When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models

Supplementary Material

AMY RECHKEMMER, Purdue University, USA

MING YIN, Purdue University, USA

1 ANOVA RESULTS

In addition to the confidence interval analysis presented in the paper, we have conducted two-way ANOVA (**RQ1** and **RQ2**) and three-way ANOVA (**RQ3**) analysis, and results are reported in Tables 1–3. Effect sizes and p-values are reported for each measure of trust with regards to each research question presented in the paper.

	belief		**agreement**		**switch**		**self-report**	
	η^2	p	η^2	p	η^2	p	η^2	p
RQ1: confidence	0.111	$< 0.001^{***}$	0.003	0.160	0.005	$0.080^{\dagger}$	0.000	0.973
RQ2: stated accuracy	0.016	$< 0.001^{***}$	0.007	0.035^{*}	0.008	0.032^{*}	0.035	$< 0.001^{***}$
RQ3: confidence × stated	0.005	$0.054^{\dagger}$	0.000	0.790	0.000	0.856	0.001	0.373

Table 1. Two-way ANOVA test results for Phase 1 data obtained from T1–T4 (i.e., treatments with an observed accuracy of 55% in Phase 1). η^2 reports the size of the effect. †, *, **, and *** represents the statistical significance level of 0.1, 0.05, 0.01, and 0.001, respectively.

	belief		**agreement**		**switch**		**self-report**	
	η^2	p	η^2	p	η^2	p	η^2	p
RQ1: confidence	0.054	$< 0.001^{***}$	0.006	0.049^{*}	0.001	0.412	0.000	0.782
RQ2: stated accuracy	0.020	$< 0.001^{***}$	0.005	$0.066^{\dagger}$	0.022	$< 0.001^{***}$	0.011	0.011^{*}
RQ3: confidence × stated	0.008	0.019^{*}	0.001	0.468	0.001	0.487	0.002	0.292

Table 2. Two-way ANOVA test results for Phase 1 data obtained from T5–T8 (i.e., treatments with an observed accuracy of 95% in Phase 1). η^2 reports the size of the effect. †, *, **, and *** represents the statistical significance level of 0.1, 0.05, 0.01, and 0.001, respectively.

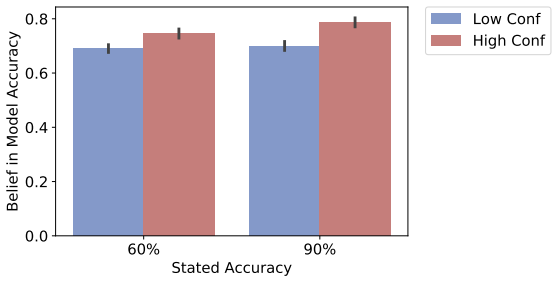
	belief		**agreement**		**switch**		**self-report**	
	η^2	p	η^2	p	η^2	p	η^2	p
RQ4: confidence	0.071	$< 0.001^{***}$	0.001	0.298	0.001	0.408	0.000	0.602
RQ5: stated accuracy	0.018	$< 0.001^{***}$	0.006	0.008^{**}	0.009	$< 0.001^{***}$	0.002	$0.079^{\dagger}$
RQ5: observed accuracy	0.044	$< 0.001^{***}$	0.029	$< 0.001^{***}$	0.126	$< 0.001^{***}$	0.125	$< 0.001^{***}$
RQ6: confidence × stated	0.003	0.029^{*}	0.000	0.999	0.000	0.747	0.001	0.384
RQ6: confidence × observed	0.001	0.251	0.001	0.435	0.000	0.711	0.003	$0.062^{\dagger}$
RQ6: stated × observed	0.000	0.728	0.001	0.209	0.010	$< 0.001^{***}$	0.009	$< 0.001^{***}$
RQ6: confidence × stated × observed	0.001	0.325	0.000	0.490	0.000	0.889	0.001	0.415

Table 3. Three-way ANOVA test results for Phase 2 data. η^2 reports the size of the effect. †, *, **, and *** represents the statistical significance level of 0.1, 0.05, 0.01, and 0.001, respectively.

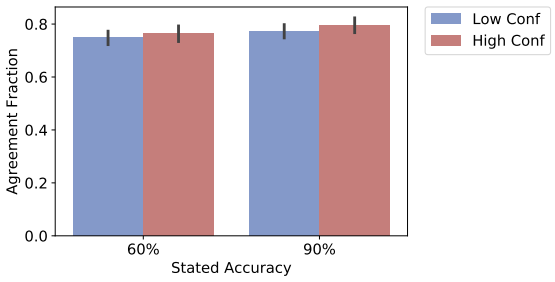
Authors' addresses: Amy Rechkemmer, Purdue University, West Lafayette, Indiana, USA, arechke@purdue.edu; Ming Yin, Purdue University, West Lafayette, Indiana, USA, mingyin@purdue.edu.

2 RAW DATA DISTRIBUTIONS

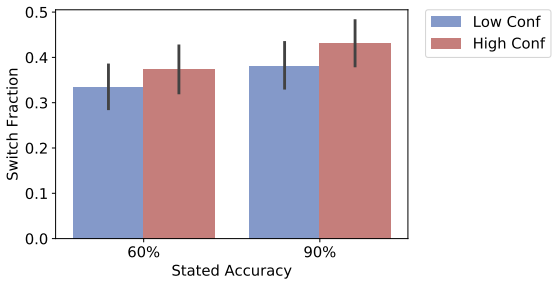
We have plotted the raw mean values for each measure of trust with regards to each experimental treatment, along with the 95% bootstrap confidence intervals. Figure 1 shows the mean values for treatments T1–T4 in Phase 1, and Figure 2 shows the mean values for treatments T5–T8 in Phase 1. Finally, Figure 3 shows the mean values for all treatments in Phase 2.



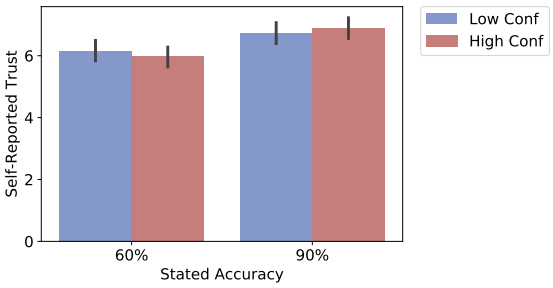
(a) Subject's belief in model accuracy



(b) Agreement fraction

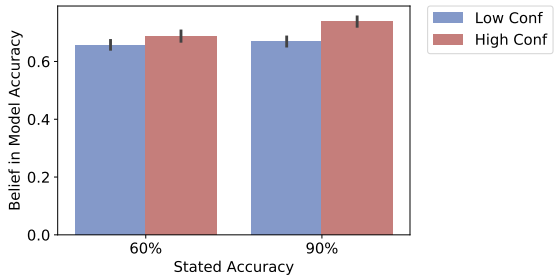


(c) Switch fraction

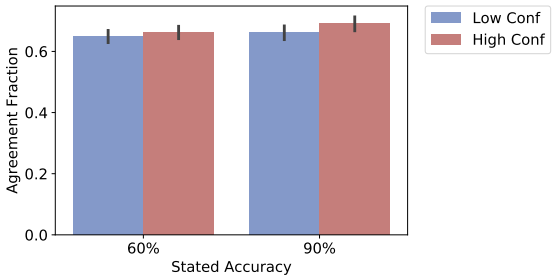


(d) Self-reported trust

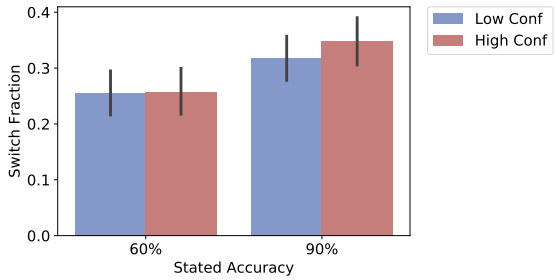
Fig. 1. Comparing how much subjects trust the ML model in Phase 1 (T1–T4). Average values of different trust measures are plotted for each treatment, and error bars represent the 95% bootstrap confidence intervals.



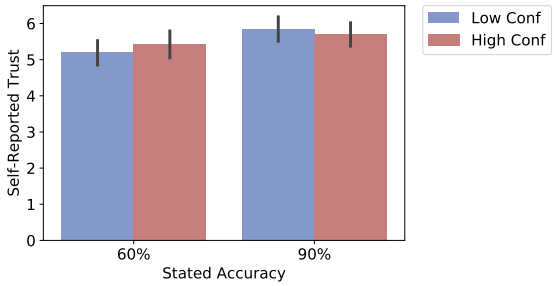
(a) Subject's belief in model accuracy



(b) Agreement fraction

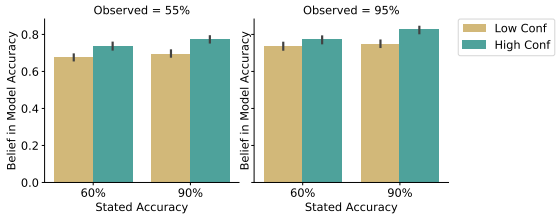


(c) Switch fraction

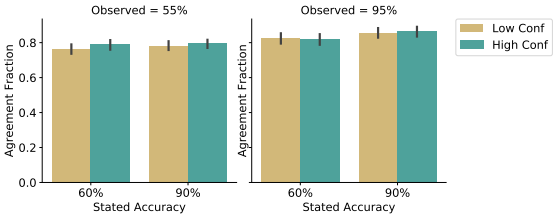


(d) Self-reported trust

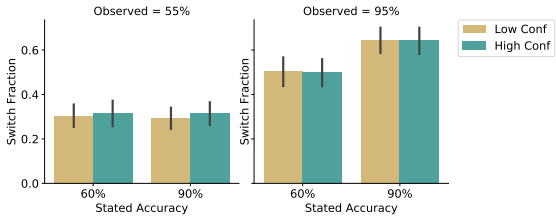
Fig. 2. Comparing how much subjects trust the ML model in Phase 1 (T5–T8). Average values of different trust measures are plotted for each treatment, and error bars represent the 95% bootstrap confidence intervals.



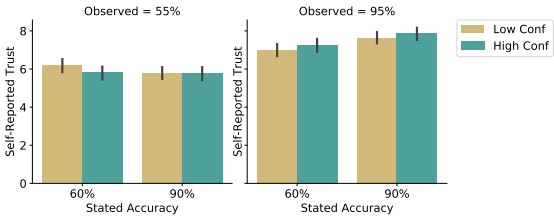
(a) Subject's belief in model accuracy



(b) Agreement fraction



(c) Switch fraction



(d) Self-reported trust

Fig. 3. Comparing how much subjects trust the ML model in Phase 2. Average values of different trust measures are plotted for each treatment, and error bars represent the 95% bootstrap confidence intervals.

3 ESTIMATED EFFECT SIZES AND THEIR 95% BOOTSTRAP CONFIDENCE INTERVALS

Cohen's d effect size values and their 95% bootstrap confidence intervals are summarized below for Phase 1 T1–T4 (Table 4), Phase 1 T5–T8 (Table 5), and Phase 2 (Table 6).

	belief		**agreement**		**switch**		**self-report**	
	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI
RQ1: confidence	0.72	[0.55, 0.89]	0.12	[-0.04, 0.28]	0.15	[-0.01, 0.31]	0.00	[-0.16, 0.16]
RQ2: stated accuracy	0.26	[0.09, 0.42]	0.17	[0.02, 0.34]	0.17	[0.01, 0.34]	0.37	[0.20, 0.53]

Table 4. Cohen's d values and their 95% bootstrap confidence intervals for Phase 1 data obtained from T1–T4 (i.e., treatments with an observed accuracy of 55% in Phase 1.)

	belief		**agreement**		**switch**		**self-report**	
	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI
RQ1: confidence	0.47	[0.31, 0.64]	0.16	[-0.01, 0.32]	0.06	[-0.10, 0.22]	0.01	[-0.14, 0.18]
RQ2: stated accuracy	0.29	[0.12, 0.44]	0.15	[-0.01, 0.31]	0.30	[0.14, 0.47]	0.21	[0.05, 0.38]

Table 5. Cohen's d values and their 95% bootstrap confidence intervals for Phase 1 data obtained from T5–T8 (i.e., treatments with an observed accuracy of 95% in Phase 1.)

	belief		**agreement**		**switch**		**self-report**	
	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI	Cohen's d	CI
RQ4: confidence	0.56	[0.44, 0.67]	0.06	[-0.05, 0.17]	0.04	[-0.08, 0.15]	0.02	[-0.09, 0.14]
RQ5: stated accuracy	0.27	[0.16, 0.38]	0.15	[0.04, 0.26]	0.19	[0.08, 0.31]	0.10	[-0.01, 0.21]
RQ5: observed accuracy	0.45	[0.33, 0.57]	0.35	[0.23, 0.46]	0.75	[0.63, 0.88]	0.77	[0.64, 0.90]

Table 6. Cohen's d values and their 95% bootstrap confidence intervals for Phase 2 data.