Using ChatGPT to Generate Human-Value User Stories as Inspirational Triggers

Agnieszka Marczak-Czajka
Computer Science and Engineering
University of Notre Dame
South Bend, USA
amarczak@nd.edu

Jane Cleland-Huang
Computer Science and Engineering
University of Notre Dame
South Bend, USA
JaneClelandHuang@nd.edu

Abstract—Recent work has recognized the importance of developing and deploying software systems that reflect human values and has explored different approaches for eliciting these values from stakeholders. However, prior studies have also shown that it can be challenging for stakeholders to specify a diverse set of product-related human values. In this paper we therefore explore the use of ChatGPT for generating user stories that describe candidate human values. These generated stories provide inspiration to stakeholder discussions and enrich the human-created user stories. We engineer a series of ChatGPT prompts to retrieve a list of common stakeholders and candidate features for a targeted product, and then, for each pairwise combination of role and feature, and for each individual Schwartz value, we issue an additional prompt to generate a candidate user story reflecting that value. We present the candidate user-stories to stakeholders and, as part of a creative requirements engineering session, we ask them to assess and prioritize the generated user-stories, and then use them as inspiration for discussing and specifying their own product-related human values. Through conducting a series of focus groups we compare the human-values created by stakeholders with and without the benefit of the ChatGPT examples. Results are evaluated with respect to coverage of values, clarity of expression, internal completeness, and through feedback from our participants. Results from our analysis show that the ChatGPT-generated user stories are able to provide creativity triggers that help stakeholders to specify human values for a product.

Index Terms—Human values, Creative requirements elicitation, User stories

I. INTRODUCTION

Human values are defined by Schwartz as 'guiding principles of what people find important in life' [1]. Embedding them into software systems allows an organization to deliver products that reflect ethical principles, social norms, or cultural values relevant to the project stakeholders, often serving as clear differentiators from competitors' products. The challenge for organizations and groups developing software systems is, therefore, to identify relevant human values as part of the requirements elicitation process, to analyze their trade-offs, specify them alongside other requirements, and ultimately weave them into the product throughout the design, implementation, and test phases of a project to deliver a value-enhanced product.

However, eliciting human-values in the early stages of a project and operationalizing them as project-specific features is a well-documented challenge [2]. Several authors have proposed and validated techniques to address this challenge in either the requirements or design phase of a project. For example, Thew and Sutcliffe's Values Based Requirements Engineering (VBRE) [3] addressed values related to politics, culture, and stakeholder conflicts which included human-values. Kheirandish focused on human values during the design process [4], while Harbers et al., [5] proposed 'Value Story Workshops' to elicit value-imbued user stories, which we refer to as HV-Stories throughout the remainder of the paper. Finally, Perera et al., conducted an extensive study that compared a feature-driven approach versus a values-driven approach for eliciting human values [6]. They reported that both approaches were effective, and that eliciting human-values in the early phases of a project empowered stakeholders to consider 'why' features were important so that connections could ultimately be developed between features and human-values throughout the project. All of these techniques and their associated studies, highlighted the importance of considering human values early in the requirements process but also indicated that significant effort is required.

Perera's study also highlighted two additional findings. First, they showed that the majority of stakeholders were able to assign values to given scenarios but that more experienced stakeholders (i.e., those with an average of nine-years of experience in their study) were more capable of identifying 'out-of-the-box' requirements that went beyond the initially provided scenario. Second, they found ample evidence that stakeholders drew examples from real-world applications. Based on these observations, this paper sets out to explore whether ChatGPT can achieve the same goals by generating project-specific user stories imbued with human values. As a generative AI framework, ChatGPT is capable of synthesizing past knowledge to create new ideas [7], [8] and therefore to propose ways that human values might impact a software product. However, we do not intend for the generated stories to replace the need for stakeholders to engage in the requirements processes of elicitation, prioritization, analysis, and negotiation of value-based requirements, but rather, we intend them to provide ideas and inspiration to stakeholders as they write their own human-value stories within a workshop type setting [9].

While various frameworks have been proposed (e.g., [10],

[11], [12], [13], [14], we utilize the Schwartz theory of basic human values as this has been shown to be effective for eliciting value-related requirements in a large body of recent work [6], [15], [16]. Notably, the Schwartz theory has been validated cross-culturally, further strengthening our choice of this framework. Introduced and validated in 1992 across diverse cultural settings, the Schwartz framework identifies ten different categories encompassing 58 distinct human values [17]. These values are grouped along two primary dimensions: self-transcendence versus self-enhancement, and openness to change versus conservation. In the first dimension, *self-transcendence* prioritizes the welfare and interests of others, nature, and the broader universe. It includes values associated with universalism, benevolence, and spirituality, while *self-enhancement* focuses on personal goals and self-enhancement and includes values such as achievement and power. In the second dimension, *openness to change* prioritizes innovation, creativity, and new experiences, and includes values such as self-direction, stimulation, and hedonism, in contrast to *conservatism* which focuses on preserving the status quo and encompasses values such as security, conformity, and tradition.

Recently, there has been a growing interest in exploring the advantages and disadvantages of using ChatGPT in various fields. ChatGPT is an advanced language model created by OpenAI that utilizes cutting-edge deep learning algorithms to produce responses to natural language inputs that are remarkably similar to those of a human. Its learning capability is based on its ability to process vast amounts of data, which enables it to generate responses for a wide range of domains and languages. ChatGPT has been investigated as an assistive tool in medical writing and supporting individuals and communities in making informed decisions about their health [18], [19], [20]. It has also shown advantages in education [21], with provided guidelines on its classroom use [22].

To illustrate the potential for ChatGPT 4.0 (referred to simply as ChatGPT from here on) to support creativity in the requirements process based on human values, rather than to simply regurgitate information about existing products, we issued the prompt depicted in Figure I. The prompt asked ChatGPT to generate human-values associated with a futuristic product that would visualize brainwaves as 3D holograms. ChatGPT generated five potentially relevant values associated with self-direction, stimulation, achievement, hedonism, and universalism, demonstrating its potential for generating ideas that serve as creativity triggers to support requirements elicitation of human values.

Given these results we engineered a series of prompts with the aim of using ChatGPT to generate Schwartz-inspired user-stories that could serve as creativity triggers for specifying product-specific values. We adopt an approach that is akin to creativity triggers when utilizing them. We refer to these as Human Value Stories, or HV-Stories for short. As part of our investigation, we explored several different prompts, and sequences of prompts; however, due to space constraints in this paper we report the approach that we judged to produce the best results based on our own observations guided by

TABLE I: As a proof-of-concept to assess whether ChatGPT 4.0 could reason about Schwartz human values in a futuristic product, we prompted it to output values for a product that generated 3D visualizations from brainwaves. It output creative ideas for associating five Schwartz values with the targeted product.

ChatGPT Prompt	
Imagine that I want to create a new product that reads my brainwaves and generates a 3D visualization in the real world. What are some Schwartz human values related to my new invention?	
ChatGPT Responses	
Self-Direction	Your product could empower individuals to gain a deeper understanding of their own brain activity, enabling self-reflection and personal growth.
Stimulation:	By visualizing brainwaves in 3D, your product could provide users with a unique and stimulating experience, exploring the intricacies of their own mind in a visually captivating way.
Achievement	Your invention could facilitate goal-setting and progress tracking, allowing individuals to monitor their cognitive patterns and work towards optimizing their brain activity.
Hedonism	By creating an immersive and visually appealing 3D visualization of brainwaves, your product could offer users a pleasurable and engaging experience.
Universalism	While not directly tied to your invention, considering the ethical implications and potential societal benefits of your product could align with this value. For example, if your invention has applications in healthcare, it could contribute to understanding and treating various neurological conditions.

the professional experience of both authors in working with stakeholders within agile development environments.

We addressed the following research question: **(RQ): Can the HV-Stories generated for a product by ChatGPT serve as effective creativity triggers in the stakeholder focused requirements elicitation process?**

Notably in this paper we do not propose a new AI algorithm, but instead we explore the idea of using ChatGPT as an AI tool to automatically generate inspirational triggers that support the currently challenging task of eliciting human-values for a given product. The motivation for this study is derived from operational principles of trained generative models, such as ChatGPT. Such models build representations of training data in their latent spaces, which are often sparsely populated by representations of training examples and lack fully discoverable internal structures. The nature of generative

models is that they have the ability to generate countless outputs by traversing the model's latent space, and therefore their synthesized outputs may, or may not, be meaningful, and furthermore, some could even be non-sensical with incorrect, bizarre logic. These are referred to as 'hallucinations'. As a result, ChatGPT's outputs should be reviewed critically as a knowledge source, and the studies we describe in this paper therefore evaluate their quality, relevance, and utility in supporting HV-story writing.

The remainder of the paper is structured as follows. Section II describes the process we followed to generate user stories imbued with human values. Sections III and IVthen describe our evaluation of using the ChatGPT generated HV-Stories as creativity triggers for eliciting and specifying human values. Section V discusses related work, Section VI discusses threats to validity, and Section VII draws conclusions.

II. Generative Process for Human-Value Enhanced User Stories

We engineered a series of three ChatGPT prompts to identify stakeholder roles and key features, and then to generate a specific HV-Story for a given role, feature, and Schwartz value. An alternate approach could start with a list of stakeholder roles and/or key features; however, using ChatGPT to generate them means that the process can be fully automated. Furthermore, we found that the key stakeholder roles matched those that we would have been likely to identify anyway.

We therefore utilized the following series of prompts:

1) 'What are [N] common stakeholders for [SYSTEM]'
2) 'What are [N] important features for [SYSTEM]'.
3) 'What is a user story for a <u>customer</u> that reflects Schwartz's basic human value of <u>benevolence</u> for [SYSTEM FUNCTIONALITY] in the [SYSTEM]'

where [N] specifies the number. Issuing this series of prompts for an Airline Reservation SYSTEM, with flight search as the SYSTEM FUNCTIONALITY, returned the following HV-Story including a rationale for why the user story supports the stated Schwartz human value of benevolence:

"As a customer, I want to be able to easily select and book a flight for myself and my loved ones, with the option to add special requests such as wheelchair assistance or extra baggage, because *I value being able to care for and accommodate the needs* of those I am traveling with."

However, reissuing the prompt results in alternate stories with nuanced differences. Two alternate stories are shown as follows:

"As a customer of an airline reservation system, I want to be able to filter my flight search results by airlines that prioritize safety, sustainability, and social responsibility, so that I can feel good about supporting a company that *shares my values of benevolence* and *contributes positively to society and the environment*."

"As a customer using the airline reservation system, I want to be able to easily search for flights that align with my values of benevolence, such as choosing airlines that *prioritize environmental sustainability, social responsibility, and charitable giving*. This will allow me to feel good about my travel choices and contribute to causes I care about, while also meeting my practical travel needs."

Finally, we provide a second example using a different configuration. In this case the prompt was configured with the airline management role, booking and ticketing feature, and Schwartz's basic human value of achievement:

"As airline management, we understand that our customers value the sense of achievement that comes with successfully booking a flight and securing their travel plans. Therefore, we strive to provide an easy-to-use reservation system on our website that allows travelers to search for and book flights quickly and efficiently. We also recognize the importance of clear communication and transparency in the booking process, so we ensure that customers receive prompt and detailed confirmation of their booking, as well as easy access to their ticket information. By prioritizing these aspects of the booking and ticketing process, we *aim to provide our customers with a positive experience and build their trust in our brand*."

III. Evaluation: Human Created User Stories

Our evaluation was designed to compare the coverage and quality of HV-user stories created with and without creativity triggers generated by ChatGPT. We refer to these groups as CT (with creativity trigger) and WT (without creativity trigger) throughout the remainder of the discussion. The tasks all focused around specifying HV-Stories for *HoloWaves*, a novel product that would be able to read human brainwaves and visualize them in the form of 3D holographs. The first researcher on this paper hosted a separate focus group for each of the two groups (CT and WT), in which participants wrote HV-Stories and evaluated the stories generated by ChatGPT. The focus groups were conducted under IRB (Institutional Review Board) approval from our University.

A. Participants

All recruited participants were recruited from the Computer Science or Computer Engineering program during the Summer of 2023. They included six undergraduates (5 female and 1 male), and five graduate students (1 female and 4 male). We used stratified random selection to randomly assign approximately equal numbers of undergraduates and graduates to each of the two groups ultimately leading to 6 people in the CT group and 5 in WT.

B. Preliminary Activities

Prior to each focus group we sent an email including preliminary information about HoloWaves, Schwartz values, and HV-Stories. Materials included a printed chart of Schwartz'

(a) High level ChatGPT Prompt.　　　　(b) Detailed ChatGPT Prompt that includes role, feature, and Schwartz Value for a targeted system
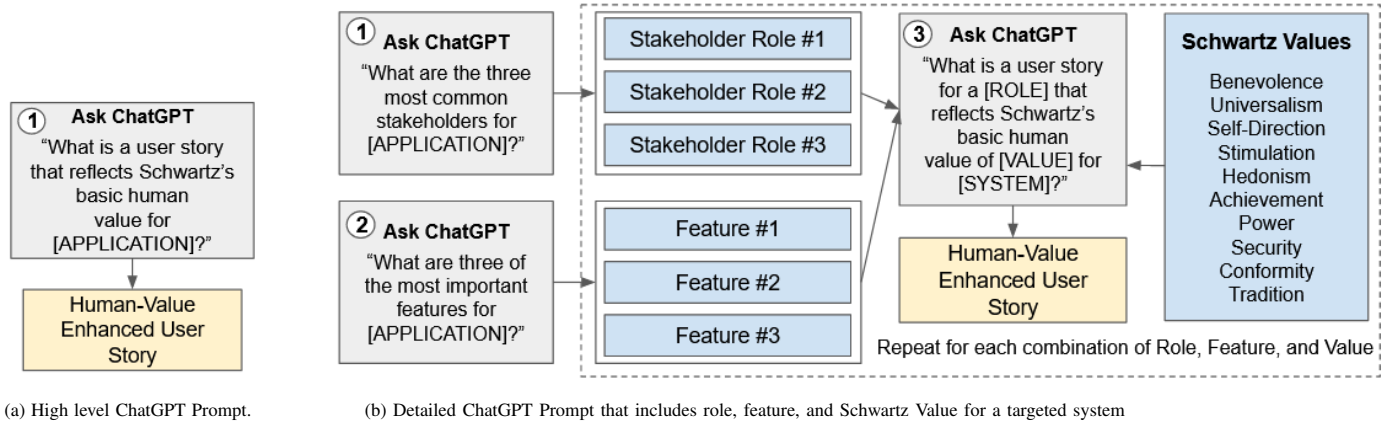
Fig. 1: Two different approaches for utilizing ChatGPT to generate human-value imbued user stories (HV-UserStories)

values with definitions. In addition, the CT group members were asked to assess the HV-Stories generated by ChatGPT, while the WT group were given examples of human-created HV-Stories. The assessment activities are described later in more detail.

C. Focus Group Activities

During each session, we elicited HV-Stories for HoloWaves from the study participants and then facilitated open discussion to explore participants' experiences in the activity. We started by explaining the ideas behind the HoloWaves product as follows: 'Imagine that, based on your qualifications, skills, and potential contributions, you have been selected as a member of a final team responsible for designing an innovative and world-changing system. The product, HoloWaves, that you design reads brainwaves and generates a 3D visualization in the real world. We are starting our design process by creating user stories.'

In the case of the CT group, each participant selected two ChatGPT generated HV-stories that most closely resonated with their own values for HoloWaves and shared these with the group. This activity reinforced the use of the generated user stories as *creativity triggers*. For the WT group we spent time discussing the HV-Stories provided as examples in the initial email. All participants were then asked to write at least three user stories based on the following template: **As a [ROLE], who values [VALUE] I want HoloWaves to [DO X] so that [BENEFIT]**.

We added the phrase 'who values [VALUE]' to the typical user-story template as it can otherwise be hard to infer the user's intent behind a feature.

Furthermore, we suggested that the participants either start with a feature that they deemed to be important, and to consider their personal values associated with that feature; or alternatively, to start with a Schwartz value and consider how it might be realized in the targeted product. These two approaches align well with Parera's findings [2] which showed that either approach was effective for developing requirements

that mapped to human values. Participants were given 6 minutes to write 3 HV-stories.

Following the focus group session, we sent a link to the WT group to ask them to assess the ChatGPT generated HV-Stories. The CT Group had already completed this task prior to the group meeting.

D. Open Discussion

Once HV-stories were written, the first author of this paper led an open discussion which was recorded with permission of the participants (as per the IRB approval process). During the discussion we asked open-ended questions that allowed participants to describe the process they followed to create HV-stories and to describe any challenges they experienced. We encouraged everyone to speak, made sure that the discussion was inclusive and respectful, and used active listening techniques to provide feedback and summarize comments, in order to demonstrate that we were interested in and valued their opinions.

E. Assessment of ChatGPT Generated HV-Stories

The survey done by each subject included twenty ChatGPT-generated user stories, which participants were asked to assess on a 5-point Likert scale to indicate whether they believed that each story captured realistic values of one or more potential stakeholders. This provided an indication of whether each generated user story was realistic. The rubric was structured as: "Strongly agree" (5) , " Somewhat agree" (4) , "Neither agree nor disagree" (3), "Somewhat disagree" (2), and "Strongly disagree"(1), based on the extent to which they believed the stories captured the values of one or more stakeholders.

F. Outcomes

As a result of the focus groups and surveys we collected a total of 35 HV-stories, along with broad feedback and assessment of twenty ChatGPT-generated user stories. We describe the analysis of the data in the next section.

TABLE II: Coverage of Values by HV-Stories in CT (Focus Group With Creativity Trigger) and WT (Focus Group Without Creativity Trigger) groups.

Value	CT	WT
Achievement	3	4
Benevolence	0	2
Conformity	0	0
Tradition	0	2
Hedonism (Entertainment)	2	0
Power	3	2
Security (Privacy)	3	3
Self-direction (Freedom, Creativity)	1	4
Stimulation (Thinking)	3	1
Universalism	1	0
Learning	1	0

IV. COMPARISON OF HV-VALUES GENERATED BY CT VERSUS WT GROUPS

To compare HV-Stories created by the CT group with those created by the WT group, we first assessed the quality and coverage of the HV-Stories using a set of metrics. We then used these metrics to answer our research question.

A. Quantitative and Qualitative Analysis

We selected three qualities to evaluate for the CT and WT created HV-stories. While we could have evaluated the quality of ChatGPT generated HV-Stories too, we chose not to so as not to imply that these stories were intended as end results.

- **Coverage of values:** We assessed coverage of the Schwartz value categories by CT and WT created HV-Stories. It is not necessary for every value to be relevant to every product, and in fact prior studies have shown that users often only specify requirements for a limited group of value; however, it provide insights into the scope of HV-stories produced with and without ChatGPT creativity triggers.

- **Internal Completeness:** Participants were asked to use a template for all their HV-stories; and we therefore evaluated completeness of each individual story according to whether it included a role (As an ...), value (who values ...), functionality (I want a ...), and benefit (so that ...).

- **Internal Clarity:** Finally, we assessed each individual HV-Story for clarity by assessing its writing level using the Grade Level score from the Flesch Reading Ease Formula (named "New GL" on p. 14 in [23]). Since this score makes an attempt to assess the number of years of education needed to understand a given text, the lower the value the greater the internal clarity. For simplicity, this score is denoted later as FREF-GL.

We also collected qualitative feedback from the participants concerning the process of creating HV-stories. This provided insights into the collected data.

B. Coverage of Schwartz Values

We started by analyzing the coverage of Schwartz values for each of the different sets of HV-Stories (i.e., CT versus WT). Focus group participants produced a total of 35 HV-Stories, with the CT group (N=6) creating 17 HV-Stories, and the WT focus group (N=5) creating 18 HV-Stories.

Results and Analysis: The majority of participants used the Schwartz's values provided in the handouts. However, some of them chose to use sub-values included in parentheses in Table II (e.g., Self-determination was expressed as Freedom or Creativity). In addition, one participant expressed a desire for learning and understanding, which did not directly align with a specific value in Schwartz's theory.

Participants from the CT group most frequently selected Achievement, Power, Security, and Stimulation as their preferred values. However, no HV-Stories were created for Benevolence and Tradition, even though there were generated by ChatGPT. For the WT group, the participants predominantly selected Achievement and Self-direction; while not including Hedonism, Universalism, and Learning.

The range of selected values in both groups showed minimal differences in terms of the number of values covered. The CT group (N=6) covered a total of eight values, while the WT group (N=5) covered seven values. Given the unequal number of participants in each group, it is challenging to draw definitive conclusions regarding any differences between them.

C. Analysis of HV-Story Quality

To analyze the quality of HV-Stories, we assessed their internal completeness and calculated sentence complexity using the FREF-GL score.

All user stories from both groups were grammatically correct sentences, and all HV-stories created by both groups included values. However, five out of the eighteen user stories from the WT group, and one out of seventeen from the CT group lacked explicit statements of benefits (e.g., "so that..."). This is illustrated in the following examples:

> "As a user who values stimulation, I want HoloWaves to open new doors for artistic expression."

> "As a user who values security, I want HoloWaves to be safe to use."

This finding aligns with the feedback received from participants, indicating their difficulty in identifying and articulating the benefits associated with creating user stories.

In terms of the process adopted by individual participants, we observed notable differences between the two groups. While both groups were given exactly the same instructions, the CT group incorporated a total of six distinct stakeholders, with the majority adopting perspectives such as "As a user..." or "As a stakeholder..." During focus group discussions, participants reported that they primarily considered the stakeholder from their own perspective, filtering the user stories through their own personal values, as they found this easier than attempting to envision other stakeholder perspectives. In contrast, the WT group employed a broader range of stakeholders, incorporating a total of eleven distinct perspectives. Some stakeholders were directly linked to the participants' personal experiences and identities, such as "As a student ...", "As

a researcher ...", "As a young adult ...", and "As a devout daughter of a traditional family"; however, other stakeholders diverged from the participants' personal perspectives, such as "As a businessman ...", "As a government official ...", and "As a regulator ..."
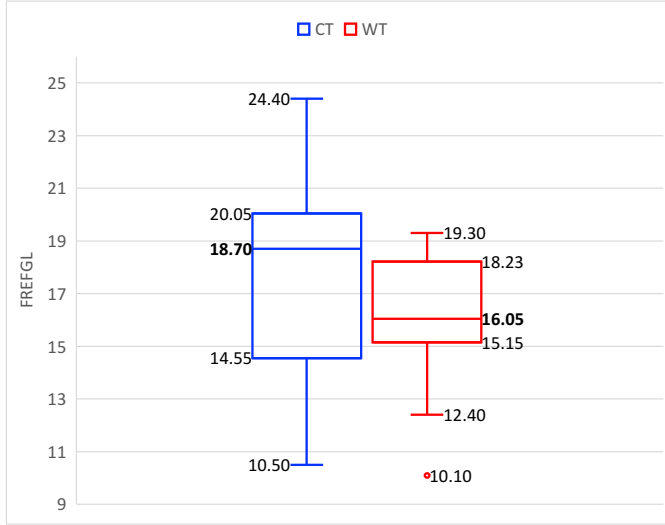

Fig. 2: Boxplots summarizing Flesch Reading Ease Formula - Grade Level score (FREFGL) of user stories created by participants of two focus groups: CT = focus group with creativity trigger, and WT = focus group without creativity trigger. Median values are shown in bold font. Height of each box corresponds to the inter-quartile range (IQR) spanning from the first (Q1) to the third (Q3) quartile (median value included in IQR calculation). Whiskers span from Q1-1.5*IQR to Q3+1.5*IQR. Outliers are shown as circles.

Boxplots summarizing the FREF-GL for both groups are shown in Figure 2. Note that due to the way the FREF-GL formula is constructed, this score has no upper bound, and the lower bound is -3.4. In our experiments this score ranged between approx. 10 and 25. The FREF-GL scores exhibit a wider range in the CT focus group, with the minimum value at 10.50 and maximum value at 24.40. In comparison, the WT focus group has a minimum value of 10.10 and a maximum value of 19.30. The median score for the CT focus group is 18.70, which is higher than the median score for the WT focus group at 16.05. This may suggest that the presence of the creativity trigger resulted in user stories of higher intricacy compared to those prepared by subjects not exposed to creativity triggers. Due to the limited data, it is not possible to determine if the observed differences are statistically significant.

D. Evaluating the ChatGPT generated HV-Stories

In addition to evaluating the HV-Stories created by the CT and WT groups, we also elicited participant's feedback on the relevance and general quality of the ChatGPT-generated HV-Stories. These were the HV-Stories evaluated through the survey given to the CT group before the focus groups and the WT group afterwards. The survey included two ChatGPT generated HV-Stories for each of the ten Schwartz values, resulting in a total of 20 HV-Stories. We used the Qualtrics

survey instrument for participants to rate each HV-Story on a 5-score Likert scale in response to the following question :

To what extent do you believe that the following user stories capture the values of one or more stakeholder?

We collected responses from all participants (6 CT, 5 WT) and calculated the average scores per user story separately for each group, as reported in Figure 3.
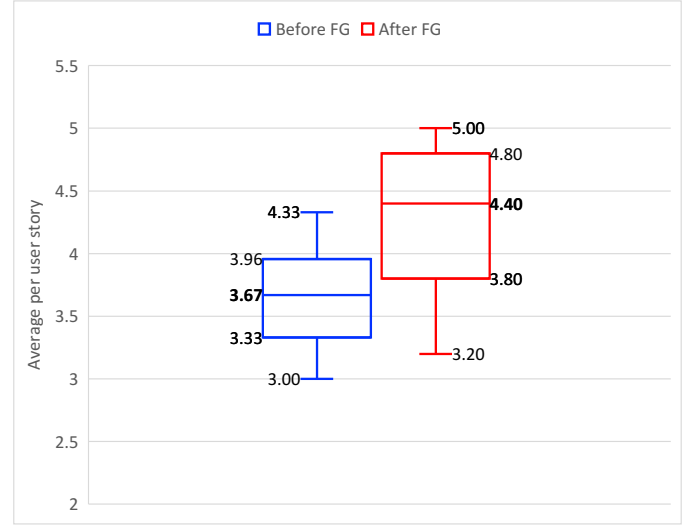

Fig. 3: Boxplots summarize the ranking of ChatGPT-generated user stories by two groups: Before FG = CT, which ranks the user stories before participating in the focus group, and After FG = WT, which ranks the user stories after participating in the focus group. Higher values constitute stronger agreement with the HV-Story. Boxplots properties as in Figure 2.

The participants who completed the survey before attending the CT focus group reported a narrower range of scores, ranging from 3.00 to 4.33. In contrast, the participants who ranked the user stories after participating in the WT focus group received a wider range of scores, ranging from 3.20 to 5.00. The median score for the group that submitted the survey prior to the FG was 3.67, which is lower than the median score for the group that submitted the survey after the FG, at 4.40. This result was unexpected and requires additional thought; however, it may simply mean that after engaging in the task of writing HV-Stories the participants had greater appreciation for the ones generated by ChatGPT. Given the limited amount of data, it is not feasible to establish the statistical significance of these observed differences.

During the CT focus group, participants were asked to select two user stories that resonated with them for various reasons. As a result, seven out of the 20 user stories were chosen as creativity triggers. One user story that focused on security, and stated "As a stakeholder, I need assurance that my brainwave data is encrypted and secure so that my privacy and confidentiality are respected," was selected by three participants. Additionally, three user stories related to hedonism, stimulation, and security were selected twice by independent participants. These user stories are as follows: "As a stakeholder, I wish to enjoy a variety of visual and audio

effects while exploring the 3D visualization so that the process is both informative and entertaining," "As a stakeholder, I want the software to surprise me with new visualization patterns based on my brainwaves so that I am continually stimulated by fresh experiences," and "As a stakeholder, I want to have the ability to delete my data at any point so that I can maintain control over my personal information."

E. Analysis of Collected Feedback from Focus Groups

We structured the feedback using the following question for both groups: How did you identify the four components of the user stories (stakeholder, value, feature, and benefit)? What was difficult, and what was easy?" In addition, for the CT group we asked "How helpful were the selected and highlighted user stories for you?"

In the process of creating user stories, three participants from the WT group started by identifying the stakeholders as the foundation for their stories. One participant drew upon their personal knowledge from an online discussion forum, where they had seen a similar topic discussed, for inspiration in story writing. Finally, one person focused on their own personal values to write their user stories. The CT group found it easy to identify diverse stakeholders; however, they found it difficult to identify and describe benefits of the identified values. One participant specifically found it challenging to describe the functionality of the product due to its general description stating that:

> "I think there's a lot of really important implementation details that we don't have here and that really impact like . I think the user story."

Another participant expressed that connecting the value with the corresponding functionality also required more time and careful consideration.

All participants in the CT group initiated the process of creating user stories by focusing on the values. Some of them selected the same values from the creativity triggers and developed their own features based on those chosen values. For example, one said that

> "the first value was from the highlighted value and then I tried to ask myself why that value"

they then created the benefit by explaining why they appreciated the value.

Two participants reported that it was easy to identify specific features with statements such as "It was easy to tell what I wanted from the system, like obviously, I wanted security and I didn't want my information to be leaked." Another participant agreed, stating, "I found it easier to think about the functionality I wanted from the system."

The remaining three participants all stated that they were able to make connections between the ChatGPT stories and their own values, which facilitated easy identification of stakeholders and creation of the entire user story. They found it harder to think about values from perspectives of different stakeholders, and to identify benefits associated with other stakeholders' values.

The main themes that emerged from the participant responses revolved around the helpfulness of creativity triggers, as exemplified by their positive feedback on the impact of specific examples, clear guidance, and the ability to effectively align values with functionalities during the creation of user stories. Participants expressed how the provided examples stimulated their thinking and inspired them to generate innovative ideas with comments such as:

> "It really guided me it gave me inspiration." .

The guidance offered throughout the process aided participants in understanding the purpose and structure of the user stories:

> "I think it also guides the structure of user stories and, it makes you think why are you using this." .

Also, the ability to align their personal values with the desired functionalities allowed participants to create meaningful and relevant narratives:

> "For me one of them was something that I wanted but I wouldn't have thought of. When I saw it, I was like I want this but I wouldn't have thought of this on my own - so it . gave you some idea."

Additionally, three participants expressed that without the presence of creativity triggers, the process of creating user stories would be more time-consuming and challenging, resulting in the generation of less meaningful narratives.

> "It would be more difficult, demands more time, and more thinking. Maybe you would be misguided to another direction . that doesn't really follow that value selected first."

F. Answering the Research Question

We now analyze the results with respect to the research question **Can the HV-Stories generated for a product by ChatGPT serve as effective creativity triggers in the stakeholder focused requirements elicitation process?**

We conducted an evaluation of the coverage and quality of the final HV-Stories generated by each group, yielding the following conclusions. Firstly, the coverage analysis (cf. II) revealed minimal differences in terms of the number of values addressed by both groups. The CT group incorporated one additional value from the provided list. However, the WT group exhibited a higher frequency of sub-values and expressed a wider range of stakeholder perspectives.

Secondly, the assessment of internal completeness and internal clarity indicates that the inclusion of creativity triggers in the requirement elicitation process may influence the quality of the generated user stories in a sense of observing more complex verbiage. Namely, users exposed to creativity triggers generated stories that obtained higher Flesch Reading Ease Formula Grade Level scores, meaning that they were more complex, and require higher educational levels to comprehend. ChatGPT tends to be rather verbose, and therefore the participants were exposed to more verbose examples. On the other

hand, the templated HV-stories required four distinct parts, which inherently does create a fairly complex sentence to start with.

Furthermore, the feedback received and the analysis of the user stories suggest that the implementation of creativity triggers guided the participants towards identifying their own values associated with the system's features. Moreover, the creativity triggers provided inspiration and facilitated the discovery of previously unconsidered features by the participants.

In conclusion, this study offers compelling evidence that user stories generated by ChatGPT can effectively facilitate the elicitation and specification of well-structured and meaningful HV-Stories for the participants. However, we did not observe a significant impact on the number and diversity of elicited values.

V. RELATED WORK

Research about human values in software engineering seeks to identify and incorporate the values that are important to stakeholders, including users, developers, and other members of society, to create more inclusive and responsible software products. While social science scholars have debated human values for centuries, their implementation in software engineering is a recent area of investigation. It is only in the past few decades that researchers have shown increasing interest in exploring the presence or absence of human values in publications [16], mobile apps [24], [25], [15], [26], health [27] and legal regulations [28].

Some of the recent research in this area has identified various human values that are relevant to software development, including social responsibility, fairness, privacy, security, transparency, and accountability, among others [29], [30]. Researchers have also explored how these values are impacted by different factors, such as cultural and organizational contexts [31], stakeholder expectations, and legal and regulatory requirements [32].

A large number of studies related to human values in software engineering (SE) is based on Schwartz's theory of basic values introduced in 1992 [17]. After refining in 2012, this theory recognized 19 human values measured by 58 value items [33]. The value items are organized in a circular motivational structure. A small distance between values translates to their complemantarity, and large distance between values suggests tension or conflict among them. The Schwartz's theory is considered as the most prominent one [34], [25] and characterized by its cultural universality [31], [25], [35], [28], [29]. The quantified relationship between different values is the greatest contribution of the Schwartz's theory [36].

However, studies in Human-Centered Interaction, which focus on introducing human values from indirect stakeholders into design processes, use different value theories, not exclusively the Schwartz's basic values. For example, Value Sensitive Design (VSD), which is "a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" [37]. It provides a list of 13 human values:

human welfare, ownership and property, privacy, freedom from bias, universal usability, trust, autonomy, informed consent, accountability, courtesy, identity, calmness, and environmental sustainability. This list of values has an ethical import that is often implicated in system design. VSD is characterized by proactivity, that is, the introduction of human value into the design process early and respecting them throughout the life circle [37], [38].

In some research, the VSD is commented as presenting incomplete list of values [39]. For this reason, in recent studies this approach is rejected [40], [4]. In order to create a final list of values for design process the researchers followed meta-inventory of human values [41] and reviewed the existing 13 human values theories described over the last century. As a final results, they identified 45 values organised in 9 clusters (carefulness, justice, ecology, respect for others, meaningfulness, status, pleasure, respect for oneself, and personal development) represented on a value wheel called HuValue. HuValue is a tool created to educate design students and facilitate the process of introducing human values into design process. Also the complexity of use of the compilation of empirical and design-based approaches resulted in creating a tool with a holistic view of values including different aspects of human life. This tool contains four sets of cards (values, activities, personas, and products) what makes it easy and intuitive to use.

The main goal for HuValue design was to create the most inclusive and representative list of values. Inclusiveness is also important for Value-led participatory design [42], [43] and Value-centred design [38]. However, both approaches did not use any list of values and elicit the values from stakeholder based on participatory design approach.

Finally, Values Q-Sort is a tool designed to study human values in the SE field [44], [45], [46]. Based on the mapping of the third draft of the AMC Code Ethics to Schwartz's basic values theory, the authors designed a tool dedicated to SE community. Values Q-Sort allows for extraction of human values from universal level (L1). It also allows for collection of semi-structured qualitative data on personal (L2) and instantiation (L3) levels.

Human values are highly sensitive to the context [31]. Depending on the context, an individual can choose different things that play an important role in their life. The context can influence the changes of the rate of values of end-users. Regarding this contextual nature of values, the question has been raised about the use of generic instruments to measure human values in specific context domain [47].

VI. THREATS TO VALIDITY

First, our study represents a preliminary investigation into the use of ChatGPT as creativity triggers for eliciting human values and generating associated HV-Stories. This paper explores the process of generating creativity triggers for only one system; however, while not reported in this paper, we have generated similar creativity triggers for Google Maps and a Patient's medical charts and observed that it is able to

generate meaningful values for almost every combination of feature and Schwartz Value. Second, we conducted our study using students as participants. The use of students is often criticized in Software Engineering studies; however, in this case, our participants were stakeholders of a futuristic product, and students can clearly serve this role. Future studies should include more expert stakeholders such as business analysts, hologram specialists, and pyschologists. Finally, ChatGPT is itself in flux, and prompts that work well one day may return a different set of results another day. For this reason the tooling proposed in the future work section is necessary in order to create harnesses around the prompts to provide more stability in how results are returned and formatted.

VII. CONCLUSION

Generative Artificial Intelligence, based on Large Language Models such as the recently popular GPT-3 and GPT-4, demonstrates increased quality of generated texts in terms of correctness of grammar and language clarity. While ChatGPT is continually improving its technical fidelity, it is particularly skilled at language generation, and therefore generates HV-Stories rather effectively. The underlying models provide a rather convincing simulation of more abstract human-specific concepts, such as candidate human values and their connections with system properties. On the other hand we cannot simply use these generated HV-stories as-is, as they simply demonstrate that it is possible to make a connection between a given Schwartz value and a potential feature in a product – even when that product is only in the imaginary stage. Instead, this paper used the automatically generated HC-Stories as creativity triggers for inspiring stakeholders with ideas in a human-intensive requirements elicitation process, which would ultimately lead to discussion, analysis, and prioritization activities.

Our experimental results show that ChatGPT did inspire the attendees of the CT group creating HV-stories. We found that (a) the coverage of Schwartz' values was similar in both groups, while (b) the quality of user stories, measured by sentence internal completeness and complexity, is potentially higher in the group exposed to ChatGPT-generate creativity triggers. This provides an initial proof-of-concept demonstration of the potential usefulness of such modern generative AI tools for facilitating the process of creating user stories focused on human values and system features.

ACKNOWLEDGMENT

REFERENCES

[1] S. H. Schwartz, "An overview of the schwartz theory of basic values," Online Readings in Psychology and Culture, vol. 2, p. 11, 2012.

[2] W. Hussain, H. Perera, J. Whittle, A. Nurwidyantoro, R. Hoda, R. A. Shams, and G. Oliver, "Human values in software engineering: Contrasting case studies of practice," IEEE Transactions on Software Engineering, vol. 48, no. 5, pp. 1818–1833, 2022.

[3] S. Thew and A. G. Sutcliffe, "Value-based requirements engineering: method and experience," Requir. Eng., vol. 23, no. 4, pp. 443–464, 2018. [Online]. Available: https://doi.org/10.1007/s00766-017-0273-y

[4] S. Kheirandish, M. Funk, S. Wensveen, M. Verkerk, and M. Rauterberg, "HuValue: a tool to support design students in considering human values in their design," International Journal of Technology and Design Education, vol. 30, no. 5, pp. 1015–1041, May 2019. [Online]. Available: https://doi.org/10.1007/s10798-019-09527-3

[5] M. Harbers, C. Detweiler, and M. A. Neerincx, "Embedding stakeholder values in the requirements engineering process," in Requirements Engineering: Foundation for Software Quality - 21st International Working Conference, REFSQ 2015, Essen, Germany, March 23-26, 2015. Proceedings, ser. Lecture Notes in Computer Science, S. A. Fricker and K. Schneider, Eds., vol. 9013. Springer, 2015, pp. 318–332. [Online]. Available: https://doi.org/10.1007/978-3-319-16101-3_23

[6] H. Perera, R. Hoda, R. A. Shams, A. Nurwidyantoro, M. Shahin, W. Hussain, and J. Whittle, "The impact of considering human values during requirements engineering activities," CoRR, vol. abs/2111.15293, 2021. [Online]. Available: https://arxiv.org/abs/2111.15293

[7] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Scao, S. Gugger, and A. Rush, "Transformers: State-of-the-art natural language processing," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 01 2020, pp. 38–45.

[8] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," CoRR, vol. abs/2005.14165, 2020. [Online]. Available: https://arxiv.org/abs/2005.14165

[9] C. Burnay, J. Horkoff, and N. Maiden, "Stimulating stakeholders' imagination: New creativity triggers for eliciting novel requirements," in 2016 IEEE 24th International Requirements Engineering Conference (RE), 09 2016, pp. 36–45.

[10] K. Dahlsgaard, C. Peterson, and M. Seligman, "Shared virtue: The convergence of valued human strengths across culture and history." Review of General Psychology, vol. 9, pp. 203–213, 09 2005.

[11] M. Rokeach, The nature of human values. Free press, 1973.

[12] R. M. Williams Jr, "Individual and group values," The Annals of the American Academy of Political and Social Science, vol. 371, no. 1, pp. 20–37, 1967.

[13] W. A. Scott and R. Scott, "Values and organizations : a study of fraternities and sororities," American Sociological Review, vol. 30, p. 971, 1965.

[14] W. A. Scott, "Cognitive consistency, response reinforcement, and attitude change," Sociometry, vol. 22, no. 3, pp. 219–229, 1959.

[15] A. Nurwidyantoro, M. Shahin, M. R. Chaudron, W. Hussain, R. Shams, H. Perera, G. Oliver, and J. Whittle, "Human values in software development artefacts: A case study on issue discussions in three android applications," Information and Software Technology, vol. 141, p. 106731, Jan. 2022. [Online]. Available: https://doi.org/10.1016/j.infsof.2021.106731

[16] H. Perera, W. Hussain, J. Whittle, A. Nurwidyantoro, D. Mougouei, R. A. Shams, and G. Oliver, "A study on the prevalence of human values in software engineering publications, 2015 – 2018," in Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering. ACM, Jun. 2020. [Online]. Available: https://doi.org/10.1145/3377811.3380393

[17] S. H. Schwartz, "Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries," in Advances in Experimental Social Psychology, ser. Advances in experimental social psychology. Elsevier, 1992, pp. 1–65.

[18] F. C. Kitamura, "ChatGPT is shaping the future of medical writing but still requires human judgment," Radiology, vol. 307, no. 2, Apr. 2023. [Online]. Available: https://doi.org/10.1148/radiol.230171

[19] T. Panch, J. Pearson-Stuttard, F. Greaves, and R. Atun, "Artificial intelligence: opportunities and risks for public health," The Lancet Digital Health, vol. 1, no. 1, pp. e13–e14, May 2019. [Online]. Available: https://doi.org/10.1016/s2589-7500(19)30002-0

[20] O. Baclic, M. Tunis, K. Young, C. Doan, and H. Swerdfeger, "Challenges and opportunities for public health made possible by advances in natural language processing," Canada Communicable Disease Report, pp. 161–168, Jun. 2020. [Online]. Available: https://doi.org/10.14745/ccdr.v46i06a02

[21] X. Zhai, "ChatGPT user experience: Implications for education," SSRN Electronic Journal, 2022. [Online]. Available: https://doi.org/10.2139/ssrn.4312418

[22] E. R. Mollick and L. Mollick, "New modes of learning enabled by AI chatbots: Three methods and assignments," SSRN Electronic Journal, 2022. [Online]. Available: https://doi.org/10.2139/ssrn.4300783

[23] J. P. Kincaid, R. P. Fishburne, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel," U. S. Naval Air Station, Memphis - Millington, TN, Tech. Rep. 8-75, February 1975, Research Branch Report.

[24] H. O. Obie, W. Hussain, X. Xia, J. Grundy, L. Li, B. Turhan, J. Whittle, and M. Shahin, "A first look at human values-violation in app reviews," in 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS). IEEE, May 2021. [Online]. Available: https://doi.org/10.1109/icse-seis52602.2021.00012

[25] C. Li, H. O. Obie, and H. Khalajzadeh, "A first step towards detecting human values-violating defects in android APIs," in 2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW). IEEE, Nov. 2021. [Online]. Available: https://doi.org/10.1109/asew52652.2021.00022

[26] H. O. Obie, I. Ilekura, H. Du, M. Shahin, J. Grundy, L. Li, J. Whittle, and B. Turhan, "On the violation of honesty in mobile apps," in Proceedings of the 19th International Conference on Mining Software Repositories. ACM, May 2022. [Online]. Available: https://doi.org/10.1145/3524842.3527937

[27] A. Sutcliffe, P. Sawyer, W. Liu, and N. Bencomo, "Investigating the potential impact of values on requirements and software engineering," in 2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS). IEEE, May 2021. [Online]. Available: https://doi.org/10.1109/icse-seis52602.2021.00013

[28] H. Perera, "Towards human values traceability in software: A goal modeling approach," in 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE, Sep. 2019.

[29] W. Hussain, M. Shahin, R. Hoda, J. Whittle, H. Perera, A. Nurwidyantoro, R. A. Shams, and G. Oliver, "How can human values be addressed in AgileMethods a case study on SAFe," IEEE Transactions on Software Engineering, pp. 1–1, 2022. [Online]. Available: https://doi.org/10.1109/tse.2022.3140230

[30] Q. Lu, L. Zhu, X. Xu, J. Whittle, and Z. Xing, "Towards a roadmap on software engineering for responsible AI," in Proceedings of the 1st International Conference on AI Engineering: Software Engineering for AI. ACM, May 2022. [Online]. Available: https://doi.org/10.1145/3522664.3528607

[31] H. O. Obie, M. Shahin, J. Grundy, B. Turhan, L. Li, W. Hussain, and J. Whittle, "Does domain change the opinion of individuals on human values? a preliminary investigation on eHealth apps end-users," in 2021 28th Asia-Pacific Software Engineering Conference (APSEC). IEEE, Dec. 2021. [Online]. Available: https://doi.org/10.1109/apsec53868.2021.00063

[32] H. Perera, "Towards human values traceability in software: A goal modeling approach," in 2019 IEEE 27th International Requirements Engineering Conference (RE). IEEE, Sep. 2019. [Online]. Available: https://doi.org/10.1109/re.2019.00060

[33] S. H. Schwartz, J. Cieciuch, M. Vecchione, E. Davidov, R. Fischer, C. Beierlein, A. Ramos, M. Verkasalo, J.-E. Lönnqvist, K. Demirutku, O. Dirilen-Gumus, and M. Konty, "Refining the theory of basic individ-ual values," J. Pers. Soc. Psychol., vol. 103, no. 4, pp. 663–688, Oct. 2012.

[34] J. Whittle, M. A. Ferrario, W. Simm, and W. Hussain, "A case for human values in software engineering," IEEE Software, vol. 38, no. 1, pp. 106–113, Jan. 2021. [Online]. Available: https://doi.org/10.1109/ms.2019.2956701

[35] E. Winter, S. Forshaw, L. Hunt, and M. A. Ferrario, "Advancing the study of human values in software engineering," in 2019 IEEE/ACM 12th International Workshop on Cooperative and Human Aspects of Software Engineering (CHASE). IEEE, May 2019. [Online]. Available: https://doi.org/10.1109/chase.2019.00012

[36] G. R. Maio, "Mental representations of social values," in Advances in Experimental Social Psychology. Elsevier, 2010, pp. 1–43. [Online]. Available: https://doi.org/10.1016/s0065-2601(10)42001-8

[37] B. Friedman, P. H. Kahn, and A. Borning, "Value sensitive design: Theory and methods," UW CSE Technical Report 02-12-01, 2002.

[38] G. Cockton, "A development framework for value-centred design," in CHI '05 Extended Abstracts on Human Factors in Computing Systems. ACM, Apr. 2005. [Online]. Available: https://doi.org/10.1145/1056808.1056899

[39] I. van de Poel, "Design for value change," Ethics and Information Technology, vol. 23, no. 1, pp. 27–31, Jun. 2018. [Online]. Available: https://doi.org/10.1007/s10676-018-9461-9

[40] S. Kheirandish, " HuValue: a tool to enrich design concepts with human values," Ph.D. dissertation, Technische Universiteit Eindhoven, 2018.

[41] A.-S. Cheng and K. R. Fleischmann, "Developing a meta-inventory of human values," Proceedings of the American Society for Information Science and Technology, vol. 47, no. 1, pp. 1–10, Nov. 2010. [Online]. Available: https://doi.org/10.1002/meet.14504701232

[42] O. S. Iversen and T. W. Leong, "Values-led participatory design," in Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design. ACM, Oct. 2012. [Online]. Available: https://doi.org/10.1145/2399016.2399087

[43] O. S. Iversen, T. W. Leong, P. Wright, J. Gregory, and G. Bowker, "Working with human values in design," in Proceedings of the 12th Participatory Design Conference: Exploratory Papers, Workshop Descriptions, Industry Cases - Volume 2. ACM, Aug. 2012. [Online]. Available: https://doi.org/10.1145/2348144.2348191

[44] E. Winter, S. Forshaw, and M. A. Ferrario, "Measuring human values in software engineering," in Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement. ACM, Oct. 2018. [Online]. Available: https://doi.org/10.1145/3239235.3267427

[45] E. Winter, S. Forshaw, L. Hunt, and M. A. Ferrario, "Towards a systematic study of values in SE: Tools for industry and education," in 2019 IEEE/ACM 41st International Conference on Software Engineering: New Ideas and Emerging Results (ICSE-NIER). IEEE, May 2019. [Online]. Available: https://doi.org/10.1109/icse-nier.2019.00024

[46] M. A. Ferrario and E. R. Winter, "Applying human values theory to software engineering practice: Lessons and implications," IEEE Transactions on Software Engineering, pp. 1–1, 2022. [Online]. Available: https://doi.org/10.1109/tse.2022.3170087

[47] H. O. Obie, M. Shahin, J. Grundy, B. Turhan, L. Li, W. Hussain, and J. Whittle, "Does domain change the opinion of individuals on human values? a preliminary investigation on ehealth apps end-users," in 2021 28th Asia-Pacific Software Engineering Conference (APSEC). IEEE, Dec. 2021.