To Err is Human, To Correct is Algorithmic: People Trust Algorithms' Corrections More Than Humans' Corrections

Chengyao Sun
Yale University

Cynthia Cryder
Washington University in St. Louis

RESEARCH QUESTION

If an algorithmic task-performer and a human task-performer both make corrections following the same mistake, which corrected task-performer is more likely to be trusted for subsequent tasks?

MAIN FINDINGS

People trust algorithms' corrections more than humans' after they erred at similar levels.

STUDY 1: People continued to trust the algorithm, but lost confidence in the human judge, after they both made corrections following the same errors.

Design: 302 participants read one of three hypothetical scenario: **1)** an algorithm erred and made corrections; **2)** a human erred and made corrections; **3)** neither erred on the same task.

DV: Choice between the algorithm and its human counterpart.

Results: Following the same error, people maintained trust in the corrected algorithm (33% vs. 28%, p = .5) but lost confidence in the human after correction (56% vs. 28%, p < .01).

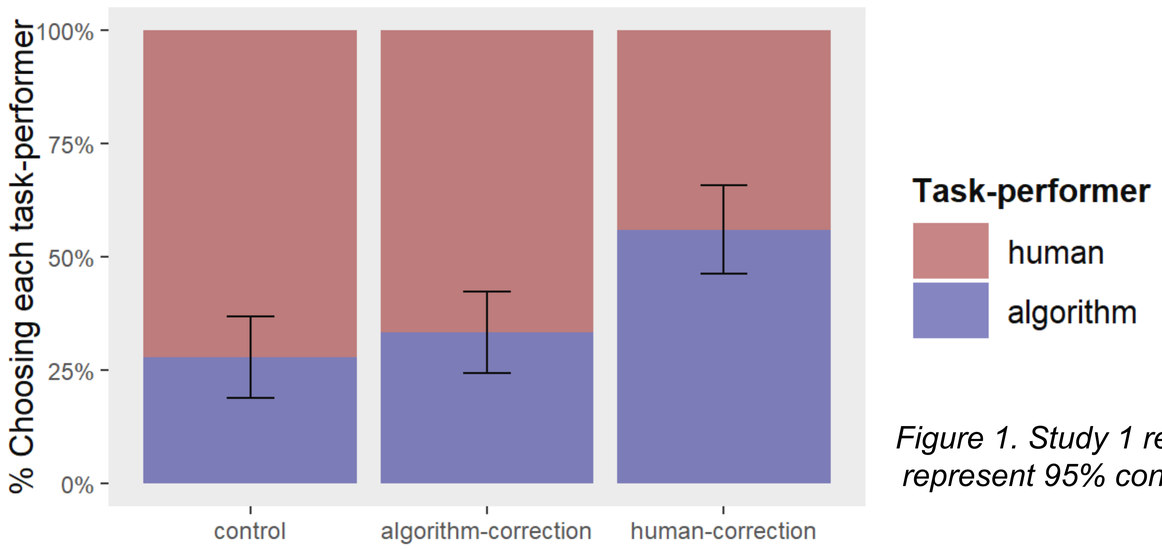

Figure 1. Study 1 results. Error bars represent 95% confidence intervals

STUDY 2: Testing Study 1's findings in different domains.

Design: 761 participants read a hypothetical scenario in 1 of 6 different domains. Each domain had the same design as Study 1.

DV: Choice between an algorithm and a human.

Results: Consistent with Study 1's

	Dependent variable
	Choice (1=Algorithm, 0=Human)
Stimulus fixed-effects	Included
Human correction	0.453*
	(0.186)
Algorithmic correction	0.486**
	(0.185)
N	761
Significance codes:	**p < 0.05, **p < 0.01, ***p < 0.001*

Table 1. Logistic regression results of Study 2. In parentheses are standard errors.

STUDY 3: Algorithms' correction trusted more than humans' in joint evaluation & with real incentivized behaviors

Design:

- 476 participants predicted the annual incomes of 10 U.S. residents and received bonus for accuracy.
- 3 between-subject conditions: they saw either **1)** an algorithm and themselves both perform and err, **2)** the algorithm and themselves both err but learn from errors, or **3)** neither err.

DV: Choice between themselves and the algorithm.

Results:

a) the algorithm outperformed participants in all three conditions;
b) people were averse to algorithm after seeing the algorithm err;
c) trust in the algorithm was restored after both humans and the algorithm could learn from errors, implying that people trust algorithm's correction more than their own correction;
d) relative performance did not attenuate their preference for algorithmic correction. (Relative performance = algorithm's accuracy – human's accuracy.)

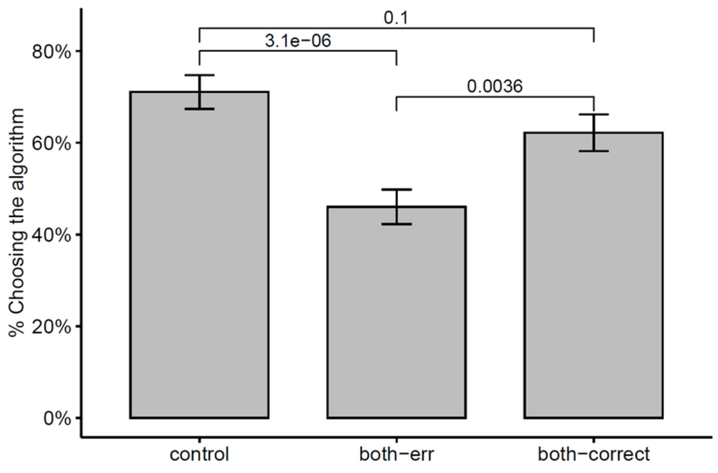

Figure 2. % choosing the algorithm. Error bars are 95% C.I. Above comparison bars are the p-values.

Figure 3. Predicted-likelihood plot of the logit model. X-axis is the relative performance.

	Dependent variable
	Likelihood of choosing the algorithm
Relative performance	0.640***
Correction	0.159***
Correction × relative performance	0.246
Constant	0.439***
Significance codes:	**p < 0.05, **p < 0.01, ***p < 0.001*

Table 2. Logit model's results of Study 3.