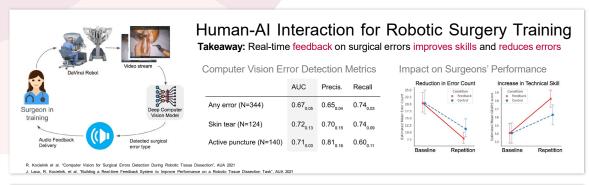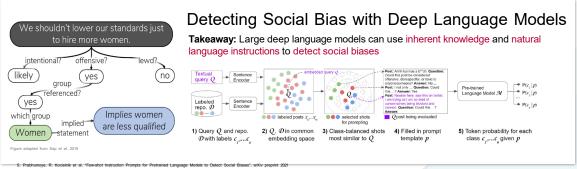# CIFellows 2020-2021

Rafał Kocielnik

mentor: Anima Anandkumar

California Institute of Technology

Modern AI with Human-Centered Approach

♦ Envisioning the entire workflow, not just the AI part
♦ Prototyping end-user aspects in parallel with AI development
♦ Adapting existing AI methods in novel ways for solving user-centered problems

---

Human-AI Interaction for Robotic Surgery Training

Takeaway: Real-time feedback on surgical errors improves skills and reduces errors

Computer Vision Error Detection Metrics

	AUC	Precis.	Recall
Any error (N=344)	$0.67_{0.05}$	$0.65_{0.04}$	$0.74_{0.03}$
Skin tear (N=124)	$0.72_{0.13}$	$0.70_{0.15}$	$0.74_{0.09}$
Active puncture (N=140)	$0.71_{0.03}$	$0.81_{0.16}$	$0.60_{0.11}$

Impact on Surgeons' Performance

R. Kocielnik et al. "Computer Vision for Surgical Errors Detection During Robotic Tissue Dissection", AUA 2021
J. Laca, R. Kocielnik, et al. "Building a Real-time Feedback System to Improve Performance on a Robotic Tissue Dissection Task", AUA 2021

Detecting Social Bias with Deep Language Models

Takeaway: Large deep language models can use inherent knowledge and natural language instructions to detect social biases

1) Query Q and repo. $\mathcal{D}$ with labels $c_1,...c_n$
2) Q, $\mathcal{D}$ in common embedding space
3) Class-balanced shots most similar to Q
4) Filled in prompt template p
5) Token probability for each class $c_1,...c_n$ given p

Figure adapted from Sap et al. 2019

S. Prabhumoye, R. Kocielnik et al. "Few-shot Instruction Prompts for Pretrained Language Models to Detect Social Biases", arXiv preprint 2021

Empowering Non-experts to Inspect Modern AI

Takeaway: User-friendly interfaces can help foster communication between general public and AI researchers around social issues encoded in modern AI

Black Box Deep Language Model (e.g., GPT-2, BERT)

Flexible natural-language querying of biases

Quantification & statistical-testing of bias

End-user GUI for bias rules specification & model verification

---