Effectiveness and Information Quality Perception of an AI Model Card: A Study Among Non-Experts

Vanessa Bracamonte
Usable Trust Group
KDDI Research, Inc.
Saitama, Japan
va-bracamonte@kddi-research.jp

Sebastian Pape*, Sascha Löbner† and Frederic Tronnier‡
Chair of Mobile Business & Multilateral Security
Goethe University Frankfurt
Frankfurt, Germany
ORCID: *0000-0002-0893-7856, †0000-0001-9164-1919 and ‡0000-0002-6202-4871

***Abstract*—With the rising popularity of artificial intelligence (AI) applications, the use of the underlying models has spread to the general public. These AI models have limitations and biases, and knowing about their characteristics could promote their safe use. Although there is some information about AI models available, in the form of AI Model Cards, there is little research on how useful this information is for non-expert users. In this paper, we conduct an experiment to evaluate the effectiveness and perception of information quality of the Model Card of a currently available AI and compare it with shorter versions. The results show that participants can use the Model Card to answer questions about the AI, but they are less confident about their answers compared to shorter versions. In addition, the full Model Card is considered less understandable and interpretable compared with a short version. On the other hand, a short version had a negative effect on perceived trustworthiness of the AI, but in all cases the participants had a positive attitude towards seeking information about the AI.**

***Index Terms*—AI, model card, information quality, intention to use, trust, understandability, user study**

I. Introduction

Machine learning and artificial intelligence (AI) are umbrella terms describing one of the recent hot topics which causes a gold rush in several industries. Certainly, ChatGPT – an artificial-intelligence chatbot developed by OpenAI [1] – has strongly contributed to that trend. However, a similar important area is the generation of images from natural language description such as OpenAI's DALL-E 2 [2], Google Brain's Imagen [3] and StabilityAI's Stable Diffusion [4]. As the technology is on the edge from research to becoming business, companies start to get more restrictive about the details, i. e., they do not provide details about the training method and the training dataset anymore [5].

On the other hand, there is also a broad discussion about ethical issues [6] and general shortcomings of the technology and how to deal with it in the future, i. e., if a strong regulation is needed. An important step to address that issue is to make the technology more transparent [7], [8], to allow the users to assess potential risks by their own as it is already common in other domains. For example, medicine contains package inserts with information about potential risks and some guidance how to use it; groceries have a nutrition label with information about the ingredients and nutritional values. There have already been proposals for documentation introduced for artificial intelligence, such as Model cards [9] and FactSheets [10], which contain some information about the characteristics of the AI. While this type of documentation would not improve the AI itself, it would at least provide some transparency about it and could help prevent some problems such as reproducibility problems [11], when the technology is updated with a newer set of training data. It would also allow the public to map certain AI typical issues such as biases [12], potential copyright problems – when the training data included some copyright protected material [13] – or "lying" [14] to all applications making use of the same AI model.

Model Cards [9] are one proposal for AI model documentation which has been adopted to some extent by developers. The provided information not only includes technical aspects such as the performance of an AI model, but also guidance information such as the intended use of the AI and its biases and limitations. Besides providing information, Model Cards could also allow stakeholders to compare different AI models. The stakeholders also include individuals who may not be AI experts but who would be impacted by the AI in some way. For these individuals, the Model Card may be used to understand the AI and use it to "pursue remedies" [9]. Nevertheless, it is not sufficient that information about the model is public, as the example of privacy policies demonstrates [15]. Model Cards in their current version include information that individual non-expert users may find difficult to understand. Non-expert users must be able to understand and interpret the given information, but the perception of these users has not often been investigated. To address this, in this paper, we investigate the effectiveness and perception of information quality of a real AI Model Card among non-practitioners and evaluate how it compares to shorter versions. The results show that participants answered questions about the AI with a similar rate of effectiveness in all Model Card versions. For some questions, this rate was low. On the other hand, they perceived the original, full Model Card as less understandable and interpretable than shorter versions. The results also show that the Model Cards did not have an effect on intention to use the AI. We found a slight although significant decrease of perceived trustworthiness of the AI in participants who viewed a shortest Model Card, but no effect for other versions. In addition, participants in every condition had a positive attitude

with regards to seeking information about the AI.

The remainder of the paper is structured as follows: Sect. II discusses related work. Sect. III introduces the research questions and describes the setup of the experiment. Results are presented in Sect. IV and discussed in Sect. V. Sect. VI concludes the paper.

II. Related Work

To aid in providing transparency about different aspects of an AI model, different formats of documentation have been proposed [9], [10]. User studies have been conducted on different aspects of this type of documentation, in particular Model Cards. However, although the AI documentation proposals make clear that the AI stakeholders include users with different levels of expertise, user studies have most often been conducted among machine learning experts and practitioners. Shen et al. [16] conducted workshops to test a Model Card toolkit to support understanding and choosing a machine learning model. In the study, although they condensed technical information contained in the Model Cards, they report that the terminology in the cards was still difficult for participants to understand. Crisan et al. [17] conducted semi-structured interviews with participants who had used machine learning or NLP model, in which they showed the participants a static model card and their proposed interactive version and asked them their opinion. They found that even participants who had a higher level of knowledge of machine learning reported difficulty in understanding the purpose of the model from the information in the model card. Nunes et al. [18] conducted a qualitative study among developers to analyze how they made use of the Model Card when reflecting on the potential ethical issues related to the AI. They found that the Model Card itself was not enough to encourage deep reflection but did not discard the possibility that the way that ethical considerations are presented could have an impact. Chiang et al. [19] conducted an experiment to test the effect of types of tutorials on machine learning models on non-experts' reliance on these systems, using model cards a type of static tutorial. They report that participants had positive a reaction to being able to gain knowledge about the AI model, including about its limitations, but did not have an effect on reliance. The authors suggest that this is due to the difficulty of interpreting technical information about the performance of the model, which laypersons lack the expertise to do so.

III. Method

A. Research Questions

In this study, we base the design our experiment on existing research on effectiveness and perception of privacy policies [20], another type of documentation that non-expert users depend on. In particular, we are interested in evaluating non-experts opinions towards a real Model Card, for an AI model that is currently in use. We also are interested in evaluating how shorter versions of that Model Card compare to the real Model Card, since previous research indicates that users perceive Model Cards' content to be difficult to understand due to their technical content. With these considerations, we examine the following questions: (1) What is non-experts effectiveness in finding information in a Model Card?, (2) What is non-experts perception of the information quality of a Model Card? (3) Does the Model Card have an effect on non-experts' perception of the AI itself and on their attitude towards seeking information about the AI?, and (4) How do these aspects (1-3) compare when considering shorter versions of the Model Card?

B. Model Card versions

For the study, we used the Stable Diffusion v2 Model Card [21] to create three versions: the original version (Full) and two adapted versions: Medium and Short. The Full version contained all the content from the original Stable Diffusion v2 Model Card, which includes technical information about the training procedures and a plot of evaluation results. For the Medium Model Card version, we excluded content that requires a degree of expertise to understand and interpret, such as information about how the model was trained and about its evaluation results. The Short Model Card version was adapted based on the Medium version. We converted the limitations into bullet points and shortened and simplified the sentences. All versions shared the same visual design.

The readability was calculated using the Flesch Reading Ease and the Flesch-Kincaid Grade Level tests. As established, there were fewer words in the Short (491) and Medium Model Cards (670) than in the Full Model Card (1331), but the level of readability was similar for each of them. According to the Flesch Reading Ease score, the scores slightly increase from short to full (Full: 44.75, Medium: 40.65, Short: 39.42), but all Model Cards versions are difficult to read and correspond to a college reading level. The Flesch-Kincaid Grade Level scores are roughly the same (Full: 11.5, Medium: 11, Short: 11.5) and indicate that the Model Cards require an 11th US reading grade level.

C. Task

Participants were randomly assigned to an experiment condition corresponding to one of the Model Card versions. The task for all conditions consisted of reading a brief description of AI text-to-image generation and viewing some examples of input texts and their corresponding output images. Then, after asking about the opinion towards such an AI, we asked the participants to view the Model Card corresponding to their experiment condition and asked them to answer questions about the AI by referring to the Model Card.

We developed questions about the AI (content questions) to evaluate whether participants could be effective in finding the relevant information in the Model Card. There were 7 content questions, corresponding to each of the sections of the Stable Diffusion v2 Model Card. The questions and answer options are detailed in Table IV. Items for the Understandability, Interpretability, Relevancy, Completeness, Concise Representation and Appropriate Amount dimensions of Information Quality were adapted from [22]. The Information Quality measurement

TABLE I
MEASUREMENT ITEMS RELATED TO THE AI

Intention to use [23]–[26]	If I have access to an AI like this I will use it. I think my interest for an AI like this will increase in the future. I will use an AI like this as much as possible. I will recommend others to use this type of AI. I plan to use an AI like this in the future.
Perceived trustworthiness [27]–[30]	Given the provided information, I trust that the AI makes good-quality results. Based on my understanding of the information I know the AI is trustworthy. I think I can trust the AI. The AI can be trusted to carry out the task faithfully. In my opinion the AI is trustworthy.
Attitude towards seeking information about the AI [31], [32]	Seeking information about AI text-to-image generation is... Worthless ——— Valuable Bad ——————— Good Harmful ————- Beneficial Not helpful ——- Helpful Unproductive —- Productive Not useful ——— Useful

TABLE II
SAMPLE DEMOGRAPHICS BY CONDITION.

		Full (n=45)		Medium (n=44)		Short (n=42)	
Age	Mean (sd)	37.07 (11.46)		35.61	(12.58)	35.36 (10.08)	
	Median	35		32		32	
Gender	Female	24	53.3%	17	38.6%	21	50.0%
	Male	19	42.2%	22	50.0%	21	50.0%
	Non-binary	2	4.4%	4	9.1%	-	-
	Prefer not to say	-	-	1	2.3%	-	-
Education	Bachelor	22	48.9%	16	36.4%	19	45.2%
	College	9	20.0%	17	38.6%	3	7.1%
	High school	4	8.9%	2	4.5%	7	16.7%
	Master	4	8.9%	6	13.6%	7	16.7%
	No school/diploma	2	4.4%	-	-	-	-
	Profess. degree	4	8.9%	3	6.8%	4	9.5%

items had a response scale ranging from 0 ("Not at all") to 10 ("Completely"). As in [22], the middle point of the scale (5) was labeled "Average".

The questionnaire also included items to measure Intention to use (adapted from [23], originally from [24]–[26]) and Perceived trustworthiness (adapted from [27], originally from [28]–[30]) of the AI, before and after viewing the Model Card. These items had a 7-point Likert response scale. Attitude towards Seeking Information was adapted from [31], [32], and consisted of six 7-point semantic differential scales. The measurement items related to AI perception are detailed in Table I. Finally, questions on previous familiarity with AI text-to-image generation and whether the participant held an IT degree were included, as well as demographic questions (age, gender, education). Due to the length of the survey, we included 3 attention questions.

D. Pre-Test

We conducted a pre-test of the survey with 30 Amazon Mechanical Turk workers, to validate the length of the survey (which was calculated at approximately 25 minutes) and the understandability of the survey in general. The mean survey response time for pre-test participants was 19.9 minutes (sd = 8.37) with a median of 17.8 minutes. The participants' feedback indicated that they had a neutral-to-positive opinion of the length and difficulty of the survey. However, the answer options of two content questions were considered ambiguous and were revised for the final version.

E. Participant Recruitment

The participants were recruited on the Prolific platform. Participation was limited to people in the USA whose first language was English. We also set the sample to have an equal number of male and female participants. The reward was set at approximately US$4.97 (£4.00). We initially obtained 170 survey responses. Of these, 4 were not accepted due to failing 2 out of 3 attention questions. The valid responses were completed in an average time of 17.23 minutes (sd = 9.77) with a median of 14.5 minutes. The participants were compensated at an average rate of US$17.31/hr (£13.93/hr). Due to the focus of this study, participants who indicated that they were familiar with AI text-to-image generation for work or research-related reasons (6 participants), and participants who indicated they had an IT degree (32 (19.39%) participants) were not included in the analysis, but were compensated.

F. Ethical Considerations

The current research was exempt from ethical review in accordance with the guidelines of our institutions. The survey included an Informed Consent form with the details about the objective of the study, task and length of the survey, voluntary participation, treatment of the collected data, conditions for compensation and participation, and contact of the principal researcher. Only participants who selected the option to agree to participate proceeded to the survey.

IV. RESULTS

A. Sample Characteristics

The sample for analysis consisted of 131 participant responses. The number of participants per group was 45 in the Full condition, 44 in the Medium condition, and 42 in the Short condition. Gender distribution was 62 (47.33%) female, 62 (47.33%) male, 6 (4.58%) non-binary and 1 participant who preferred not to indicate gender. The mean age of the participants was 36.03 years-old (sd = 11.38), with a median of 33 years-old. The distribution of participants' demographics by condition, including education degree, is shown in Table II. Finally, the distribution of categories of familiarity with AI text-to-image generation by condition is shown in Table III.

B. Effectiveness

First, we compared the time that participants spent in the section of the survey where the Model Card was initially presented and the time they spent in the section where they had to answer the content questions. The times were recorded

TABLE III
CATEGORIES OF FAMILIARITY WITH THE AI BY CONDITION.

	Full	Med.	Short
I am unfamiliar with AI text-to-image generation.	14 (31.1%)	13 (29.5%)	9 (21.4%)
I have never used AI text-to-image generation, but I'm familiar with the concept.	19 (42.2%)	16 (36.4%)	19 (45.2%)
I have used AI text-to-image generation, for fun.	12 (26.7%)	15 (34.1%)	14 (33.3%)

in the survey platform, and we used them as an approximation of how long the participants viewed the Model Card. One-way ANOVA tests indicated that the time spent in the initial Model Card view section was not significantly different between conditions (F = 0.711, p-value = 0.49). The time the participants spent in the content questions section was also not significantly different (F = 2.215, p-value = 0.11).

We then examined how participants answered the content questions in each condition. Kruskall-Wallis tests showed that there were no significant differences between conditions with regards to the number of correct answers (chi-squared = 1.2524, p-value = 0.53), but there was a significant difference in confidence (number of questions that participants thought they answered correctly) (chi-squared = 11.842, p-value = 0.003). The post-hoc Dunn tests indicate that the confidence in the Full and Medium Model Card condition was significantly lower than in the Short condition. Spearman correlation analysis for the relationship between the number of correct answers and confidence showed a significant positive correlation for the Full Model Card, but not for the Medium and Short conditions

With regards to the answers to each question, the results are detailed in Table IV. Participants had similar a rate of effectiveness in finding the right answer in all conditions (which was low for some questions). The exception was the difference in response to Q2 on the performance of the AI, where fewer than 60% of the participants in the Full and Medium Model Cards answered correctly. More participants answered "I don't know" or "The answer is not in the information provided" in these conditions compared to the Short version, indicating the they could not effectively find the information.

Participants had the most difficulty with the Q6 on the type of images used to train the AI. More than 70% of participants in all conditions answered the AI had been trained with all images in the database mentioned in the Model Card. The reason for the results of Q6 can only be partially attributed to the fact that "None of the above" was the correct answer, since the majority of participants in all conditions agreed on a particular (incorrect) answer. We interpret these results to reflect that the description of the data used in training the AI is ambiguous and not clear to non-expert users.

C. Information Quality

Cronbach's alpha values for all the scales indicated good reliability (above 0.7), ranging between 0.76 and 0.92. The measurement items for each variable were therefore averaged.

TABLE IV
RESPONSES TO THE QUESTIONS ABOUT THE MODEL CARD CONTENT (%).

Question	Condition		
Q1. Which use is this AI intended for?	**F**	**M**	**S**
Only for research.	91	91	81
For commercial purposes.	0	2	5
To be used for Open Source projects only	0	0	5
To be used for nonprofit only	0	2	0
None of the above.	2	0	2
I don't know.	7	2	2
The answer is not in the info. provided.	0	2	5
Q2. How does the AI perform in representing the truth about people or events when generating images?	**F**	**M**	**S**
The AI always represents the truth, because it is very accurate.	0	0	0
The AI was trained with real images, so it always represents the truth.	0	2	0
The AI was not trained to represent the truth.	**58**	**52**	71
The AI might not represent the truth because it was trained with fictional images.	13	23	17
None of the above.	2	0	2
I don't know.	11	9	2
The answer is not in the info. provided.	16	14	7
Q3. Are there cases for which AI should not be used?	**F**	**M**	**S**
Should not be used to share copyrighted material without permission.	**69**	73	76
Should not be used to generate real people.	18	16	10
Should not be used to generate images that represent emotion.	0	2	2
Should not be used to generate medical images.	4	0	2
None of the above.	7	9	10
I don't know.	0	0	0
The answer is not in the info. provided.	2	0	0
Q4. What are the limitations of this AI?	**F**	**M**	**S**
The AI has problems with generating text in images, but faces of people are generated properly.	4	2	0
None. The AI always generates faces and people perfectly.	0	0	0
The AI may not generate faces and people correctly, and cannot generate readable text.	78	91	90
The AI can always generate faces and people correctly, but has difficulty with position of objects.	2	0	2
None of the above.	7	2	5
I don't know.	7	5	0
The answer is not in the info. provided.	2	0	2
Q5. Which of these statements is true about the images generated by the AI?	**F**	**M**	**S**
The AI generates images of all types of communities and cultures equally.	2	2	5
The AI has the same ability to generate images with prompts in any language.	2	2	0
The AI generates images that are always free of social biases.	2	9	2
The AI generates images of some types of communities and cultures less frequently.	**69**	**57**	**69**
None of the above.	16	11	14
I don't know.	4	14	7
The answer is not in the info. provided.	4	5	2
Q6. What images were used to train this AI?	**F**	**M**	**S**
Images with descriptions in all languages.	2	2	0
All images contained in the LAION-5B dataset.	71	75	74
Only images which contain adult and sexual content.	2	5	0
Only images in the LAION-5B dataset which are copyright-free.	7	9	12
None of the above.	**18**	**7**	**14**
I don't know.	0	2	0
The answer is not in the info. provided.	0	0	0

[a] Grey row indicates the correct answer.
[b] Numbers in bold indicate a less than 70% correct response rate.

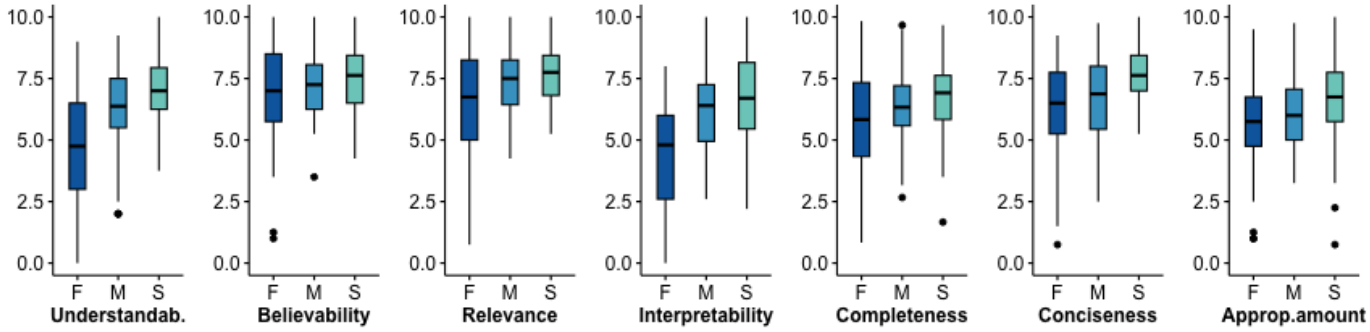

Fig. 1. Perception of the Information Quality dimensions of the Model Cards.

TABLE V
MEAN OF PERCEPTION OF INFORMATION QUALITY DIMENSIONS.

	alpha	Mean (sd) Full	Medium	Short
Understandability	0.92	4.6 (2.4)	6.2 (2.0)	7.0 (1.4)
Believability	0.84	6.9 (2.1)	7.2 (1.4)	7.5 (1.4)
Relevancy	0.88	6.7 (2.1)	7.3 (1.4)	7.8 (1.2)
Interpretability	0.89	4.4 (2.2)	6.3 (1.8)	6.8 (1.7)
Completeness	0.90	5.7 (2.3)	6.3 (1.6)	6.6 (1.6)
Conciseness	0.89	6.1 (2.1)	6.7 (1.9)	7.6 (1.2)
Appropriate amount	0.76	5.6 (1.9)	6.0 (1.5)	6.6 (1.9)

TABLE VI
DIFFERENCE IN PERCEPTION OF INFORMATION QUALITY DIMENSIONS.

	Kruskal-Wallis		Dunn Test (p-value)		
	chi-sq.	p-value	F-M	F-S	M-S
Understandability	23.89	**0.000**	**0.004**	**0.000**	0.061
Believability	2.43	0.297	0.560	0.367	0.502
Relevancy	7.56	**0.023**	0.189	**0.018**	0.222
Interpretability	25.95	**0.000**	**0.000**	**0.000**	0.268
Completeness	5.54	0.063	0.333	0.056	0.255
Conciseness	11.25	**0.004**	0.251	**0.003**	**0.046**
Appropriate amount	6.13	**0.047**	0.446	**0.046**	0.144

Table V shows the Cronbach's alpha values, means and standard deviation for each variable.

To evaluate differences in perceived Information Quality dimensions between the conditions, we conducted Kruskal-Wallis tests with Dunn post-hoc comparisons, with p-values adjusted with the Benjamini-Hochberg method. The results are detailed in Table VI and Figure 1. The results indicate that the Full Model Card was perceived as significantly less understandable and less interpretable than the Medium and Short versions. Perception of the relevancy of the content, conciseness and appropriate amount of information in the Full Model Card were significantly lower compared to the Short version, but not compared to the Medium Model Card. The Medium Model Card was also perceived as significantly less concise than the Short version, but there were no significant differences between these versions in terms of the other dimensions of information quality. Finally, we found no significant differences in the perception of completeness or believability of the Model Card between any of the conditions.

It is important to note that all versions were perceived positively overall, but the Full Model Card version had the lowest mean for the dimensions of Understandability and Interpretability, below the middle point of the scale.

D. Perception towards the AI

We tested the impact of the Model Card versions on Intention to use the AI and Perceived trustworthiness of the AI. First, we examined Cronbach's alpha for all variables. The values were above 0.7, indicating acceptable reliability, and ranged between 0.94 and 0.96. The high Cronbach's alpha values can be explained as being the result of well-established scales. The items for each variable were averaged. The Cronbach's alpha values, means and std. deviation for each variable are detailed in Table VII.

TABLE VII
INTENTION TO USE AND PERCEIVED TRUSTWORTHINESS OF THE AI BEFORE AND AFTER VIEWING THE MODEL CARD.

	alpha	Mean(sd) Full	Medium	Short
Intention to use (Before)	0.95	4.8 (1.4)	5.0 (1.4)	4.6 (1.6)
Intention to use (After)	0.95	4.6 (1.5)	4.9 (1.5)	4.5 (1.5)
Trustworth (Before)	0.94	4.5 (1.4)	4.9 (1.1)	4.5 (1.4)
Trustworth (After)	0.96	4.2 (1.4)	4.7 (1.2)	4.1 (1.4)

We compared the level of each variable between conditions before the participants had seen the Model Card. The results of Kruskal-Wallis tests indicate that there were no significant differences in Intention to use (chi-squared = 3.049, p-value = 0.218) or Perceived trustworthiness (chi-squared = 2.904, p-value = 0.234) between conditions in the "before" state. Then, we evaluated any differences between the "before" and "after" states within each condition, using Wilcoxon signed rank tests. Intention to use the AI was positive in all participants, and it did not significantly change for any condition after viewing the Model Card (Full: p = 0.12, Medium: p = 0.80, Short: p=0.31). Perceived trustworthiness also did not significantly decrease after viewing the Full Model Card (p = 0.052), nor the Medium version (p = 0.06). However, there was a slight (although significant) decrease in perceived trustworthiness in the Short Model Card condition (p = 0.027).

E. Attitude towards Seeking Information about the AI

Finally, we also examined the attitudes towards seeking information about the AI. The Cronbach's alpha for the items in the Attitude scale was 0.94. We calculated the average of the items and conducted a Kruskal-Wallis test. The means and std. deviation for each condition were: Full Model Card mean = 5.4 (sd = 1.4), Medium mean = 5.7 (sd = 0.8) and Short mean = 5.7 (sd = 0.9). No significant differences were found (chi-squared = 0.363, p-value = 0.83) between the conditions. The results indicate that participants had a positive attitude towards seeking information, regardless of the Model Card version they viewed.

V. Discussion

The results show that participants in all conditions had a similar rate of effectiveness in finding the answer to the content questions by referring to the Model Card, with some sections having a very low rate. Confidence was lower in the Full and Medium Model Cards conditions than in the Short condition, but participants in the Full Model Card condition judged their own accuracy better in the Full Model Card condition. In contrast, the confidence of participants in the Medium and Short Model Card conditions had no relationship to their actual accuracy.

With regards to the perception of information quality of the Model Card, on the dimensions examined in this study, the results indicate that perception of the Full Model Card as well as of the shorter versions is generally positive. However, participants in the Full version condition found this Model Card to be lower in understandability and interpretability compared to those shorter versions. It may be that the technical content signals to non-expert participants that they are not the main intended audience for the Model Card, and therefore participants do not expect that they will be able to understand it. It is important to note that the questions did not at any point ask about the technical sections. This suggests that only their presence in the document was sufficient to reduce perception of understandability and interpretability of the Model Card.

Also notable, the results showed no difference in perceived completeness between the Model Card versions. Participants viewed only one version of the Model Card, but the perception that the information was sufficient for using the AI was at a similar level for all versions. This may be due to the fact that participants did not have a reference to compare the Model Card with, and therefore could only assume what a complete Model Card would look like. Although we did not further ask about which information they would expect in the AI Model Card, future research is planned to validate whether participants have concrete expectations of the information that the this type of document should include.

With regards to the impact on opinions about the AI itself, the results indicate that viewing the Full Model Card had no effect on intention to use the AI or on how trustworthy the AI was perceived to be. And only participants who viewed the Short Model Card had a slight decrease in the level of perceived trustworthiness of the AI. About the effect of the Short Model Card, one hypothesis for this decrease in trustworthiness is that participants in the Short Model Card condition were more engaged with the content of the Model Card, since participants in the Short Model Card condition spent more time in the sections of the survey with access to the document. The knowledge they obtained about the AI might have resulted in this adjustment of their perception. However, the effect is small, and the participants did no interact directly with the AI; therefore, further research is required to validate these results.

As the pace of AI development accelerates, access to AI models by the general public will increase. It is important for developers to provide understandable and usable information to non-experts that will allow them to safely make use (directly or indirectly) of these AI models while avoiding overconfidence. The results indicate that Model Cards such as the one used in this study have room for improvement. In real use scenarios, users would not be required to search and read the document. Therefore, if the Model Card is perceived to lack understandability and interpretability due to the technical information, as suggested by the findings, then users might ignore it altogether.

Another issue is that although the type of content can influence perception of the Model Cards, there remains the difficulty of generating the Model Cards in the first place. This is a difficulty shared by various types of documentation. For example, Pushkarna et al. [33], who proposed Data Cards that focus on information about datasets used in machine learning as a complement to Model Cards, found that producers of datasets would create data cards from already completed data cards of earlier data collections – resulting in inaccuracies and error propagation. In the Stable Diffusion v2 Model Card used in the current study, it is also indicated that certain sections were adapted from the DALLE-MINI model card. Although we do not investigate the accuracy of the Model Card in this study, this type of practice illustrates the challenges of creating the Model Cards. While future regulation in this space is likely, it will be important to make use of the lessons learned from research on documentation and on privacy policies and consider the requirements of users and the difficulties of producers of this type of documentation.

A. Limitations

The present study has a number of limitations. First, we focus on only one example of a Model Card, which may not be representative of other AI model information documents. Nevertheless, the Model Card for the experiment corresponds to a real example of an AI currently in popular use and so the findings can provide useful insights. Second, although we described the AI to participants and provided examples, this may still be limited to truly give an idea about the AI, in particular for those participants who had no experience with it. Future research will include testing the effect of the Model Card information on the behavior of participants. Third, participants were recruited online using the Prolific platform, and the sample was not set be to representative. The findings

might not generalize to other populations, and therefore further research is needed to validate the results.

VI. Conclusion

In this study, we evaluate the effectiveness and the perception of information quality of an existing AI Model Card among non-expert participants and compare it to shorter versions. The results show that participants can find answers to non-technical questions about the AI at a similar rate of effectiveness with the Full Model Card as with shorter versions, but that rate is low for some questions. However, although participants were not asked to rely on technical content, the perception of the understandability and interpretability of the Full Model Card (which contained technical content) was lower than for shorter versions (which did not). Neither the Full Model Card nor the shorter versions had a significant effect on intention to use the AI, and shortest version had a slight significant negative effect on perceived trustworthiness of the AI. Finally, participants in general had a positive attitude towards seeking information about the AI. Future research is planned to evaluate the effectiveness and perception of Model Cards for other AI models currently in use, such as AI for text generation, and in addition to evaluate the effect of the Model Card on how users handle the output of the AI.

References

[1] J. Schulman, "ChatGPT: Optimizing language models for dialogue," OpenAI, 11 2022. [Online]. Available: https://web.archive.org/web/20221130180912/https://openai.com/blog/chatgpt/

[2] OpenAI, "Dall·e 2 is an AI system that can create realistic images and art from a description in natural language," OpenAI, 04 2023. [Online]. Available: https://web.archive.org/web/20230407013913/https://openai.com/product/dall-e-2

[3] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," 2022.

[4] Stability AI, "Stable diffusion github repository," 04 2023. [Online]. Available: https://github.com/Stability-AI/stablediffusion

[5] OpenAI, "GPT-4 technical report," 2023. [Online]. Available: https://arxiv.org/pdf/2303.08774v3.pdf

[6] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[7] "The OECD Artificial Intelligence (AI) Principles," https://www.oecd.ai/ai-principles.

[8] European Commission, "Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts (COM(2021) 206 final)," 2021.

[9] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 220–229.

[10] M. Arnold, R. K. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilović, R. Nair, K. N. Ramamurthy, A. Olteanu, D. Piorkowski *et al.*, "Factsheets: Increasing trust in AI services through supplier's declarations of conformity," *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 6–1, 2019.

[11] S. Kapoor and A. Narayanan, "OpenAI's policies hinder reproducible research on language models," Substack, 03 2022. [Online]. Available: https://aisnakeoil.substack.com/p/openais-policies-hinder-reproducible

[12] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and mitigating unintended bias in text classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 2018, pp. 67–73.

[13] B. Brittain, "Lawsuits accuse AI content creators of misusing copyrighted work," Reuters, 01 2023. [Online]. Available: https://www.reuters.com/legal/transactional/lawsuits-accuse-ai-content-creators-misusing-copyrighted-work-2023-01-17/

[14] J. Turley, "Defamed by ChatGPT: My own bizarre experience with artificiality of "artificial intelligence"," 04 2023. [Online]. Available: https://jonathanturley.org/2023/04/06/defamed-by-chatgpt-my-own-bizarre-experience-with-artificiality-of-artificial-intelligence/

[15] A. M. McDonald and L. F. Cranor, "The cost of reading privacy policies," *ISJLP*, vol. 4, p. 543, 2008.

[16] H. Shen, L. Wang, W. H. Deng, C. Brusse, R. Velgersdijk, and H. Zhu, "The model card authoring toolkit: Toward community-centered, deliberation-driven AI design," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 440–451.

[17] A. Crisan, M. Drouhard, J. Vig, and N. Rajani, "Interactive model cards: A human-centered approach to model documentation," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 427–439.

[18] J. L. Nunes, G. D. Barbosa, C. S. de Souza, H. Lopes, and S. D. Barbosa, "Using model cards for ethical reflection: a qualitative exploration," in *Proceedings of the 21st Brazilian Symposium on Human Factors in Computing Systems*, 2022, pp. 1–11.

[19] C.-W. Chiang and M. Yin, "Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models," in *27th International Conference on Intelligent User Interfaces*, 2022, pp. 148–161.

[20] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal, "How short is too short? Implications of length and framing on the effectiveness of privacy notices," in *Twelfth symposium on usable privacy and security (SOUPS 2016)*. USENIX Association, 2016, pp. 321–340.

[21] Stability AI, "Stable Diffusion Version 2," 12 2022. [Online]. Available: https://github.com/Stability-AI/stablediffusion/

[22] Y. W. Lee, D. M. Strong, B. K. Kahn, and R. Y. Wang, "AIMQ: a methodology for information quality assessment," *Information & management*, vol. 40, no. 2, pp. 133–146, 2002.

[23] C. B. Nordheim, "Trust in chatbots for customer service–findings from a questionnaire study," Master's thesis, University of Oslo, 2018.

[24] V. Venkatesh, F. D. Davis *et al.*, "Theoretical acceptance extension model: Field four studies of the technology longitudinal," *Management science*, vol. 46, no. 2, pp. 186–204, 2000.

[25] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS quarterly*, pp. 425–478, 2003.

[26] T. Zarmpou, V. Saprikis, A. Markos, and M. Vlachopoulou, "Modeling users' acceptance of mobile services," *Electronic Commerce Research*, vol. 12, pp. 225–248, 2012.

[27] J. Schöffer, Y. Machowski, and N. Kühl, "A study on fairness and trust perceptions in automated decision making," in *Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA*, 2021, p. 170005.

[28] L. Carter and F. Bélanger, "The utilization of e-government services: citizen trust, innovation and acceptance factors," *Information systems journal*, vol. 15, no. 1, pp. 5–25, 2005.

[29] C.-M. Chiu, H.-Y. Lin, S.-Y. Sun, and M.-H. Hsu, "Understanding customers' loyalty intentions towards online shopping: An integration of technology acceptance model and fairness theory," *Behaviour & Information Technology*, vol. 28, no. 4, pp. 347–360, 2009.

[30] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, p. 2053951718756684, 2018.

[31] I. Ajzen, "Perceived behavioral control, self-efficacy, locus of control, and the theory of planned behavior 1," *Journal of applied social psychology*, vol. 32, no. 4, pp. 665–683, 2002.

[32] L. Kahlor and S. Rosenthal, "If we seek, do we learn? predicting knowledge of global warming," *Science Communication*, vol. 30, no. 3, pp. 380–414, 2009.

[33] M. Pushkarna, A. Zaldivar, and O. Kjartansson, "Data cards: Purposeful and transparent dataset documentation for responsible ai," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1776–1826.

All URLs have been last visited on April 15th, 2023.