DeepPhish: Understanding User Trust Towards Artificially Generated Profiles in Online Social Networks (Supplementary Materials)

Jaron Mink*, Licheng Luo*, Natã M. Barbosa*, Olivia Figueira†, Yang Wang*, Gang Wang*
*University of Illinois at Urbana-Champaign †Santa Clara University
{jaronmm2, ll6, natamb2}@illinois.edu, ofigueira@alumni.scu.edu, {yvw, gangw}@illinois.edu

A Pilot Studies

Before running the final study (presented in Section 3), we ran several pilot studies to identify potential errors and survey presentation issues. These pilots included two MTurk studies in September of 2020 ($n = 27$) and January of 2021 ($n = 96$), and a virtual, "think-aloud" study with colleagues in February of 2021 ($n = 8$). These pilots have led to key methodology improvements, such as adding different prompt levels (detailed in Section 3.4) and adding the generalized-trust questions (**Q9**–**Q11**). In addition, these pilots have helped us to identify certain confusing UIs and improve our framing of the questions. For instance, we initially had "negatively-worded" questions for **Q1**–**Q3**, but they caused major confusions during the pilots. In our final design, all questions are positively framed. We also changed our wording for the "benevolence" question (**Q3**). Our initial version was "*This person will go out of his/her way to help in case of an emergency during our collaboration.*" However, a pilot participant mentioned that the wording seems entirely situation-dependent (i.e., what emergency was happening). We ultimately converged on the current version: "*This person will make newcomers feel welcome*." Note that the definition of benevolence is "the extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive" [4]. Our framing of making a stranger feel welcome fits this definition.

B Trust Factor Correlation

In our study, participants rated the trustworthiness of each provide via three questions related to the perceived integrity [**Q1**], ability [**Q2**], and benevolence [**Q3**] of the profile. In Figure 1, we show the pairwise Pearson product-moment correlation between the three ratings of the profiles. While these factors are theoretically orthogonal [4], we find a strong positive pair-wise correlation between the measured results. Similar to prior work [1–3], we use the mean of the three ratings given by a participant to a profile to form a single "trust score" for our analysis.

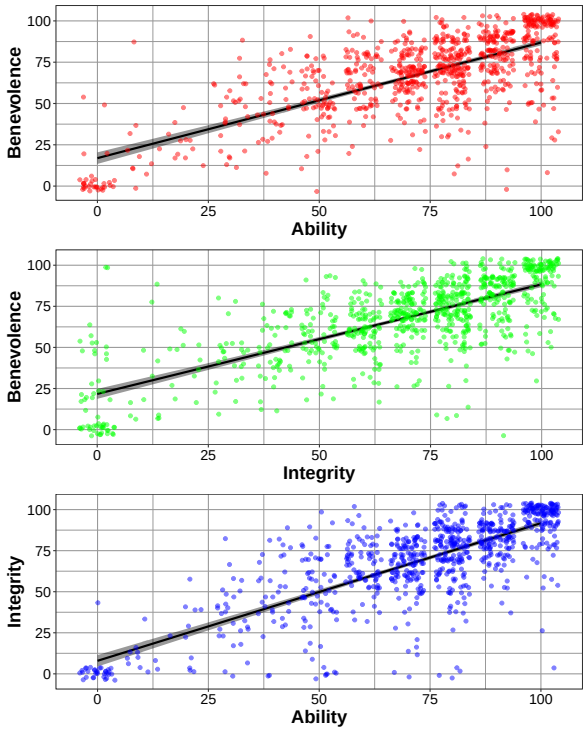

Figure 1: **Correlation of Three Trust Scores** – We show the pair-wise scatter plots of each trust factor along with the regression line and a 95% confidence interval. Under a Pearson product-moment correlation test, these trust factors of ability (A), benevolence (B), and integrity (I) have a significant correlation with one another: B-A ($r = 0.7117, n = 858, p < 0.001$), B-I ($r = 0.7349, n = 858, p < 0.001$), and I-A ($r = 0.7729, n = 858, p < 0.001$).

C Analysis of Additional Factors

To evaluate the effect of education, experience, and prior social media usage has on the reported trust and acceptance of profiles, we extend the quantitative analysis in Section 4.1 and 4.2 to cover additional factors. Specifically, we include participants self-reported experience (**Q12**), educational degree (**Q17**), prior usage of LinkedIn (**Q13**), and number of

Variable	Estimate (β)	Std. Err.	p-value
Intercept	72.888	4.054	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	5.153	2.011	0.011*
Hard Prompt	-4.399	1.997	0.028*
Artifact (Reference = Consistent)			
Inter Image	-2.112	2.091	0.313
Inter Text	-10.609	2.070	<0.001***
Intra Image	-12.138	2.078	<0.001***
Intra Text	-7.406	2.089	<0.001***
Gender (Reference = Female)			
Male	-1.803	1.721	0.296
Non-Binary	-9.611	8.230	0.244
Age	0.155	0.438	0.723
Generalized Trust	0.421	0.054	<0.001***
Highest Education (Reference = Bachelors)			
No Bachelors	2.947	2.370	0.215
Graduate	-2.682	2.206	0.225
IT/Comp. Sci. Background (Reference = No Experience)			
Has Experience	-2.198	1.828	0.230
Human Resources Background (Reference = No Experience)			
Has Experience	-2.640	2.132	0.217
Business/Finance Background (Reference = No Experience)			
Has Experience	-1.510	2.018	0.455
LinkedIn Background (Reference = Does Not Use)			
Uses/Has Used	1.974	2.053	0.337
Friend Count	0.564	0.422	0.182

Table 1: **Trust - Extended Analysis**—Linear mixed-effects regression model extended with participants professional and social background (in italics). "Friend Count" (in units of 100 friends) is numeric and thus doesn't have a reference group. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

Variable	Estimate (β)	Std. Err.	p-value
Intercept	1.498	0.452	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	0.809	0.229	<0.001***
Hard Prompt	-0.179	0.200	0.370
Artifact (Reference = Consistent)			
Inter Image	-0.556	0.282	0.048*
Inter Text	-1.162	0.264	<0.001***
Intra Image	-1.520	0.264	<0.001***
Intra Text	-1.103	0.265	<0.001***
Gender (Reference = Female)			
Male	-0.074	0.183	0.686
Non-Binary	0.500	0.950	0.599
Age	0.018	0.047	0.710
Generalized Trust	0.019	0.006	<0.001***
Highest Education (Reference = Bachelors)			
No Bachelors	-0.085	0.242	0.725
Graduate	0.522	0.253	0.039*
IT/Comp. Sci. Background (Reference = No Experience)			
Has Experience	0.314	0.197	0.112
Human Resources Background (Reference = No Experience)			
Has Experience	0.077	0.233	0.743
Business/Finance Background (Reference = No Experience)			
Has Experience	0.037	0.218	0.867
LinkedIn Background (Reference = Does Not Use)			
Uses/Has Used	-0.036	0.218	0.870
Friend Count	0.020	0.045	0.652

Table 2: **Acceptance - Extended Analysis**—Logistic mixed effects regression model extended with participants professional and social background (in italics). The unit for estimate and standard error is log odds scaled. "Friend" is numeric (in units of 100 friends) and thus doesn't have a reference group. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

active social media connections (**Q14**) as additional fixed effects within the previously evaluated models. The results are shown in Table 1 and Table 2. Note that, to reduce noise from the "education" categories with few respondents, we group participants into three bins, namely "No Bachelor's", "Bachelor's", and "Graduate".

This analysis shows that most of these additional factors (italicized in the tables) did not significantly affect either the perceived trust or acceptance rate. Also, the significance of all the previously evaluated factors in the main paper remain unchanged. The only exception is shown in Table 2 where we find a graduate degree significantly increased the log-likelihood of profile acceptance, compared with a bachelor's degree. One plausible explanation is that a participant with a graduate degree may be more willing to network with our experimental profiles given 2/3 profile templates also held graduate degrees (a homophily effect); however, we did not find qualitative data to support this conjecture and thus leave further evaluation to future work.

D Artifact Comparisons Results

To understand how different artifact conditions compare to one another, we run a post-hoc analysis on our main regression models for perceived trustworthiness and profile acceptance, respectively. More specifically, to determine the differences between treatments (i.e., artifact conditions), we compare the estimated marginal means produced by the models presented in Section 4.1 and Section 4.2. We use a Bonferroni correction to allow us to make multiple comparisons without inflation of the false-positive rate. The results are shown in Table 3 and Table 4.

Contrast	Estimate	Std. Err.	p-value
Inter Image - Inter Text	8.681	2.516	0.004**
Inter Image - Intra Image	10.315	2.446	<0.001***
Inter Image - Intra Text	5.189	2.458	0.210
Inter Text - Intra Image	1.634	2.430	1.000
Inter Text - Intra Text	-3.492	2.437	0.914
Intra Image - Intra Text	-5.126	2.522	0.255

Table 3: **Pairwise Artifact Analysis on Perceived Trustworthiness**—To compare the effects between different conditions within the trust model (Table 1), a pairwise comparisons test is performed. Displayed are the results of a Bonferroni method post-hoc comparison of the estimated marginal means. A significant result means that the two factors have a statistically different effect upon perceived trustworthiness of a profile. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

Contrast	Estimate	Std. Err.	p-value
Inter Image - Inter Text	0.572	0.290	0.291
Inter Image - Intra Image	0.955	0.287	0.005**
Inter Image - Intra Text	0.522	0.290	0.433
Inter Text - Intra Image	0.383	0.263	0.876
Inter Text - Intra Text	-0.050	0.270	1.000
Intra Image - Intra Text	-0.433	0.270	0.649

Table 4: **Pairwise Artifact Analysis on Profile Acceptance**—To compare the effects between different artifacts conditions within the acceptance model (Table 2), a pairwise comparisons test is performed. Displayed are the results of a Bonferroni method post-hoc comparison of the estimated marginal means. A significant result means that the two factors have statistically different effect upon the acceptance of a profile request. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

We first compare different artifacts for their effects on the perceived trustworthiness of a profile. The results (Table 3) show that profiles with inter-field image artifacts are significantly more trusted by participants compared to profiles with intra-field image artifacts and inter-field text artifacts.

Similarly, for acceptance rate analysis (Table 4), the only significant difference is found between inter-field image and intra-field image artifacts.

E Full Survey Questions

E.I Main Tasks.

Participants are instructed to review three profiles (Database Administrator, Human Resource Manager, and Billing Manager). Under each profile, the participants will answer 5 questions ($5 \times 3 = 15$ questions in total).

Q1–Q3 Based on the profile, please rate to what extent you agree with each of these statements:

Q1 This profile is an accurate depiction of the user

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q2 This user is knowledgeable in their role as a [Database Administrator / Human Resource Manager / Billing Manager]

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q3 This person will make newcomers feel welcome

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q4 What aspect of this profile most influenced your ratings?

Q5 Based on the profile above, would you accept or ignore the connection request?

- Accept
- Ignore

E.II Attention Check Questions

Q6 What is your role in this scenario?

- I am receiving friend requests from coworkers I haven't met before on LinkedIn
- I am moderating LinkedIn content
- I am choosing who to message for advice on LinkedIn
- I am applying for jobs on LinkedIn
- I am sending friend requests to familiar coworkers on LinkedIn

Q7 This is a control question. Please leave the question below blank.

- 0 times
- 1-2 times
- 3-4 times
- 4+ times

E.III Follow-up Questions

Q8 Please describe the strategies you used to assess profiles.

Q9–11 Please rate to what extend you agree with each of these statements:

Q9 In general, most people are honest

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q10 In general, most people are qualified for their job

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q11 In general, most people are good and kind

- 0% (Strongly Disagree)
- 10%
- 20%
- 30%
- 40%
- 50% (Undecided)
- 60%
- 70%
- 80%
- 90%
- 100% (Strongly Agree)

Q12 Do you have any significant experience in, or knowledge of, any of the following fields (Select all that apply)

- Business or Financial Operations
- Information Technology or Computer Science/Development
- Human Resources
- No significant experience or knowledge in any of the above fields

Q13 Do you currently use, or have previously used, LinkedIn?

- Yes
- No

Q14 (If Q13==Yes): How many 1st-degree (or direct) connections do you have on LinkedIn? (If Q13==No): How many friends or connections do you have on your most actively used social media application?

- 0-100
- 101-200
- 201-300
- 301-400
- 401-500
- 501-600
- 601-700
- 701-800
- 801-900
- 901-1000
- 1001+
- Prefer not to answer

Q15 What is your age?

- 18-19
- 20-24
- 25-29
- 30-35
- 35-39
- 40-44
- 45-49
- 50-54
- 55-59
- 60-64
- 65-69
- 70+
- Prefer not to answer

Q16 What is your gender?

- Female
- Male
- Non-Binary
- Prefer to self-describe
- Prefer to not disclose

Q17 What is your highest educational degree?

- High School Graduate or Less
- Some College or Two-Year Associate Degree
- Bachelor's Degree
- Master's Degree
- PhD
- Professional Degree
- Prefer to not disclose

Q18 What Ethnicity do you most identify with?

- White
- Hispanic or Latino
- Black or African American
- Asian
- American Indian or Alaska Native
- Native Hawaiian or Pacific Islander
- Other
- Prefer to not disclose

F Participant Demographics

We presented the self-reported demographic information [**Q15** - **Q18**] of participants for each treatment.

Attributes	None	Soft	Hard	**Total**
Gender				
Female	34	38	39	**111**
Male	60	56	55	**171**
Non-Binary	2	1	0	**3**
Prefer not to say	0	0	1	**1**
Age				
18-20	0	0	0	**0**
20-24	0	5	7	**12**
25-29	21	17	20	**58**
30-34	25	24	23	**72**
35-39	19	18	22	**59**
40-44	10	13	8	**31**
45-49	8	7	4	**19**
50-54	2	8	7	**17**
55-59	7	2	1	**10**
60-64	2	0	2	**4**
65-69	1	0	0	**1**
70+	0	1	1	**2**
Prefer not to say	1	0	0	**1**
Ethnicity				
White	68	72	65	**205**
Hispanic or Latino	5	1	6	**12**
Black or African American	12	16	19	**47**
Asian	6	3	3	**12**
American Indian or Alaska Native	2	2	0	**4**
Native Hawaiian or Pacific Islander	1	0	0	**1**
Other	2	0	1	**3**
Prefer not to say	0	1	1	**2**
Highest Education				
High School or Less	2	5	4	**11**
Some College or Two Year Degree	19	14	17	**50**
Bachelor's	60	53	54	**167**
Master's	12	22	16	**50**
PhD	0	1	2	**3**
Professional Degree	3	0	1	**4**
Prefer not to say	0	0	1	**1**
Total	96	95	95	**286**

Table 5: **Participant Demographics**— We show the demographic information of participants for different prompt treatments.

G Qualitative Codebooks

We presented the codebooks and code counts for the open-form questions in our study. Table 6 is for [**Q4**] and Table 7 is for [**Q8**].

Primary Code	Prim. Freq.	Secondary Code	Sec. Freq.	Description	Participant Sample
ABOUT	94,91,119	ARTIFACT	11,23,29	The grammar or references in "About" are incorrect	"The grammar in the description is messed up"
		CONSISTENT	7,4,16	The grammar or references in "About" are correct	"Their About Me section seems accurate and well-written"
		CONTENT	47,30,52	Skills, information, or quotes from the "About"	"Her bio indicates flexibility in her role and dedication to service."
		META	11,6,6	Relating to the writing quality of "About"	"They had a well-written paragraph introducing themselves"
EDUCATION	53,49,37	ARTIFACT	1,8,6	The degree is not correct for position	"Thomas Edison State College sounds fake"
		CONSISTENT	3,4,2	The degree is correct for position	"The work experience and education seem to match up and be correct"
		CONTENT	20,14,11	Referencing the field or university of the degree	"She has a masters that is persuasive as well."
		META	3,1,0	Referencing the time it took to achieve the degree	"She has many years of ... schooling"
EXPERIENCE	200,154,124	ARTIFACT	6,7,12	The experience doesn't match or is otherwise suspicious	"The description indicates they are not working at the position that is stated."
		CONSISTENT	7,7,9	The experience looks appropriate	"His experience adds up to the time stated and he has worked for the company mentioned. "
		CONTENT	71,55,38	Referencing a position of employment	"She has wide experience in the same field"
		META	72,54,28	Referencing the length of time or amount of experience	"She hasn't worked with the company for very long."
FEELING	135,148,120	REAL	15,17,25	The profile is an accurate representation of a user	"The profile seems credible, I do not observe any major issues"
		FAKE	5,23,22	The profile does not accurately represent a user	"I don't think this person is real"
		POSITIVE	86,84,53	Positive feelings about the profile	"This person seems friendly and like a hard-worker who has mastered many different important parts of her job."
		NEGATIVE	14,20,13	Negative feelings about the profile	"I think his profile is not fit for this job."
		MIXED	24,6,9	Mixed feelings/unsure of the profile	"I'm willing to give him a shot, but barely."
NAME	0,2,4	ARTIFACT	0,2,3	Name is inconsistent with other aspects	"The picture shows a Black woman but the name seems to belong to a White man"
		CONSISTENT	0,0,1	Name is appropriate for person	"Being Chris. M the third lessens the likelihood of the profile being made by a bot simply because I doubt an AI would make a name like that."
PICTURE	48,37,68	ARTIFACT	5,15,30	The picture looks strange or incorrectly formed	"The image of the person seems to look like it was altered in my opinion."
		CONSISTENT	1,5,14	The picture looks to be a legitimate photograph	"The picture looks real and she looks nice enough. "
		COMPOSITION	4,4,0	Related to the image clarity or person's gender/age	"You don't see many women in computer science fields, so it would be nice to have connections with a more diverse group of people in this area."
		EXPRESSION	32,7,6	Related to the person's facial expression or style	"She has a very friendly smile"
CONNECTIONS	0,1,0	–	–	Related to the friends or connections of profile.	"This person has ... no connections"
NON RESPONSE	10,18,12	–	–	Did not respond to the question	

Table 6: **Factors that Influence User Trust Rating.**—We show the codebook and counts for the open-ended question **Q4** (κ=0.87, Appendix E). We show the factors that influence users' trust rating on a given profile. Under Primary Code Frequency ("Prim. Freq.") and Secondary Code Frequency ("Sec. Freq."), we show the occurrences of codes for no prompt ($n = 288$), soft prompt ($n = 285$), and hard prompt ($n = 285$), respectively.

Primary Code	Prim. Freq.	Secondary Code	Sec. Freq.	Description	Participant Sample
Inconsistencies	15,33,41	INTRA-FIELD	0,17,28	Searches for inconsistencies/artifacts *within* a field	"I read through the profile for any errors in logic, such as repetition"
		INTER-FIELD	11,12,16	Searches for inconsistencies/artifacts *between* fields	"I tried to match up with their About profiles to their experiences and maybe education."
Qualities	42,22,15	ABILITY	35,16,13	Searches for how professionally competent a person is	"I evaluated the profile in terms of experience and familiarity with the job."
		PERSONALITY	17,7,4	Searches for some personal trait not related to competence	"If someone seemed pretty grounded and professional... I decided to accept them"
		CLEAR COMMUNICATION	0,2,1	Searches for clear wording and logic in text	"If the profile description was well-written this indicated, to me, a well-rounded person"
Reason	6,3,1	POTENTIAL BENEFIT	2,2,1	Acceptance influenced by the benefit the user may provide.	"If their position was important then that is all the more reason to talk with them."
		NO REASON TO DECLINE	2,0,0	Doesn't see a reason to decline a connection request	"It would be bad form to decline someone as that would only hurt and not help me."
		OBLIGATION TO ACCEPT	3,0,0	Feels obligated to accept based on social expectations	"These were all future coworkers ... I feel it's incumbent upon me to accept their connection requests."
		JOB IS IN DEMAND	0,1,0	Acceptance influenced by position demand	"I accepted based on ... demand for positions I thought there might be."
UI	79,65,74	EXPERIENCE	65,50,44	Uses experience in decision making	"I used years of work experience and positions held"
		ABOUT	43,39,45	Uses about self-description in decision making	"Looked primarily at "About" section."
		EDUCATION	25,27,22	Uses educational background in decision making	"Based it on their ... level of education."
		IMAGE	18,25,43	Uses profile image in decision making	"I looked at their general attitude and appearance in their profile picture. "
		NAME	0,1,1	Uses the user's name in decision making	"I looked at the name"
NON RESPONSE	12,19,12	–	–	Did not respond to the question	

Table 7: **Strategies for Assessing Profiles**—We show the codebook and counts for the open-form question **Q8** (κ=0.72, Appendix E). We show different strategies users mentioned to assess profiles. Under Primary Code Frequency ("Prim. Freq.") and Secondary Code Frequency ("Sec. Freq."), we show the occurrences of codes for no prompt ($n = 96$), soft prompt ($n = 95$), and hard prompt ($n = 95$), respectively.

References

[1] Tommy Bruzzese, Irena Gao, Griffin Dietz, Christina Ding, and Alyssa Romanos. Effect of Confidence Indicators on Trust in AI-Generated Profiles. In *Proc. of CHI EA*, 2020.

[2] Maurice Jakesch, Megan French, Xiao Ma, Jeffrey T. Hancock, and Mor Naaman. AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness. In *Proc. of CHI*, 2019.

[3] Xiao Ma, Jeffery T. Hancock, Kenneth Lim Mingjie, and Mor Naaman. Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles. In *Proc. of CSCW*, 2017.

[4] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 1995.