DeepPhish Analysis

This analysis provides the quantitative results reported in Section 4 and Appendix B

Section 3.6: Demographics Background of Participants

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 Graduate or Less	2	5	4	**11**
Some College or Two Year Degree	19	14	17	**50**
Bachelors	60	53	54	**167**
Masters	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**

Demographics – We show the demographics information of participants for different prompt treatment groups.

Section 3.6: Time distribution of participants

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   4.083   9.250  12.950  13.852  17.100  33.050      23
```

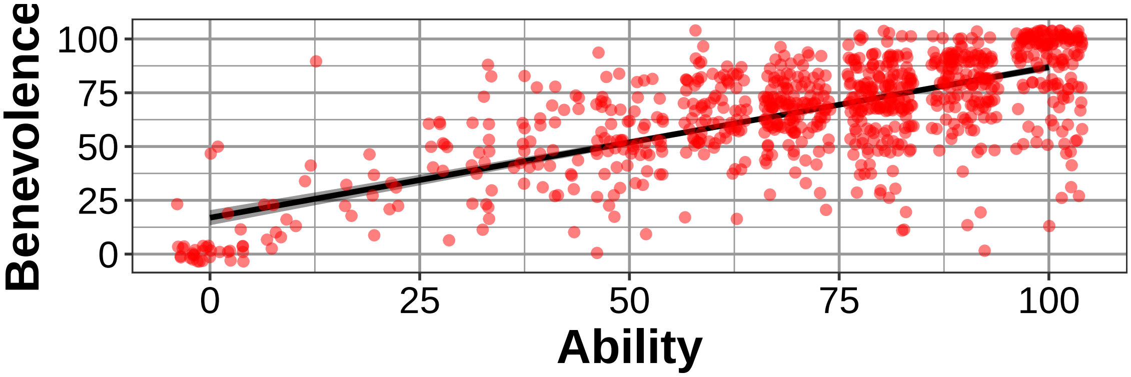
The median time spend on the survey for $n = 286$ participants was 12.95 minutes.

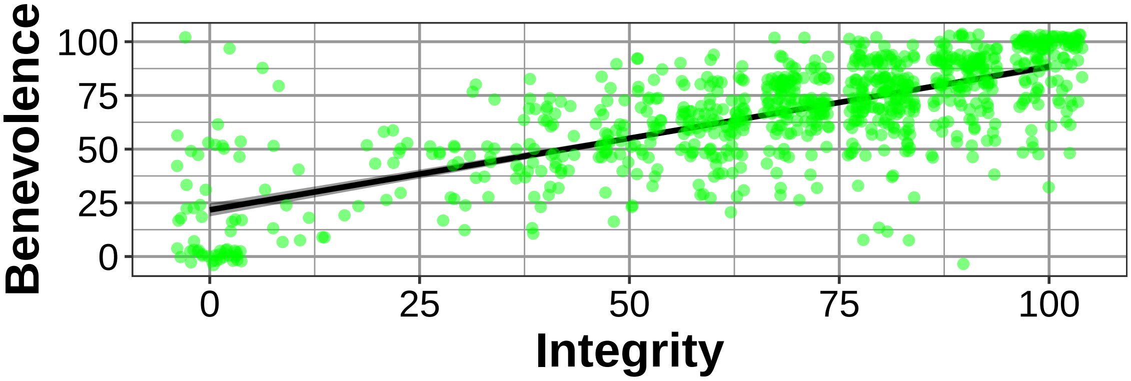
Section 4.1: pairwise correlation of measured factors (Ability, Benevolence, Integrity)

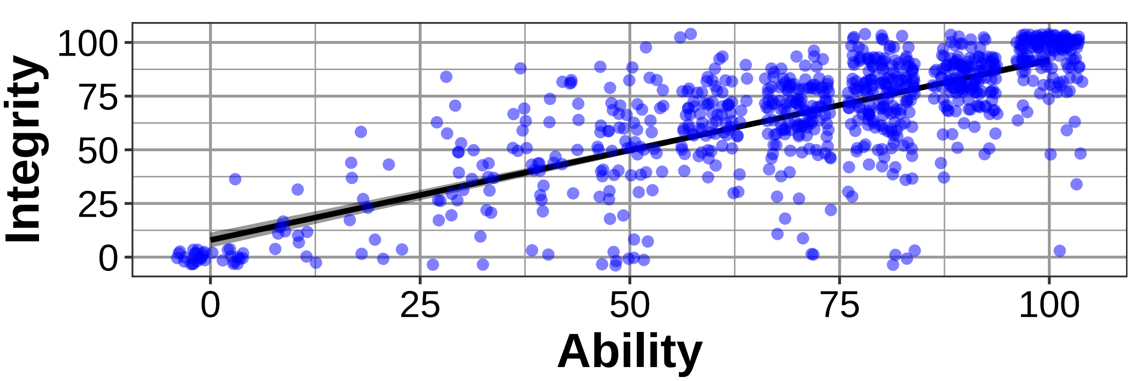
Ability-Benevolence Correlation: 0.71167, p-value: <0.001

Integrity-Benevolence Correlation: 0.73494, p-value: <0.001

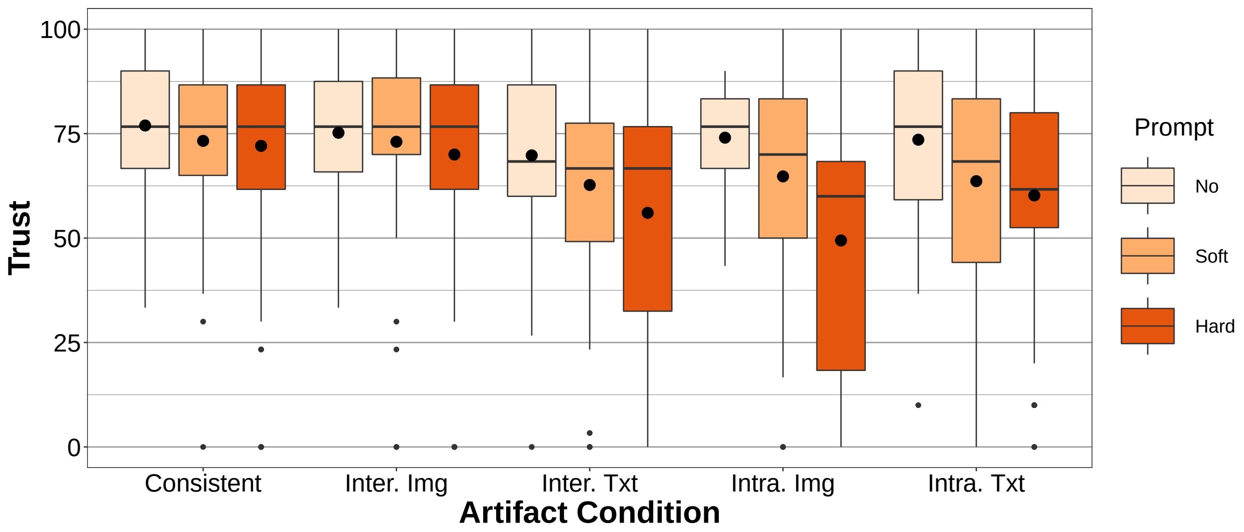
Ability-Integrity Correlation: 0.77294, p-value: <0.001







Section 4.1: Figure 5



Descriptive statistics of Figure 5

prompt	trust_mean	trust_standard_dev
No	74.4	17.4
Soft	68.4	22.8
Hard	63.3	27.3

Section 4.1: Table 1

Variable	Estimate	Std. Err.	p
Intercept	75.118	3.005	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	5.187	2.019	0.011*
Hard Prompt	-4.447	2.024	0.029*
Artifact (Reference = Consistent)			
Inter Image	-2.018	2.093	0.335
Inter Text	-10.700	2.072	<0.001***
Intra Image	-12.334	2.079	<0.001***
Intra Text	-7.207	2.089	<0.001***
Gender (Reference = Female)			
Male	-2.317	1.703	0.175
Non-Binary	-8.664	8.287	0.297
Age	0.027	0.436	0.951
Generalized Trust	0.388	0.053	<0.001***

Trust Rating Analysis – Linear mixed-effects regression model. The unit for estimate and standard error is the aggregated trust score. “Age” (in units of 5 years) and “Generalized Trust” are numeric and thus do not have a reference group. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

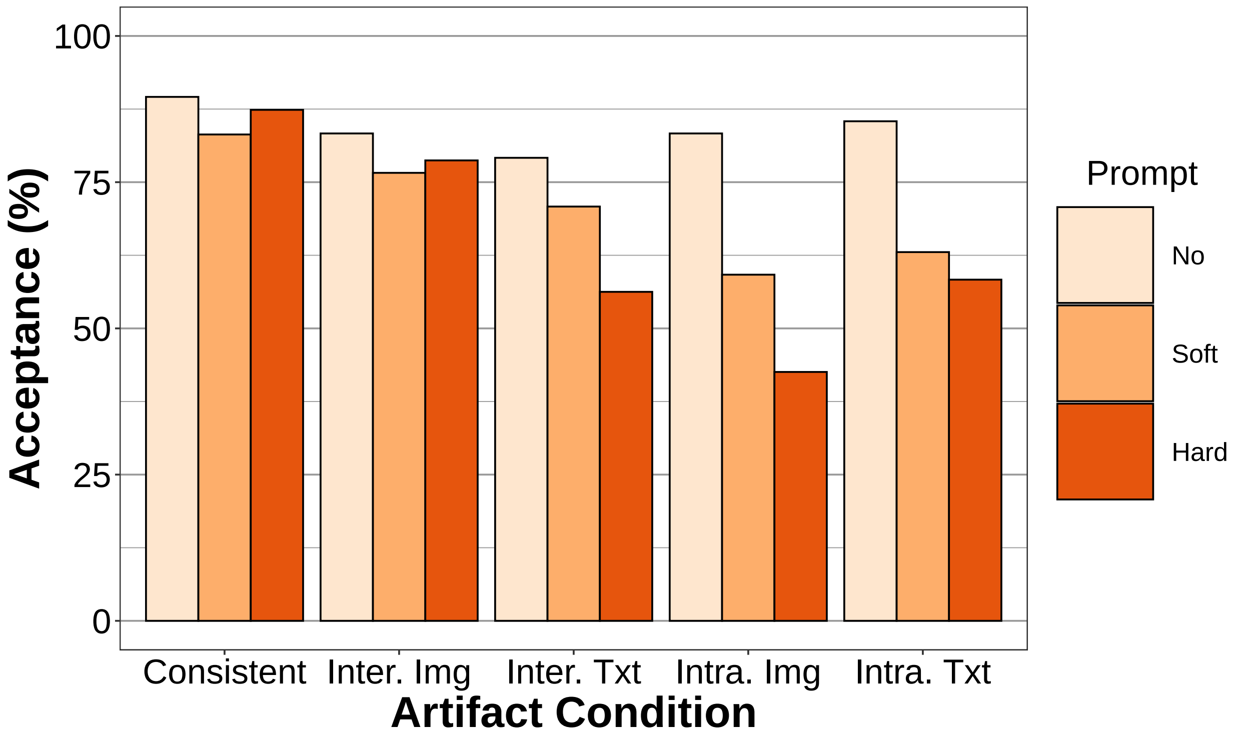
Section 4.1: ANOVA Test for differences in consistent trust

prompt	p_subcond	trust_mean	trust_standard_dev
No	Consistent	76.8	16.0
Soft	Consistent	73.3	18.0
Hard	Consistent	72.1	21.6

ANOVA results

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## prompt        2   1110     555   1.595  0.205
## Residuals   281  97784     348
```

Section 4.2: Figure 6



Descriptive statistics for Figure 6

prompt	p_subcond	req	Percentage
No	Consistent	Accept	90
No	Inter. Img	Accept	83
No	Inter. Txt	Accept	79
No	Intra. Img	Accept	83
No	Intra. Txt	Accept	85
Soft	Consistent	Accept	83

prompt	p_subcond	req	Percentage
Soft	Inter. Img	Accept	77
Soft	Inter. Txt	Accept	71
Soft	Intra. Img	Accept	59
Soft	Intra. Txt	Accept	63
Hard	Consistent	Accept	87
Hard	Inter. Img	Accept	79
Hard	Inter. Txt	Accept	56
Hard	Intra. Img	Accept	43
Hard	Intra. Txt	Accept	58

Section 4.2: Table 2

Variable	Estimate (Prob)	Std. Err. (Prob)	p-value
Intercept	1.702	0.350	<0.001***
Prompt (Reference = Soft Prompt)			
No Prompt	0.761	0.228	<0.001***
Hard Prompt	-0.163	0.201	0.418
Artifact (Reference = Consistent)			
Inter Image	-0.570	0.282	0.043*
Inter Text	-1.142	0.263	<0.001***
Intra Image	-1.525	0.264	<0.001***
Intra Text	-1.092	0.264	<0.001***
Gender (Reference = Female)			
Male	-0.026	0.180	0.887
Non-Binary	0.516	0.940	0.583
Age	0.025	0.047	0.589
Generalized Trust	0.023	0.006	<0.001***

Request Acceptance Analysis – Logistic mixed effects regression model. The unit for estimate and standard error is log odds scaled. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

Section 4.3: Artifact to Artifact trust comparison

contrast	estimate	SE	df	t.ratio	p.value
Inter Image - Inter Text	8.681177	2.515677	835.6082	3.4508320	0.0035213
Inter Image - Intra Image	10.315447	2.446075	728.6418	4.2171427	0.0001671
Inter Image - Intra Text	5.188987	2.457623	724.9941	2.1113842	0.2104857
Inter Text - Intra Image	1.634270	2.430172	722.0600	0.6724916	1.0000000
Inter Text - Intra Text	-3.492190	2.437012	727.6002	-1.4329804	0.9137578
Intra Image - Intra Text	-5.126460	2.522126	833.0379	-2.0325949	0.2544611

Section 4.3: Artifact to Artifact comparison for acceptance rate

contrast	estimate	SE	df	z.ratio	p.value
Inter Image - Inter Text	0.5722710	0.2900032	Inf	1.9733263	0.2907504
Inter Image - Intra Image	0.9550967	0.2867158	Inf	3.3311616	0.0051891
Inter Image - Intra Text	0.5221196	0.2903249	Inf	1.7983975	0.4326842
Inter Text - Intra Image	0.3828258	0.2633236	Inf	1.4538223	0.8759734
Inter Text - Intra Text	-0.0501514	0.2695911	Inf	-0.1860276	1.0000000
Intra Image - Intra Text	-0.4329772	0.2695806	Inf	-1.6061141	0.6494931

Section 5.1: Table 3

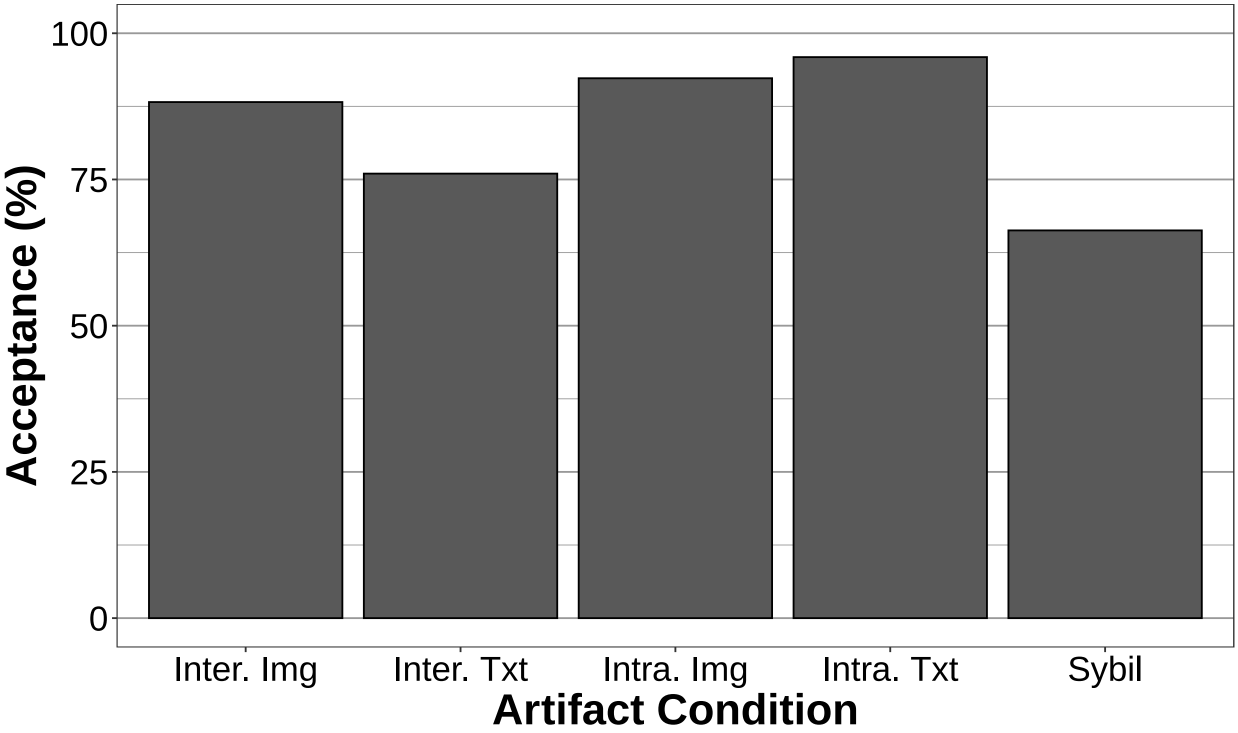
Metrics		No	Soft	Hard
Profile Image (expand)		12%	17%	40%
Experience (expand)		36%	28%	30%
About (hover over)	Mean	4841	4486	7341
	Median	611	331	989
Experience (hover over)	Mean	7760	6646	6584
	Median	3376	1872	1912
Education (hover over)	Mean	1564	1957	1695
	Median	126	133	128

Mouse Tracking Results – The top two rows report the percentage of participants that clicked on the UIs to expand profile images and the experience items. The bottom three rows report the mouse hover-over time (median and mean) in each UI section for “About”, “Experience”, and “Education” in milliseconds. After the Hard prompt, users are more likely to look for artifacts in profile images and the “About” text.

Appendix B: Deepfake Profiles vs. Real-World Sybils

$n = 101$ participants were included in the follow-up survey.

Appendix B: Figure 9



Descriptive statistics for Figure 9

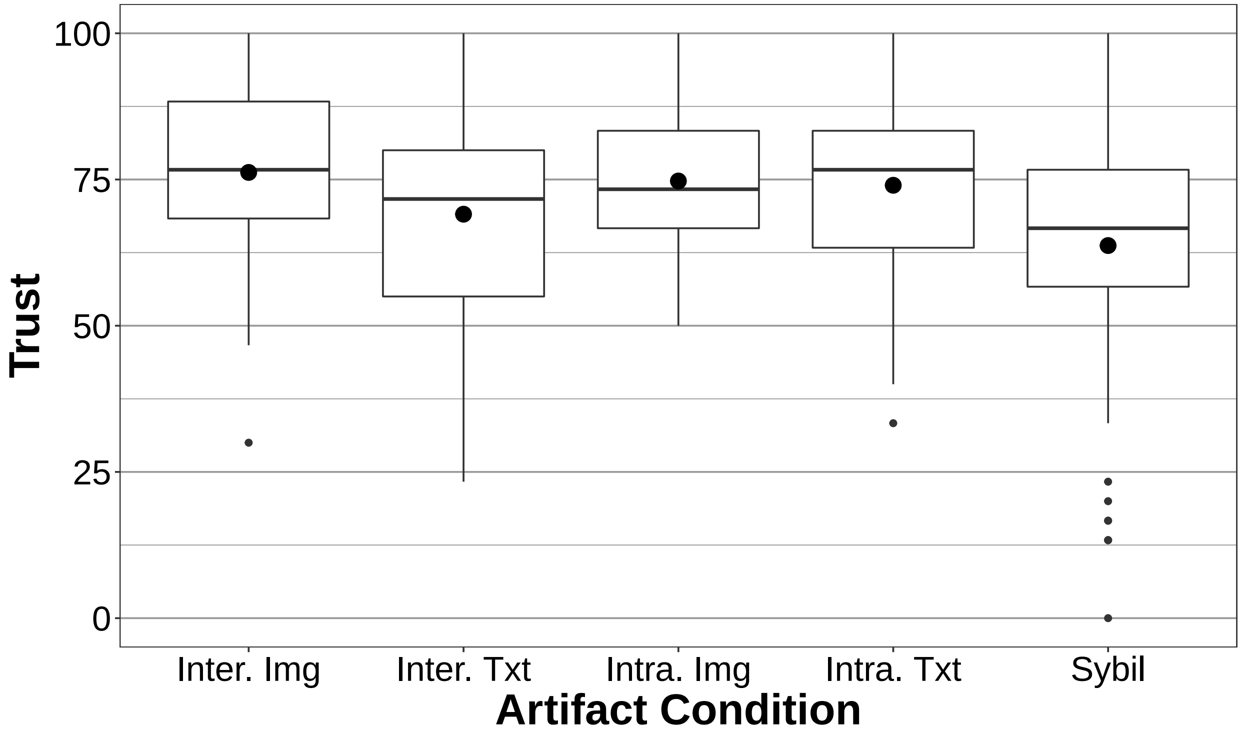
prompt	p_subcond	req	Percentage
Sybil-None	Inter. Img	Accept	88
Sybil-None	Inter. Txt	Accept	76
Sybil-None	Intra. Img	Accept	92
Sybil-None	Intra. Txt	Accept	96
Sybil-None	Sybil	Accept	66

Appendix B: Table 4

Variable	Estimate (Prob)	Std. Err. (Prob)	p-value
Intercept	1.176	0.583	0.044
Profile Type (Reference = Sybil)			
Inter Image	1.456	0.524	0.005**
Inter Text	0.449	0.420	0.285
Intra Image	1.977	0.615	0.001**
Intra Text	2.519	0.775	0.001**
Gender (Reference = Female)			
Male	0.240	0.344	0.485
Non-Binary	–	–	–
Age	-0.107	0.090	0.235
Generalized Trust	0.033	0.010	0.002**

Request Acceptance Analysis – Logistic mixed effects regression model. The unit for estimate and standard error is log odds scaled. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).

Appendix B: RESULT IS REFERENCED BUT NOT PROVIDED - Sybil trust plot



Descriptive statistics for Sybil trust plot

p__subcond	trust__mean	trust__standard__dev
Inter. Img	76.2	15.0
Inter. Txt	69.1	18.1

p_subcond	trust_mean	trust_standard_dev
Intra. Img	74.7	12.9
Intra. Txt	74.0	15.1
Sybil	63.7	20.9

Appendix B: RESULT IS REFERENCED BUT NOT PROVIDED - Sybil trust modeling

Variable	Estimate	Std. Err.	p
Intercept	64.885	3.536	<0.001***
Profile Type (Reference = Sybil)			
Inter Image	12.300	2.926	<0.001***
Inter Text	5.006	2.927	0.089
Intra Image	11.295	2.892	<0.001***
Intra Text	10.027	2.971	<0.001***
Gender (Reference = Female)			
Male	0.090	2.006	0.964
Non-Binary	–	–	–
Age	-0.205	0.531	0.701
Generalized Trust	0.272	0.062	<0.001***

Trust Analysis – Linear mixed effects regression model. The unit for estimate and standard error is log odds scaled. Significance is denoted by *** ($p < 0.001$), ** ($p < 0.01$), and * ($p < 0.05$).