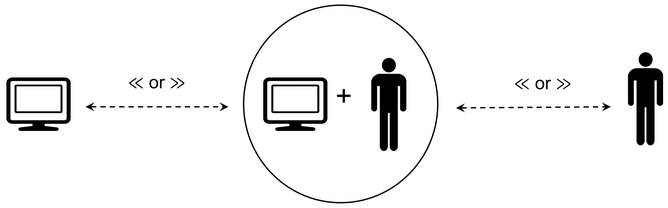
Understanding Human-AI Work Collaboration using a Randomized Field Study

Abhinav Maurya[1,2], Sunder Kekre[3], Rahul Telang[1,3]

[1]Heinz College of Information Systems and Public Policy, [2]Machine Learning Department, [3]Tepper School of Business, Carnegie Mellon University

Key Findings



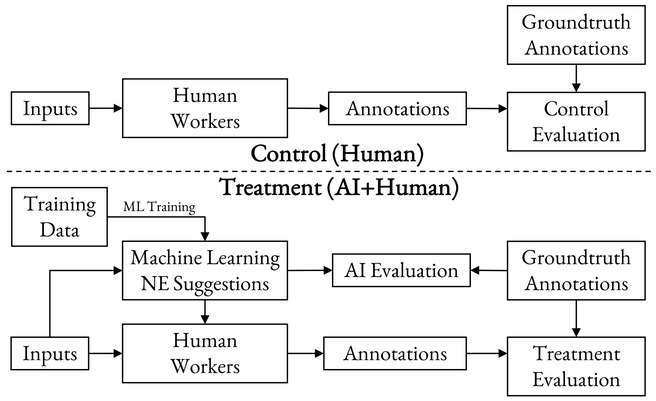
Human+AI team can exploit **complementary strengths** of both humans and machines and surpass either.

Statistically significant **improvements in work quality; improved worker productivity.**

Labor Specialization: AI does better on lower-complexity work; humans better in more complex regimes.

Evidence for J.C.R. Licklider's hypothesis that **human-machine symbiosis** can harness the "cognitive" capabilities of both humans and machines optimally.

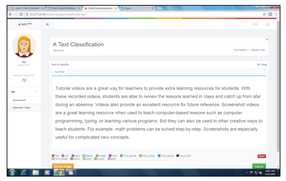
Designing a Study to Compare AI, Human, and AI+Human Team

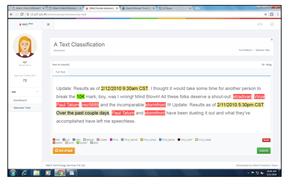


Running The 3-Day Experiment on Named Entity Annotation

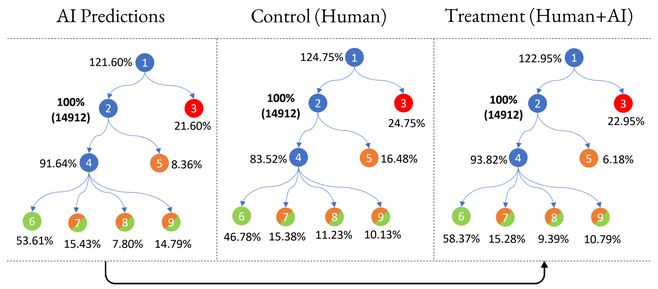
Control group (15 participants who carried out 2975 NE tasks overall)

Treatment group (15 participants who carried out the same 2975 tasks as control group)





Error Decomposition Trees for AI, Control, and Treatment



Treatment introduced fewer new mistakes/misses than it corrected for AI. However, treatment added more false positives than machine.

TP True Positives FP False Positives FN False Negatives

	Precision		Recall		F1-Score	
	Control	Treatment	Control	Treatment	Control	Treatment
Mean	0.6409	0.7146	0.4319	0.5634	0.5132	0.6290
Variance	0.0039	0.0011	0.0037	0.0022	0.0029	0.0013
Observations	15	15	15	15	15	15
Treatment - Control	0.0737		0.1315		0.1158	
t-statistic	4.0533		6.6492		6.9819	
P(T≤t) (one-sided)	0.0003 (***)		0.0000 (***)		0.0000 (***)	

Treatment group outperforms control group, differences in quality metrics are statistically significant.

Comparing AI, Control, and Treatment

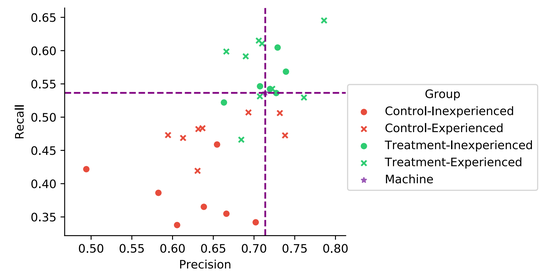
Work Quality

Metric	AI	Control	Treatment
Number of Participants	–	15	15
Total Documents Tagged	2975	2975	2975
Precision	71.38	65.40	**71.78**
Recall	53.65	46.78	**58.37**
F1-Score	61.26	54.54	**64.38**
Precision (non-strict)	**80.93**	77.14	80.35
Recall (non-strict)	91.64	83.52	**93.82**
F1-Score (non-strict)	85.95	80.20	**86.56**
Precision (token-level)	51.07	59.97	**62.21**
Recall (token-level)	42.50	57.29	**60.96**
F1-Score (token-level)	46.39	58.60	**61.58**

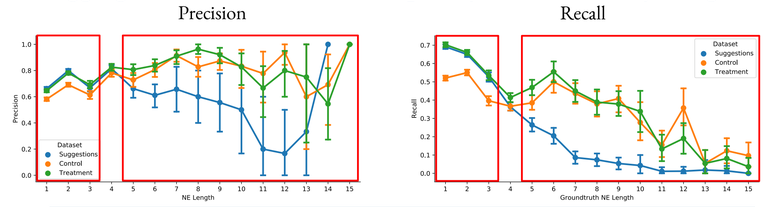
Work Productivity

Metric	Control	Treatment
Number of Participants	15	15
Total Documents Tagged	2975	2975
Median time (minutes/task)	3.37	**3.05**
Average time (minutes/task)	5.16	**4.73**
Throughput (tasks/hour)	11.63	**12.70**
Throughput (entities/hour)	62.71	**74.29**
True positive entities per hour	40.04	**52.83**
Google searches (per task)	6.72	**5.75**

Treatment Workers Do Better Than Controls on Precision and



Tagging Correctness versus Complexity (Named Entity Length)



Machine outperforms for NE length<4, treatment follows machine.
For length>4, human cognition (control/treatment) outperforms machine.

Conclusions and Future Work

- Studied mediating factors of **Human-AI collaboration in NLP data annotation services.**
- Potential **generalization beyond NLP** e.g. computer vision where work complexity in object detection might be measured by bounding box complexity.
- **Theory** for what team construction regimes are better.
- Investigating **pathological regimes** of Human-AI team dynamics.
- **Engendering trust** in AI agents in Human-AI teams.
- **Multistage work and cooperation** in Human-AI teams.

Working Paper

http://bit.ly/maurya2019understanding