determined for each twin (see Appendix 2). The system also emails the user a copy of their twin results.

The user can then view a baseball card graphic (see D in Figure 4.1) for each set of twins. The user can download the baseball card graphic for saving locally, printing, or sharing (see Appendix 3). However, there is no direct way to share the baseball card on social media as the system does not create persistent URLs as a privacy protection measure (Design Goal 3). The user can also learn more about their twins by clicking on their twin's CWPS profile links, which displays additional biographical and military records for that individual. Finally, the user has the option to contribute additional photos (e.g., from their personal collections or public sources) to further enrich underrepresented categories in our database, or continue learning about AI through the additional links provided.

4.2 Phase II: Consulting Academic Experts

After implementing the redesigned system described in Phase I, we consulted three academic experts in Civil War History (E1), Gender Studies and Ethics (E2), and Race and AI (E3) to critique our design decisions. Specifically, the experts helped us understand how the Phase I design engages with some of the sensitive topics around race, gender, history, and AI, and to have a better understanding of the societal implications of our proposed system. The experts' scholarly expertise and lived experiences helped (re)frame the design goals of our system and validate the design designs from Phase I. Based on their feedback, we iterated on our design and developed a high-fidelity prototype.

4.2.1 Expert Feedback

All three experts were tenured faculty at our university who were not previously familiar with CWT, but whom we knew through mutual research interests in HCI. In individual one-on-one sessions, each expert was first introduced to the background, motivation and goals of the CWT project, followed by a demo of the Phase I prototype, and a high-level overview of the Phase I findings and design choices. We then asked the experts about their general perceptions of the CWT prototype, along with a series of questions specific to their specialization (see Table 4.1).

Overall, the experts found the "hook" of matching with one's Civil War twin, along with the concept of a baseball card as a shareable proof of match, to be a fun and engaging experience, while appreciating the simple workflow of the application. The experts also perceived the

1. How do you feel about the overall premise, motivation and goals of the project? What value (if any) do you see in this project?
2. How can this project address rising ethical concerns related to facial recognition?
3. How might African American users respond to the way this project addresses the topic of Civil War history?
4. How might the focus on facial recognition software be perceived by people of color?
5. What are some issues regarding race, in the context of the Civil War era, this project may have overlooked?
6. What are some ethical considerations when asking users for their search preferences related to race?
7. What are some concerns regarding gender identities and gender roles this project may have overlooked?
8. How can we use this project to encourage contribution of historical photos to the CWPS database?
9. How can this website potentially be misused? Should aspect(s) of the platform be modified to prevent this?
10. How should we address the lack of photos for certain genders and ethnic groups in the CWT database?

Table 4.1: Sample questions for expert feedback.

educational goals of the application favorably, and believed that the Phase I design choices were effective in supporting these goals. At the same time, they also pointed out 4 focus areas for the system, outlined below, suggesting opportunities for improvement and updated design goals.

Experts mentioned the need to enrich our photo database and acknowledge the impact of historical inequities.

All three experts found the real-time interactive visual chart on the search preferences page to be effective in conveying the demographic distribution of the database. Even though they appreciated the design efforts towards being transparent and acknowledged the challenges of historical bias, there were concerns about the dominant representation of white Union men over other groups. E1 said, *"The number and type of photographs of African Americans are* *going to be very different. So, I think just being really transparent is the right way to go"*.

E3 pointed out the historical bias from a different lens by comparing the demographic distribution in the database to the actual population of the 1860s, stating that even if every woman in 1860s was photographed, the number would would still be disproportionately low compared to photos of men. In addition to acknowledging the disproportionate demographic distribution of the existing dataset, the experts also suggested other changes, such as using the more ethnically inclusive "Black" as a category label instead of "African American." E2 stressed the importance of being upfront about the gender binary limitation of the dataset, while recognizing more fluid representations of gender in both the historical and modern eras.

Design Goal 4: The system should highlight the contributions of minority groups.

Experts cautioned about the general public mistrust of face recognition-based systems.

E1 raised concerns about the general sentiment in the media towards facial recognition technology and warned about possible public hesitancy towards CWT: *"There will be a substantial segment of the public who are not going to upload their photo regardless of what the site says about privacy"*. E1 emphasized the need to make the privacy policy easily accessible from every page and remind users about how their data is being used. Along similar lines, E2 raised concerns of users being wary of immediately uploading their personal photo as *"that might make them a little bit uncomfortable."* E2 further predicted that some people, out of general distrust towards big technology companies, might be cautious about sharing their photo with the Microsoft Face API.

Design Goal 5: The system should be usable even without providing a personal

photo.

Experts emphasized the use of microhistories during the Civil War era to foster empathy.

All the experts identified a clear opportunity for CWT to use microhistories as a tool to tell individual stories from the Civil War. E1 encouraged us to use the CWT platform to foster connections: *"I do think there's value in just thinking about the connections between us as individuals and the people who lived through and fought the Civil War as individuals."*. E3 challenged us further to not only form these connections but to build on the notion of empathy and enable users to empathize with the various people in our database and to understand their stories. As E3 stated, *"One conceptualization of empathy is that you want to stand in another person's shoes"*. E2 added that CWT achieves a deeper level of learning that focuses on sensitivity and empathy which can be further developed through the use of microhistories.

When discussing the search preferences of military, gender, and ethnicity, experts believed that giving users more control over their potential matches was a justifiable approach. E1 talked about how users could be presented with unwanted twin results if there was no way of specifying preference: *"I think that's a really valuable feature giving users the choice, rather than you know, it is a potential minefield"*. However, E3 warned that giving users control of their search preferences can also lead to confirmation bias: *"It's quite dangerous to simply confirm people's existing biases"*.

E3 also emphasized that the system should not be *"reinforcing stereotypes, both for black people and for people of color and for women."* E3 later expanded on the idea of "empathy twin's" i.e., deliberately (with permission) showing twins results outside the user's selected preferences, which would aid in fostering connections and empathy as well as pushing back against stereotypes.

Design Goal 6: The system should utilize empathy as a tool to understand history.

Experts saw the potential to further challenge existing preconceptions of AI

E2 affirmed CWT's goal of educating users about the limitations of facial recognition. E2 said, *"It promotes public understanding of technology, and it does so not just by being a billboard or infomercial, but it gets them involved in using the technology"*. According to E2, the "learning by doing" aspect of our system could showcase the benefits of *"using the technology in order to help people understand the technology"*. While all the experts appreciated the informational panel about the AI Text, broken down into "Behind the Scenes" and "What Could Go Wrong?" content alongside the interactive interface, they also raised concerns about the verbosity of the text. E3 suggested prompting users to question the technology. E2 elaborated on a similar notion: *"It's one thing to tell someone this technology has pros and cons. It's another thing for them to experience the pros and cons by getting a result."*

Design Goal 7: The system should encourage users to speculate about consequences of facial recognition and AI.

4.2.2 Iterative System Description

Based on the feedback received by the experts and the new design goals we iterated on CWT by adding new features and modifying existing features.

AI Text & Speculative Questions

Inspired from speculative design approaches [67], we modified the AI Text throughout the website (see C in Figure 4.2) to include speculative questions around facial recognition (Design Goal 7). By employing simple elements of speculative design, users can begin to participate in the conversation about AI ethics and learn through an interactive process. The text also includes links to articles where the user can find other examples of facial recognition in the world and learn more about the technology.

Demographics Overview

We added a new visualization page prior to the initial photo uploading step (see A in Figure 4.3). This page provides the user with graphical representations of the CWT photo database (see Appendix 5) and its relationship to the 1860s US Census (see Appendix 4). We were motivated to add this page to highlight the contribution of demographic groups (Design Goal 4) and to be transparent about how our system is affected by historical bias. The page explains to the user how we collected the photographs in the dataset and the bias associated with them. The page also poses a speculative question — "Do you think your photo may be part of any such databases where facial recognition is being used?" — for the user to think critically about AI and existing public photo datasets.

The user can also view and interact with a series of population charts (i.e., one for each search preferences military, gender, and ethnicity) to show the distribution of demographics both in our photo database and in the 1860s US Census. The census information was obtained from the historical 1860 US Census records. Through the design process, we tried a variety of different charts (pie charts, horizontal and vertical bar charts, icon charts, etc.) to show the distributions of our database but we faced issues of accuracy, clarity, and simplicity. We

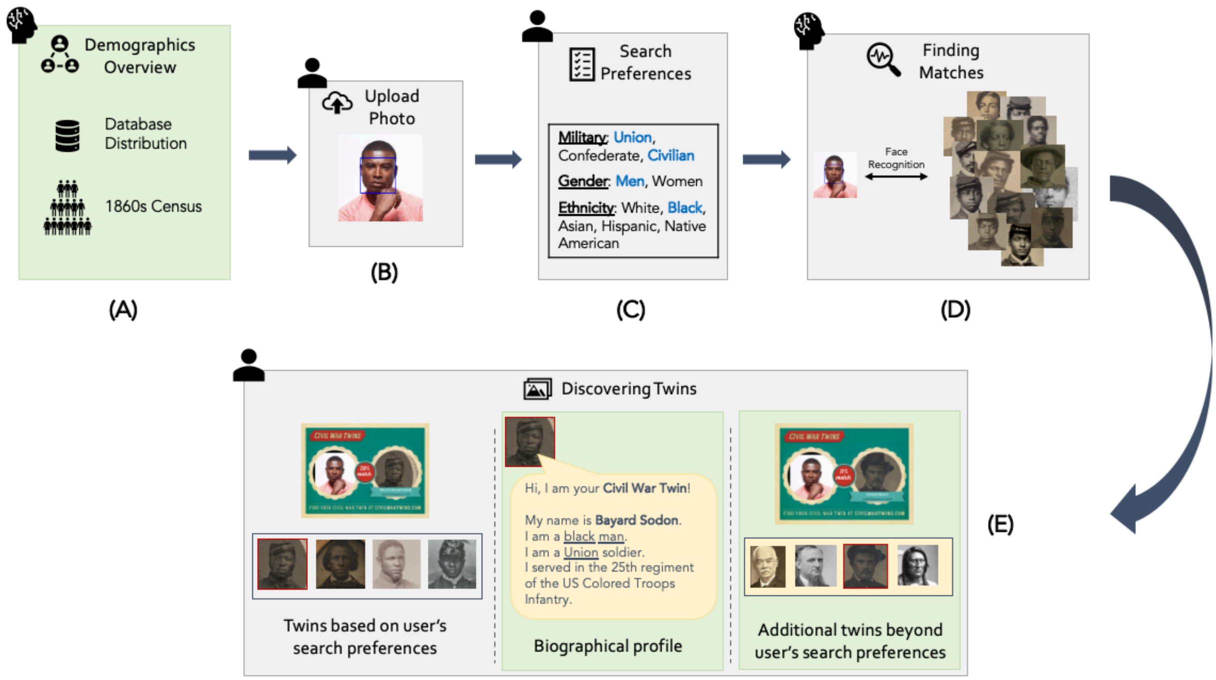

Figure 4.3: System workflow for the Civil War Twin website. The green sections were added in Phase II after receiving expert feedback.
(A) Demographics Overview: The user can view graphical visualizations of the CWT database and its relation to the US population during the 1860s.
(B) Uploading a Photo: The user can then upload a selfie of themselves to the website or select a set of stock photos to use.
(C) Selecting Search Preferences: The user can specify search preferences (Military, Gender, Ethnicity) for the twins they would like to see.
(D) Finding Matches: The user waits while our matching algorithm determines the top-four similar looking twins.
(E) Discovering Twins: The user can see the four twins along with a baseball card graphic. Biographical text is also added to provide information about each twin. The user can also view four additional twins outside their selected search preferences to learn more about the people in our database.

wanted the charts to be easy to understand and help visualize the small populations in our dataset. After experimenting with numerous visualization types, including pie charts and bar charts, we ultimately chose icon graphs as they were the most visually intuitive and allowed the user to flip between the visual representation of the photo database and the US Census.

Uploading A Photo

By continuing to be cognizant of how the system collects data (Design Goal 5), we added a new feature on this page that provides the user with a set of stock portrait photos to use instead of their own. If the user does not want to upload a personal photo, they can choose from a set of demographically diverse stock photos and continue the entire workflow to discover twins for the stock photo.

Selecting Search Preferences

We updated the exemplars for each search preference type to present microhistories that reflect more diverse experiences. For example, for the "Gender" category, we included a wartime photo and biography of Albert Cashier, a Union soldier who grew up female but lived as a man for fifty years after the war.

Discovering Twins

For each twin, we added a new biographical, first-person prose on this page for the user to learn more about their twin's life while fostering natural connections (Design Goal 6). The prose contains information such as the gender, race, birthday, and military affiliation of the twin (see E in Figure 4.3). These basic identifiers help humanize the people in the photos to offer a point of introspection for the user. We also added a new page for “empathy twins”, where the user can see additional twins beyond their search preferences. We used Celery and Redis to implement an asynchronous function that finds a user’s additional twins. This new page shows the user four additional twins from a combination of search preferences (see E in Figure 4.3), providing biographical prose about these additional twins and baseball card graphics. By giving users an opportunity to learn more about the historical people in the database with whom they have a connection via facial similarity, but are demographically outside the user’s selected search preferences, the system aims to foster empathy and perhaps change perceptions about who fought and lived during the Civil War.

4.3 Phase III: Collecting Prospective User Feedback

After implementing the designs that resulted from Phase II, we recruited nine prospective end-users (i.e., direct stakeholders) to use the CWT application and provide feedback on their experience via interviews. We recruited a demographically diverse set of participants to understand the effectiveness of the system’s learning goals and gauge public perception of the system’s ethical issues and our attempted mitigations. This qualitative study to evaluate CWT helped capture how users’ perceived learning about the strengths and limitations of facial recognition and Civil War history improved how our design successfully met users’ ethical standards, and how users interacted with the AI technology.

Table 4.2: List of Participants for Phase III

Participant	Gender	Ethnicity	Age Range	Knowledge about Facial Recognition (1-5)	Interest in US History (1-5)
P1	Man	Hispanic, Latino, or Spanish origin, Black or African American	30 - 39	4	5
P2	Man	Hispanic, Latino, or Spanish origin	30 - 39	2	5
P3	Nonbinary	Asian	18 - 29	2	3
P4	Woman	White	50 - 59	2	3
P5	Man	Black or African American	30 - 39	1	4
P6	Man	White	40 - 49	1	5
P7	Man	White	60 - 69	3	4
P8	Woman	White	70 - 79	3	5
P9	Man	Black or African American	18 - 29	3	3

4.3.1 Methods

Recruiting Participants

We recruited nine participants (see Table 4.2) from different cultural organizations at our university, as well as online forums on Facebook and Reddit. We attached an interest form to our recruiting emails and blurbs that was used to screen participants based on demographics (e.g., race, gender, age), knowledge of facial recognition technology, and interest in history. We aimed to recruit participants with a diverse range of ethnicities, genders, and ages who were interested in United States history — representing the ABT's target audience — and had limited knowledge of facial recognition technology. The selected participants completed a consent form and a demographics survey, with IRB approval. The participants were compensated with a $20 gift card at the end of the study.

Procedure

Once participants consented to the study, they were given access to the CWT website and asked to provide availability for a 60-minute remote study session. Participants were encouraged to familiarize themselves with CWT prior to the interview session. They could optionally send a screen recording of their experience on the website as they walk through the process of uploading their photo (or selecting a sample photo), selecting their search preferences, and viewing their twin results. This screen recording was intended to help us observe how new users interacted with the website and assess any unexpected behavior.

The interview session was conducted via Zoom video conferencing. We recorded the participant's video and audio with their consent. We gave the (two) participants who did not send in a screen recording ahead of time 15 minutes at the beginning of the interview to use the website while we screen-recorded their experience. The interview questions were divided into three main themes that corresponded to recurring topics throughout the project: facial recognition, Civil War history, and ethics. We also asked more targeted questions about specific features such as the database charts and the baseball card graphic.

Analysis

We fully transcribed the interview audio recordings and used MAXQDA, a qualitative analysis tool, to organize participant's feedback. We used inductive thematic analysis to categorize participant quotes based on the themes that emerged from the transcripts (e.g., prior knowledge, historical bias, empathy, accuracy of twin results) [11]. We iteratively grouped together existing themes to organize quotes related to our main research topics.

4.3.2 Findings

Based on our analysis, we synthesized three main themes from the user study of CWT: understanding of face recognition, Civil War microhistories, and ethical values.

Understanding of Face Recognition

Prior to using the website, most participants had a general but limited awareness of how facial recognition technology works and is being used in society. After using the CWT system, participants perceived learning more about how the technology works and is being used, along with its strengths and limitations. The two main ways participants perceived learning was through direct interaction and reading the AI Text in the application. Below, we look at how the participants used CWT to improve their understanding of facial recognition.

Users explored how different inputs can affect the results of the face recognition algorithm. Eight out of the nine participants tested the website by uploading their own photos and one participant instead used the sample photos. Some participants even tried the site multiple times with different pictures of themselves to see if their twins differed. A majority of participants played around with their search preferences, specifically gender and ethnicity, to see how well their picture would match a twin of a different identity. P6 stated, *"I thought it was a really neat website, to be able to kind of play around and see do I match anybody, but also just the facial recognition technology was kind of neat to play with."*

A minority of participants had a more directed, less playful approach. For example, P5 believed that changing his search preferences enabled him to refine the search results to achieve better matches: *"Maybe I have this wrong, but choosing and narrowing down the search on my end makes it that you actually could get something more accurate, is that the* *idea?"*

Users used facial features to determine the similarity of their twins. Participants focused attention on one or two high-diagnostic facial features to justify their twin results. For example, P8 noticed that her main twin had *"one eye that was more droopy than the other, as in mine, and I saw that, that was the connection."* P3 stated how the algorithm *"was able to detect accurately, like, femininity in my face"*. Participants also used such physical characteristics to gauge the accuracy of the confidence score produced by the face recognition algorithm.

Users considered how contextual factors can affect the face recognition algorithm. Some participants experienced firsthand how limitations in the dataset led to the low facial resemblance with their twin(s). For example, P2 strategically looked at multiple facial features to determine similarity and recognized the limited amount of Hispanic photos in our dataset: *"Slight similarities with like the eyes, maybe, but not too much like the chin [...] I think it was like a 36% match it wasn't a perfect match, but then again, I don't think [any] one looks like me at that time."*

P6 pointed out how the relatively small size of the dataset could affect the accuracy of twin results: *"I didn't think that those pictures look that much like me, now granted this is a database of like as you said 15,000 photos as opposed to 100 million."*

Other participants, drawing on the AI Text and speculative questions, made connections to broader societal issues. For example, P3 acknowledged the importance of the dataset when trying to determine the accuracy and how in different contexts such as in policing it could be misused. P3 said, *"With this specific set of data and images I would not be able to determine whether or not it is accurate. Like I wouldn't be using this in like police facial recognition* *or anything like that."* Along these lines, some participants considered how the app's design surfaced their own biases, which could have negative impacts in higher-stakes scenarios. P5 pointed out that the application was fun within the context of historical twins, but alluded to a possible case of confirmation bias while comparing his selfie to his twin's photo: *"...maybe my brain is also making these connections that don't exist because this website told me that it has a 36% match."*

Users speculated about how facial recognition works and is currently being used. Participants referred to the AI Text for justifying the results of the facial recognition algorithm and how it works. When uploading a photo, P4 recognized,

> *"I was lucky because I had a good picture that really, I guess, must have just identified the characteristics pretty clearly [...] I mean it even talks about it on the left side of the screen, you know if your image is not that clear, your characteristics are not that clear, then there is going to be less of a chance that it's going to be a good match."*

Many participants explored the links provided and spent time to answer the speculative questions posed. P5, for example, answered as he uploaded his photo, *"If I was trying to identify someone which facial features, what would I pay attention to? The nose and mouth right? And facial hair? I think those are the main ones right."* P7 explored the articles linked on the AI text after reading the section: *"I think it was on the left-hand side that talked about where can facial recognition go wrong — false identification and false imprisonment. I was just reading some of the links that the program provided right before I logged in here, those are interesting."* After using the application, P4 speculated about future applications of facial recognition technology: *"It just made me kind of curious how this possibly could become more utilized in the future. Like what if somebody looks very much like you, and they* *are using face recognition for security purposes or to enter a building. How does that work? How would that work?"* Furthermore, P8 and P5 suggested areas of the AI Text where they wanted more details, such as adding more information about what facial landmarks were being compared during the matching process.

Civil War Microhistories

Most participants recalled an instance in K-12 education as their main introduction to learning about the Civil War. Some participants went on to learn more about the war from local museums, landmarks, genealogy, and personal research. CWT complemented this prior exposure of participants by allowing them to learn about the perspectives of different cultural groups and individuals during the 1860s and to empathize with their experience.

Users learned about the experiences of different cultural groups during the Civil War era. The process of selecting search preferences was effective in teaching participants about the different demographic groups that participated in the war. A majority of participants were surprised to learn about the critical roles that women played in the war. P1 thought it was interesting *"to know that women served, I didn't know to what extent."* P8 said, *"It's surprising to see that there were Asians and Hispanics that participated in this, and I was unaware of that. So that kinda opened up my eyes to what was going on."* P6 also described how CWT's experiential approach of teaching history is different from traditional mediums such as classrooms, textbooks, or articles:

> *[CWT] I think, brings it to life more so, instead of just being a date and a place [...] So many kids these days, are like, 'Oh, history is boring', but there's more to it. There are real people, there are real consequences, there are real actions that happened. I think having this [CWT] kind of brought that to life more.*

Users formed a connection to their twins while trying to learn more about their lives. Participants explored biographical profiles on CWT, CWPS profiles, and other external sources (e.g., search results from Wikipedia, Google, etc.) to learn about their twins. P2 did not expect to find other sources of information about his twin, stating, *"This guy [Shawn Moffitt] was someone that was known and he lived through the war, which is a highlight to see [...] I looked up the name and stuff and then he popped up on Wikipedia and I was like, wow, because I wasn't expecting that."* P6 expressed a genuine curiosity for his twin: *"I wanted to learn more about him and I kind of wanted to go into it a little bit more to see, oh, what's his history? Do we know what happened to him? Did he survive the battles? And you know, does he have a family?"* One participant P4 even felt a familial relation to her twin, stating, *"She could have been my sister, you know. She looked like that!"* Some participants also sought more biographical information about their twin(s) directly on the website. P8 requested additional sources, saying, *"It would be nice to have some more information, you know, try these sites or, we got most of our photographs from here, try these sites."*

Users wanted to discover new people from our database. Most participants, while using the website for the first time, did not select any search preferences to exclude or tailor results. Some felt it was the best way to test the system, while others did not want to restrict the facial recognition algorithm. P7 was comfortable about being matched with any results, due to his prior use of genealogy technologies:

> *I'll be honest, I clicked on all [preferences] right away. I'm here to see what I'm matched with. I'm a little familiar with ancestry.com and their DNA database matches your DNA with your cousins. And I see all ethnicities and, of course, all sexes there, so I was accustomed to the fact that I could come up as a Hispanic or a person of color or a black person.*

Most participants, like P4, mentioned that they *"wanted to be open to anything"* out of curiosity. P6 enjoyed looking at additional twins outside of his search preferences even if the twins did not seem like an accurate match. P6 said:

> *[The additional twins] gave me females, gave me people of color, gave me other things so, even though I may or may not match with someone else like that, it did give me a chance to also see what else there was and to really kind of bring home the point that there's more than just, you know, rich white men, you know there's more to it with Native Americans being involved.*

Ethical Values

Participants shared that they felt comfortable and safe using the platform, while recognizing the efforts taken towards creating an inclusive experience and being transparent about how personal information was being used.

Users felt the platform made efforts to be inclusive. P3, who identified as non-binary, offered their perspective on the gender binary that exists on our platform: *"I'd say it's very difficult for there to be inclusion with any like system because the recording of historical information has been so white male-centered. [...] I think you've done the best that you could have done with this specific circumstance and set of data. So I thought it's pretty cool."* They were also aware of the limited number of gender and racially diverse people photographed during the 1860s. In P3's words, *"I don't think there's really like any data of gender diverse people. So, it's not necessarily something that you can accommodate for and you have only so much, like, racial data."*

Users felt the platform prioritized their privacy. Many participants expressed trust in the application because of the ubiquity of the privacy policy on the website. Most participants noticed the link to the "privacy policy" page, yet they never directly clicked to access the link, though they found comfort in its accessibility. P4 said, *"I skimmed over the privacy policy page because I was under the assumption and trust of this educational institution."* All participants felt comfortable uploading their photo to the website. The one participant who did not did not have a selfie available on his computer. P8 pointed to the website's explanation about not storing photos or faceprints, saying, *"I didn't have any problem, and particularly, when it says that your photo will not be saved."* Similarly, participants liked that the baseball cards of their twins were emailed to them, instead of being saved on the system, not only because of the privacy preservation, but also because they could easily share them with friends and family via email.

4.4 Phase IV: Conducting Usability Testing

After synthesizing the findings from Phase III, we found that users' perceived learning about the strengths and limitations of facial recognition and Civil War history improved after using CWT and that our design successfully met users' ethical standards. The last design phase involved getting the CWT website ready for a public release. The previous phases were primarily focused on establishing design goals and iterating on feature designs subsequently phase 4 looks at addressing usability issues on the website to create a seamless user experience. First, we made the website compatible with mobile devices and tablets, to maintain consistency with other facial recognition applications in the field [18, 54]. We also expected a majority of users would access the website with a mobile device so creating a personalized mobile experience was pertinent. After the mobile development, we conducted a usability

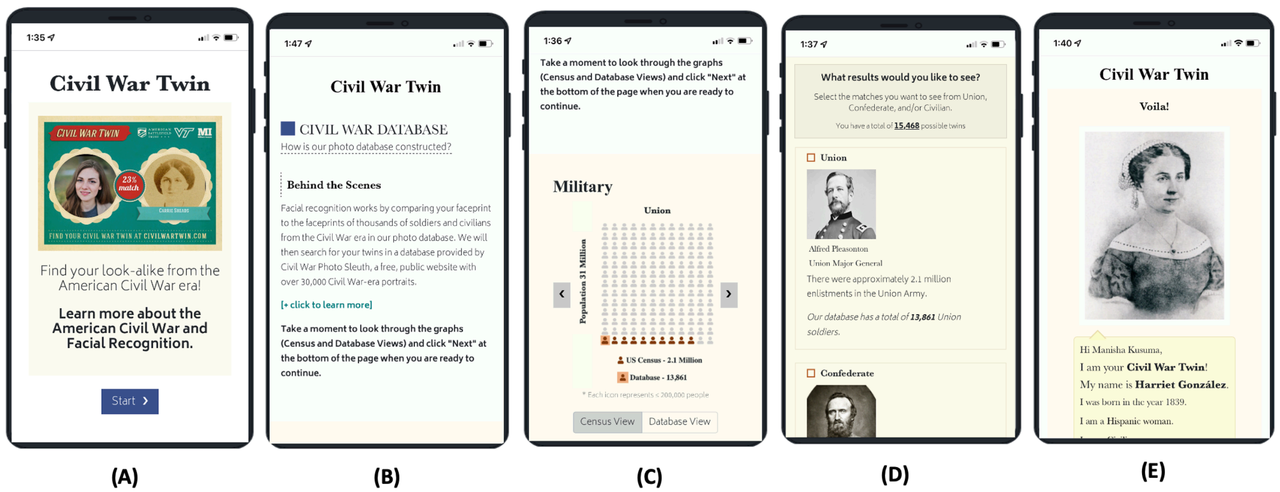

Figure 4.4: Mobile version of the Civil War Twin website.

study with 60 crowd workers on Amazon Mechanical Turk. Based on the findings of the usability study we made additional changes to the CWT website and publicly launched the platform.

4.4.1 Mobile Development

To make the desktop version of CWT compatible with mobile and tablet devices we looked at how we could streamline the process on smaller screen sizes. We consider how the dual learning goals of facial recognition and Civil War history could translate to the smaller screen size. The first major change was shifting from a dual column layout with the AI Text on the left panel, to a single column layout with the AI text followed by the interactive user component. The second change involved condensing the AI Text and historical resources so the user can now click to view more information if they like (see B in Figure 4.4). We still wanted to retain the learning goals of our system while not increasing a mobile user's cognitive load so we reduced the amount of text present and prompt engagement for further learning. The third change to simplify the process on a mobile device was to remove the real-time interactive visualization that shows how a user's preference affects the search pool (see D in Figure 4.4). Since we still wanted mobile users to understand how the pool of potential twins is affected by their choice in preferences we added a counter to show mobile users the number of possible twins in place of the interactive visualizations. The fourth change was to show one graph at a time in the Demographics Overview so the mobile user can choose to flip through the visuals if they want to learn more about the Census (see C in Figure 4.4). Finally, we made aesthetic changes to adapt the design template for the tablet and mobile screens (see A and E in Figure 4.4) and made sure users could directly take a selfie on their photo to initiate the process of discovering twins.

4.4.2 Crowd Feedback

After completing the mobile development, we tested the performance and usability of the CWT website by creating a study on Amazon Mechanical Turk. The goal of the study was to determine any additional bugs or issues with the application and ensure it holds up to usability standards.

Method

Participants We recruited 60 crowd workers on Amazon Mechanical Turk for the study. All crowd workers were located in the United States with a HIT approval rate of greater than 90%. The crowd workers consented to the IRB form.

Procedure For the study, we created two different tasks of Amazon Mechanical Turk; one for testing the desktop version, and the other for testing the mobile version of the website. Both tasks were set up similarly in that crowd workers were first asked to access the CWT website and find their lookalikes then answer questions in a post-survey. For the mobile task, crowd workers used the sample photos feature instead of uploading a personal photo to protect the anonymity of the workers. Similarly for the desktop task, crowd workers used Wikimedia Commons to upload a stock image on CWT. After viewing their twin results crowd workers could access the post-survey. To ensure the workers used the CWT website, we added an attention check to the survey asking crowd workers to list their twin matches. The survey contained mainly Likert scale questions (on a scale of Strongly Agree to Strongly Disagree) about the crowd workers' experience on the website. The questions focused on 4 major usability themes learnability, ease of use, consistency, and satisfaction. The survey also included three open-ended responses about features the crowd workers liked and disliked and any error they came across. The tasks took approximately 30 minutes and we have 30 crowd workers per task.

Data Analysis We analyzed quantitative data (system logs, Likert-scale survey responses, user clicks) using statistical software such Mouseflow and Qualtrics.

Findings

Through analyzing the survey responses we synthesized the main findings and issues that need to be addressed. Overall, the workers liked the explanations about the algorithm's limitations, being able to select search preferences, and learning about their twins. The crowd workers felt the "Behind the Scenes" and "What Could Go Wrong?" text was clear and easy to understand (Mean=1.20, SD=0.98). The crowd workers also took additional steps by clicking to read more information in the AI text and some even visited CWPS. The workers identified that uploading their photos (Mean=1.62, SD=0.71) and selecting their search preferences (Mean=1.53, SD=0.69) was an easy process. From the logs, we can

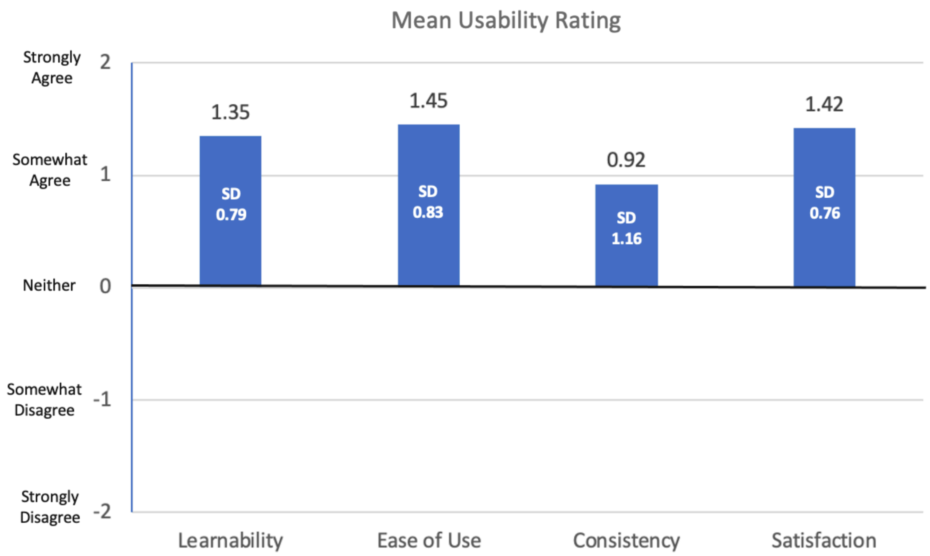

Figure 4.5: Civil War Twin usability survey overview.

see workers were clicking through the database charts and the workers generally found the Database page clear and easy to understand (Mean=1.47, SD=0.67). Based on the responses to the Likert scale questions (see Figure 4.5) we did not identify specific usability issues.

From the open-end questions on the post-survey, we were able to capture more detailed thoughts. Similar to the prospective user findings, the crowd workers wanted more options for search preferences and more photos in the database for the minority groups. Despite our prior database diversification efforts, the limitation of our historical dataset continues to limit user experience. Other crowd workers mentioned that the design and color pallet was dull and could be improved. This was a conscious decision to not use a flashy color palette because of the seriousness and sensitivity around the topics of facial recognition and Civil War history. Based on the open-end responses we added two main changes:

Add Photo Requirements Some workers had issues with a face being detected in the photo they uploaded and had to upload again. To provide more detail during the uploading step we added more information about the photo requirements such as size and resolution.

Clarify Search Preferences Additionally, some workers expressed confusion while selecting their preferences. Their main concern was whether or not it was required to select a preference for each category (e.g., military, gender, ethnicity) since the website allowed the workers to continue without selecting any preferences. To provide more consistency when users are selecting search preferences, we change the workflow to not allow users to continue to the "Gender" category until they selected at least one preference for "Military". Similarly, users cannot continue to the "Ethnicity" category until they selected at least one preference for "Gender".

4.4.3 Release

The website was officially released on November 26th, 2021. After the first two weeks we analyzed system logs and analytics to capture initial usage trends. We had over 1, 500 users in the first two weeks. Users primarily used mobile devices to access the CWT website 71% with laptop being the next highest 25% then tablet 4%. On average users spent 2 minutes and 25 seconds on the website. Users spent the most time on the Uploading a Photo page and Discovering Twins page. When users were discovering their twins, 20% of users choose to view their additional "empathy" twins.

Chapter 5

Discussion

5.1 Ethically Framing Facial Recognition Applications

Facial recognition applications are generally susceptible to ethical challenges related to privacy, gender and racial bias, and accuracy. In this paper, we employed a four-phase ethical design process to iteratively address these challenges. The feedback at each phase from a variety of internal and external stakeholders helped inform the design decisions applied to the system. In Phase I, we reviewed prior literature and industry guidelines to establish design goals from existing guidelines for AI ethics to our application. This first phase allowed us to be cognizant of the limitations pointed out by other applications, best practices and lessons learned, and make efforts towards minimizing bias and accuracy in AI systems. Additionally, we centered design goals around transparency, inclusivity, and privacy to address the three main limitations of the CWT prototype. These goals helped frame the values of the new system.

In Phase II, we consulted experts from the fields of history, gender, and race which allowed us to get specific feedback for the system compared to the more generalizable AI guidelines. The experts not only validated key Phase I design goals and features, such as the search preferences and AI Text, but also provided feedback on the existing design that led to new features, such as our database visualizations and "empathy twins".

In Phase III, we evaluated the system with potential end-users in order to receive user feedback as part of the design process. By recruiting a demographically representative set of users, we were able to understand their experience on CWT and whether the system was effective in teaching them about face recognition and Civil War history. This phase was also an opportunity to understand the user's ethical values and determine if those values were met in the system design. Based on these findings, the participants largely validated the design decisions made in Phase I and II, so there were no major design iterations in Phase III.

In Phase IV, we considered more minor design iterations around usability and responsive design. By recruiting crowd workers we were able to get a larger set of users testing the CWT system and identify any areas of confusion. Through the usability study we were able to measure the satisfaction, consistency, ease of use and learnability of the website. The crowd workers and prospective users from Phase III shared similar sentiments about enjoying the process of selecting search preferences and learning about their twins.

Phase I is consistent with existing research on synthesizing human-AI guidelines for designing specific AI applications [62], but our experience adapting these guidelines for an educational face recognition application illuminated some of the unique ethical challenges of this technology. Future facial recognition applications can similarly reflect on how transparency, fairness, inclusivity, and privacy play a role in their system. Expert feedback as in Phase II can provide areas of improvement and specialized design goals for features that might have been overlooked. Consulting, specifically, with a race and gender studies expert during Phase II provided insight into the current discourse around inclusivity and the roles of race and gender in technology [25]. Through this design process, we noticed that the feedback for industry standards, academic experts, and users were complementary, building on existing themes of privacy, inclusivity, and transparency. For example, the goals of inclusivity propagated through each phase. In Phase I, we made targeted efforts to increase diversity in our database, followed by the addition of the Demographics Overview in Phase II highlighted the minority representation in our database. Finally, in Phase III, participants recognized the system's efforts towards maintaining an inclusive database despite historical limitations and stated learning more about different cultural groups.

5.2 Fostering General Awareness of AI

A 2019 Pew survey found that a majority of Americans were aware of facial recognition technologies being used by law enforcement, advertising, and technology companies, but believed the technology to be highly accurate [58]. However, as with any AI system, it is important to consider *for whom* facial recognition technology is accurate. Buolamwini et al. analyzed several commercial facial recognition systems and showed that the technology works best for white or "lighter-skinned" males [12]. We designed the website to raise awareness about the limitations of the AI and broaden laypeople's ideas about the capabilities and drawbacks of the technology through a personal and playful experience. The CWT system workflow allowed users to interact directly with facial recognition technology, while simultaneously being informed about how the technology works and where it fails. We found that our study participants from diverse demographics enjoyed exploring the technology by using the website multiple times, either uploading new photos or trying out different search preferences, to notice how their twin results differed. Participants were also quick to observe multiple physical similarities with their twins despite the generally low "confidence score" presented ($< 50\%$) for each twin result. Cognitive science theories offer insight that the presence of a unique "high-diagnostic" feature in two different objects increases their relative similarity [64], and therefore might have caught the attention of our participants. Although participants enjoyed finding physical similarities, they understood the context of the system afforded a fun experience with low stakes. Similar to other recent surveys [29, 49], some participants voiced they were not supportive of this technology being used in other contexts such as the criminal justice system.

Our findings from Phase III showed that the supporting speculative AI Text, along with the interactive workflow, helped participants in learning more about how facial recognition works and where it fails. The questions posed in the AI Text encouraged users to speculate about how facial recognition is currently being used, and how it might be used in the future, similar to other recent work on speculative design in AI [4, 68]. This work also shows how XAI practices can be applied at the interface level, specifically in the context of facial recognition, by providing "human-consumable explanations" of AI models [35].

5.3 Interactive Digital Humanities

All participants from Phase III cited their first introduction to the Civil War was in a classroom setting. Often the approach standardized in the K-12 school system is to focus on events, historical figures, and places, taught via the "pipeline model" of lectures transmitting facts, rather than more active or constructivist learning experiences. With CWT, we wanted to present history in a more interactive medium. With visualizations, we were able to provide participants with an overview of key information about the 1860s US Census. By exploring these charts, participants were surprised to learn about various demographics and who participated in the war. Also, when selecting search preferences, users were exposed to photos, biographical profiles, and anecdotes of historical individuals representing diverse cultural perspectives and experiences. As a result many participants were surprised to learn, for example, that Native Americans, Hispanics, Asians, and women participated in the war as soldiers. This work aligns with a multicultural curricular approach that steers away from presenting only one mainstream view of American history to instead focus on the experiences and perspectives of multiple cultural groups [8].

Given the unique nature of the CWT database with stories of thousands of people during the 1860s, we leveraged the idea of microhistories to tell human stories. Participants from Phase III were able to learn about the war through the people who contributed to the effort. We were surprised to see the openness that participants had when selecting search preferences. A majority of the participants did not select any search preferences and were open to being matched with people outside their own identity fostering a notion of empathy. Empathy can be conceptualized in two different ways: putting oneself in the other's shoes and building a shared perspective [51]. Participants learned through the biographical text about the lives of civilians and soldiers living in this critical period in American history. Through the images and text, similarities were found (e.g., geographic location, ethnicity, gender, family history) to help form a bond between the twin and the user. The application of microhistories helps establish empathy as it frames large historical events into lived experiences. This concept is similar to other projects exploring how empathy can be induced through the use of AI [70].

Chapter 6

Limitations

One limitation of our user study, in Phase II, was the lack of diversity in our pool of participants. Given that we had an under-representation of women and nonbinary participants, as well as limited ethnic diversity, we have limited data on how CWT is going to be perceived by marginalized groups. A possible reason for the lack of diversity could be that we specifically recruited participants who were interested in US history and had limited experience with facial recognition technology. However, after recruiting nine participants, and upon conducting the interviews, we realized that we had achieved a level of theoretical saturation, meaning the feedback we received was largely consistent with prior participants [23].

Another limitation of our system was that even though we made targeted efforts towards enriching CWT's database to be more diverse in Phase I, we were bottlenecked by the historical biases. As a result, we do not have any photos available for certain search preferences such as Confederate-Men-African American [34], Civilian-Women-Asian, etc. As this is an ongoing process, the website also ask users to contribute Civil War portraits from their own collections to help support our efforts towards a more inclusive dataset.

Chapter 7

Conclusion

Civil War Twin is an educational web-application where users can discover their lookalike from the American Civil War while learning about facial recognition and Civil War history. We presented a three-phase ethical design process that documented how we synthesized industry guidelines, consulted with different academic experts (in history, gender, and race), and collected user feedback for validating our design choices. We found that our system met the ethical stands of users and provided them an opportunity to learn about the strengths and limitations of facial recognition technology along information about Civil War history. CWT's workflow allowed users to directly interact with facial recognition technology, while the supporting AI text encouraged them to speculate about the implications of facial recognition. Users further displayed a level of empathy for their twins and were keen on learning more about the experience of their twins during the Civil War. Our work opens the doors for research on designing ethical facial recognition applications, while demonstrating how topics such as AI and history can be incorporated together into an interactive educational experience.

Bibliography

[1] Amazon. Amazon Rekognition https://aws.amazon.com/rekognition/?blog-cards.sort-by=item.additionalFields.createdDate&blog-cards.sort-order=desc, 2016.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[3] Apple. About Face ID advanced technology https://support.apple.com/en-us/HT208108, 2021.

[4] James Auger. Living with robots: A speculative design approach. *Journal of Human-Robot Interaction*, 3(1):20–42, 2014.

[5] Brooke Auxier and Lee Rainie. Key takeaways on Americans' views about privacy, surveillance and data-sharing https://www.pewresearch.org/fact-tank/2019/11/15/key-takeaways-on-americans-views-about-privacy-surveillance-and-data-sharing/. *Pew Research Center*, 2019.

[6] Kate Bagnall and Tim Sherratt. The real face of White Australia https://www.realfaceofwhiteaustralia.net/faces/?rsort=1, 2010.

[7] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to

technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 421–433, 2019.

[8] James A Banks. Approaches to multicultural curriculum reform. *Multicultural education: Issues and perspectives*, 2:195–214, 1993.

[9] Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–31, 2021.

[10] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bia s. volume 63, pages 4:1–4:15, 2019.

[11] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[12] 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.

[13] Mike Cason. Archives agency acknowledges distorting racial history, June 2020. URL https://www.al.com/news/2020/06/archives-department-acknowledges-role-in-distorting-alabamas-racial-history.html. Section: News.

[14] Matthew Chalmers, Ian MacColl, and Marek Bell. Seamful design: Showing the seams in wearable computing. 2003.

[15] Ronald S. Coddington. *Faces of the Confederacy an album of Southern soldiers and their stories*. Johns Hopkins University Press, 2008.

[16] Sasha Costanza-Chock. *Design justice: Community-led practices to build the worlds we need*. The MIT Press, 2020.

[17] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. Translation, tracks & data: an algorithmic bias effort in practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[18] Google Arts & Culture. Art Selfie https://artsandculture.google.com/camera/selfie, 2017.

[19] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2021.

[20] Clare Garvie, Alvaro Bedoya, and Jonathan Frankle. The Perpetual Line-Up | Unregulated Police Face Recognition in America https://www.perpetuallineup.org/, 2016.

[21] Kelly A. Gates. *Our Biometric Future: Facial Recognition Technology and the Culture of Surveillance*. NYU Press, New York, January 2011. ISBN 978-0-8147-3210-6.

[22] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. Explainable ai: the new 42? In *International cross-domain conference for machine learning and knowledge extraction*, pages 295–303. Springer, 2018.

[23] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? an experiment with data saturation and variability. *Field methods*, 18(1):59–82, 2006.

[24] Hu Han and Anil K Jain. Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*, 87:27, 2014.

[25] David Hankerson, Andrea R Marshall, Jennifer Booker, Houda El Mimouni, Imani Walker, and Jennifer A Rode. Does technology have race? In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 473–486, 2016.

[26] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 501–512, 2020.

[27] Taylor Hatmaker. Microsoft's facial recognition just got better at identifying people with dark skin, 2018. URL https://social.techcrunch.com/2018/06/26/microsofts-facial-recognition-darker-skin-tones-azure-face-api/.

[28] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[29] Ada Lovelace Institute. Beyond face value: Public attitudes to facial recognition technology. 2019.

[30] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 453–456, 2008.

[31] Rachael Krishna. Asian People Are Not Impressed With Their Matches On Google's Museum Selfie Feature https://www.buzzfeednews.com/article/krishrach/asian-people-are-not-impressed-with-their-matches-googles. *BuzzFeed News*, 2018.

[32] Dovetail Labs. Emojify https://emojify.info/menu, 2018.

[33] Benjamin Charles Germain Lee. Newspaper Navigator https://news-navigator.labs.loc.gov/search, 2020.

[34] Kevin M Levin. *Searching for Black Confederates: The Civil War's Most Persistent Myth*. UNC Press Books, 2019.

[35] Q Vera Liao, Daniel Gruen, and Sarah Miller. Questioning the ai: informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

[36] Di Liu, Randolph G Bias, Matthew Lease, and Rebecca Kuipers. Crowdsourcing for usability testing. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.

[37] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[38] Microsoft. Responsible AI https://www.microsoft.com/en-us/ai/responsible-ai.

[39] Microsoft. Face API - Facial Recognition Software | Microsoft Azure https://azure.microsoft.com/en-us/services/cognitive-services/face/, 2018.

[40] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. Photo sleuth: Combining human expertise and face recognition to identify historical portraits. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 547–557, 2019.

[41] Vikram Mohanty, David Thames, Sneha Mehta, and Kurt Luther. Photo sleuth: Identifying historical portraits with face recognition and crowdsourced human expertise. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–36, 2020.

[42] Design Justice Network. Design Justice Network Principles https://designjustice.org/read-the-principles, 2018.

[43] Daughters of the American Revolution. Who Are You Most Like From the Portrait Collection? https://www.dar.org/museum/education/who-are-you-most-portrait-collection.

[44] Google PAIR. People + AI Guidebook. https://pair.withgoogle.com/guidebook, 2021.

[45] Deborah Raji. How our data encodes systematic racism, 2020.

[46] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 429–435, 2019.

[47] Inioluwa Deborah Raji and Genevieve Fried. About face: A survey of facial recognition evaluation. *arXiv preprint arXiv:2102.00813*, 2021.

[48] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. Saving face: Investigating the ethical concerns of facial recogni-

tion auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 145–151, 2020.

[49] Kay L Ritchie, Charlotte Cartledge, Bethany Growns, An Yan, Yuqing Wang, Kun Guo, Robin SS Kramer, Gary Edmond, Kristy A Martire, Mehera San Roque, et al. Public attitudes towards the use of automatic facial recognition technology in criminal justice systems around the world. *PloS one*, 16(10):e0258241, 2021.

[50] Thomas Robisheaux, Nicole Barnes, Jamal Quick, and Avrati Bhatnagar. What is Microhistory? https://sites.duke.edu/microworldslab/what-is-microhistory/.

[51] Pier Giuseppe Rossi and Laura Fedeli. Empathy, education and ai. *International Journal of Social Robotics*, 7(1):103–109, 2015.

[52] Ruth Sandwell and John Sutton Lutz. *What Has Mystery Got to Do with It?*, pages 23–42. University of Michigan Press, 2014. URL http://www.jstor.org/stable/j.ctv65swr0.5.

[53] Roger C Schank, Tamara R Berman, and Kimberli A Macpherson. Learning by doing. *Instructional-design theories and models: A new paradigm of instructional theory*, 2(2): 161–181, 1999.

[54] Tijmen Schep. How Normal Am I? https://www.hownormalami.eu/, 2020.

[55] Tijmen Schep. Are You You? https://www.areyouyou.eu/, 2021.

[56] Morgan Klaus Scheuerman, Jacob M Paul, and Jed R Brubaker. How computers see gender: An evaluation of gender classification in commercial facial analysis services. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–33, 2019.

[57] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020.

[58] Aaron Smith. More Than Half of U.s. Adults Trust Law Enforcement to Use Facial Recognition Responsibly https://www.pewresearch.org/internet/2019/09/05/more-than-half-of-u-s-adults-trust-law-enforcement-to-use-facial-recognition-respon *Pew Research Center Internet & Technology*, 2019.

[59] PAI Staff. Bringing Facial Recognition Systems to Light https://partnershiponai.org/paper/facial-recognition-systems/, 2020.

[60] Dmitry Statsenko. Star By Face https://starbyface.com/, 2018.

[61] Francesca Street. How facial recognition is taking over airports https://www.cnn.com/travel/article/airports-facial-recognition/index.html. *CNN travel*, 2019.

[62] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. Protoai: Model-informed prototyping for ai-powered interfaces. In *26th International Conference on Intelligent User Interfaces*, pages 48–58, 2021.

[63] Hariharan Subramonyam, Colleen Seifert, and Eytan Adar. Towards a process model for co-creating ai experiences. *arXiv preprint arXiv:2104.07595*, 2021.

[64] Amos Tversky. Features of similarity. *Psychological review*, 84(4):327, 1977.

[65] Steven Umbrello and Ibo van de Poel. Mapping value sensitive design onto ai for social good principles. *AI and Ethics*, pages 1–14, 2021.

[66] Kyle Wiggers. Bias persists in face detection systems from Amazon, Microsoft, and Google. *VentureBeat*, September 2021. URL https://venturebeat.com/2021/09/03/bias-persists-in-face-detection-systems-from-amazon-microsoft-and-google/.

[67] Richmond Y Wong and Vera Khovanskaya. Speculative design in hci: from corporate imaginations to critical orientations. In *New Directions in Third Wave Human-Computer Interaction: Volume 2-Methodologies*, pages 175–202. Springer, 2018.

[68] Niels Wouters, Ryan Kelly, Eduardo Velloso, Katrin Wolf, Hasan Shahid Ferdous, Joshua Newn, Zaher Joukhadar, and Frank Vetere. Biometric mirror: Exploring ethical opinions towards facial analysis and automated decision-making. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 447–461, 2019.

[69] Austin P Wright, Zijie J Wang, Haekyu Park, Grace Guo, Fabian Sperrle, Mennatallah El-Assady, Alex Endert, Daniel Keim, and Duen Horng Chau. A comparative analysis of industry human-ai interaction guidelines. *arXiv preprint arXiv:2010.11761*, 2020.

[70] Pinar Yanardag and Iyad Rahwan. Deep Empathy https://deepempathy.mit.edu/, 2017.

[71] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. Value-sensitive algorithm design: Method, case study, and lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

Appendices

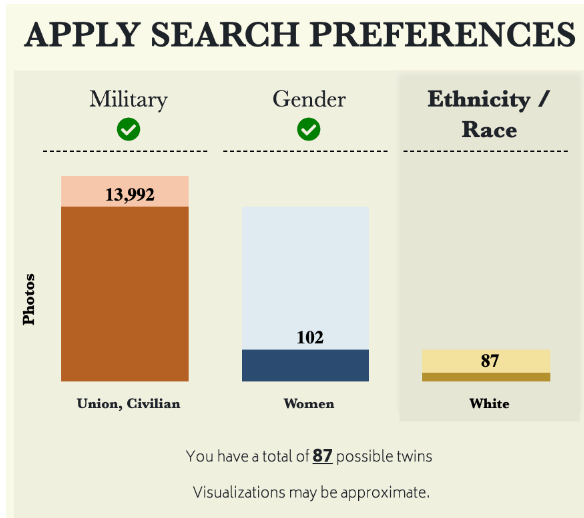

Figure 1: Graphical visualization of the database when the user selects the search preferences: Union, Civilian, Women and White.

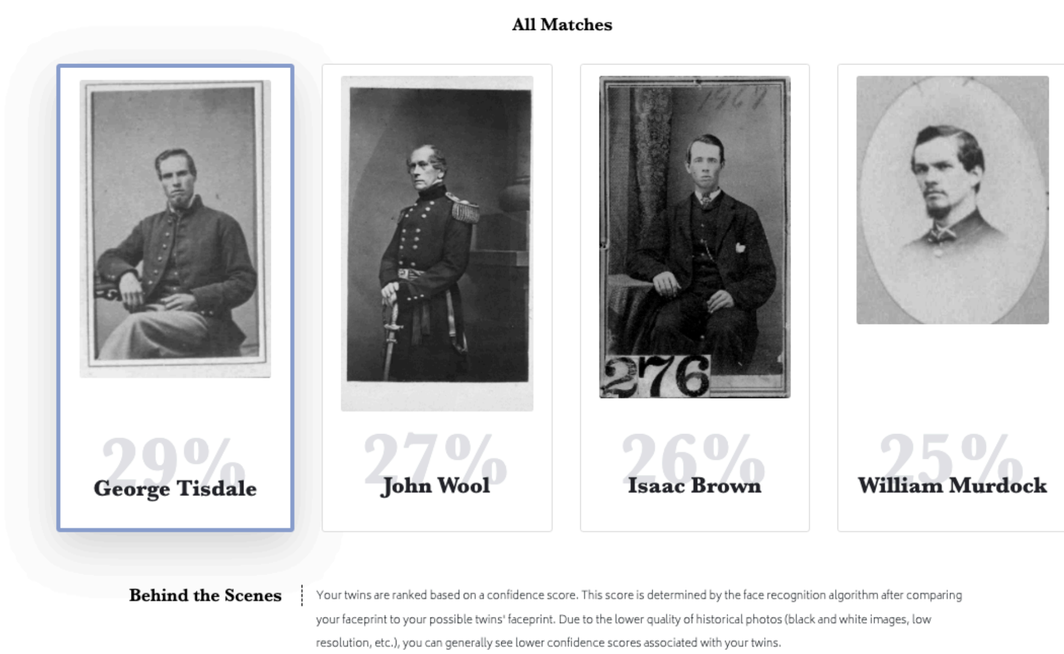

Figure 2: Screenshot of the "Discovering Twins" page with the four twin matches.

Figure 3: The baseball card created for each twin result.

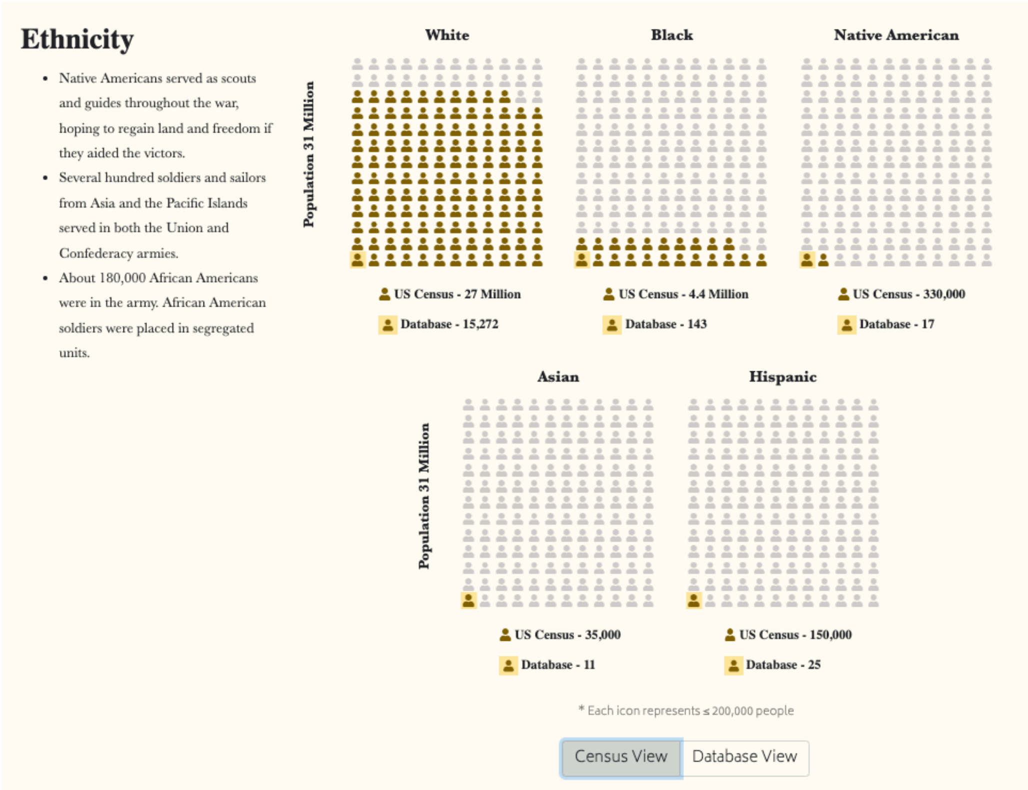

Figure 4: Screenshot of the "Demographics Overview" page with the Census visualizations of the five different ethnicities.

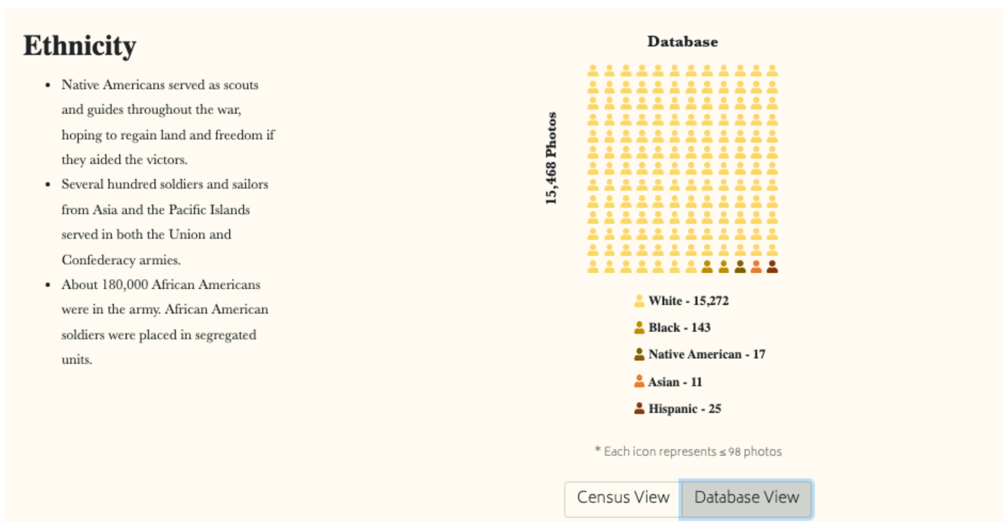

Figure 5: Screenshot of the ”Demographics Overview” page with the Database visualizations of the five different ethnicities.