

Human Interaction with ML and AI: User Evaluation Challenges

Responsible ML for Healthcare Workshop
28 Oct 2022

Enrico Costanza
UCL Interaction Centre

- HCI - “A discipline concerned with the design, evaluation and implementation of interactive computing systems for human use and with the study of major phenomena surrounding them.” (ACM)
- How do we design interfaces or interactions around AI and ML?
- How do we evaluate them?
- Critical to achieve responsible ML in healthcare

HCI as a Broad Multi-disciplinary Area

- Technology: electronics & computer science
- Design
- Cognitive Psychology
- Social Science



(Image from Sharp, Rogers, Preece, 2002)

“Can’t we simply ask people?”

- Why is it a challenge to design user studies?
- Why do we need to borrow research methods from psychology and social science?
- Is it not just about simple questionnaires?

- Introduction
- Human Bias around AI
- Comparing Computer Vision Feedback Strategies
- Classification confidence information
- Evaluating CNN explanations
- Some work in progress

- Introduction
- **Human Bias around AI**
- Comparing Computer Vision Feedback Strategies
- Classification confidence information
- Evaluating CNN explanations
- Some work in progress

Cognitive Bias on AI Performance

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, naging gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero, naging gobernador ng Zamboanga. Nagkaroon ng misyon ong mga Heswita na pabalikin si Rizal

- No-animation Condition

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, naging gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero, naging gobernador ng Zamboanga. Nagkaroon ng misyon ong mga Heswita na pabalikin si Rizal

- Animation Condition

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, **naging** gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

- Animation Condition

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, **naging** gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

- Which system works better?

- Animation Condition

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

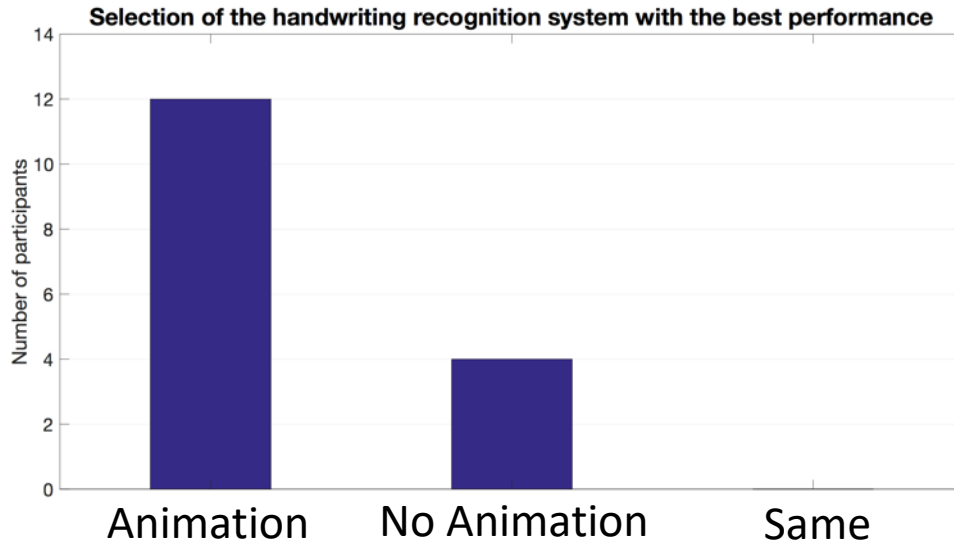
maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, **naging** gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

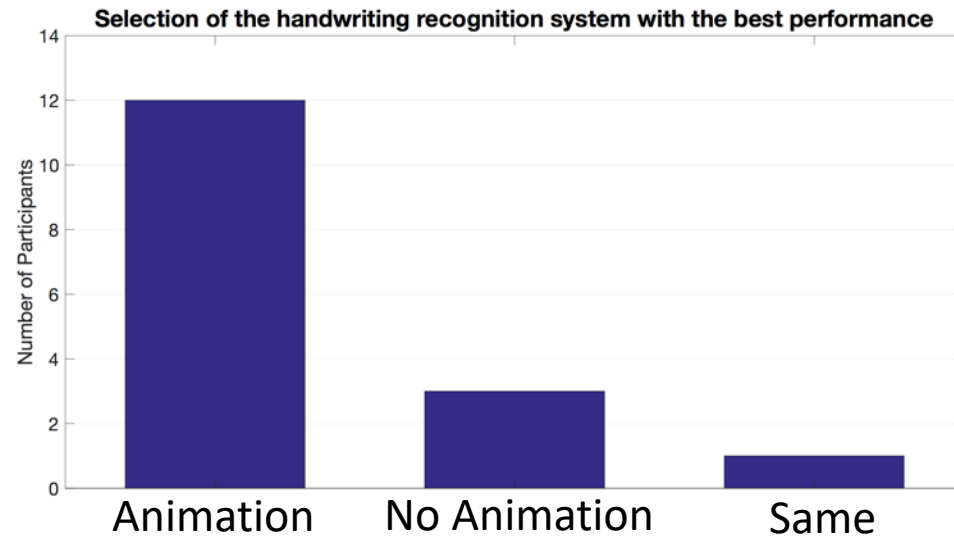
nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

- ~~Which system works better?~~
- Which one will people choose as working better?



16 participants in the lab



16 participants on m-turk

Explanation: Human-like Qualities?

Explanation: Human-like Qualities?

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, naging **gobemador** ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

Explanation: Human-like Qualities? No!

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

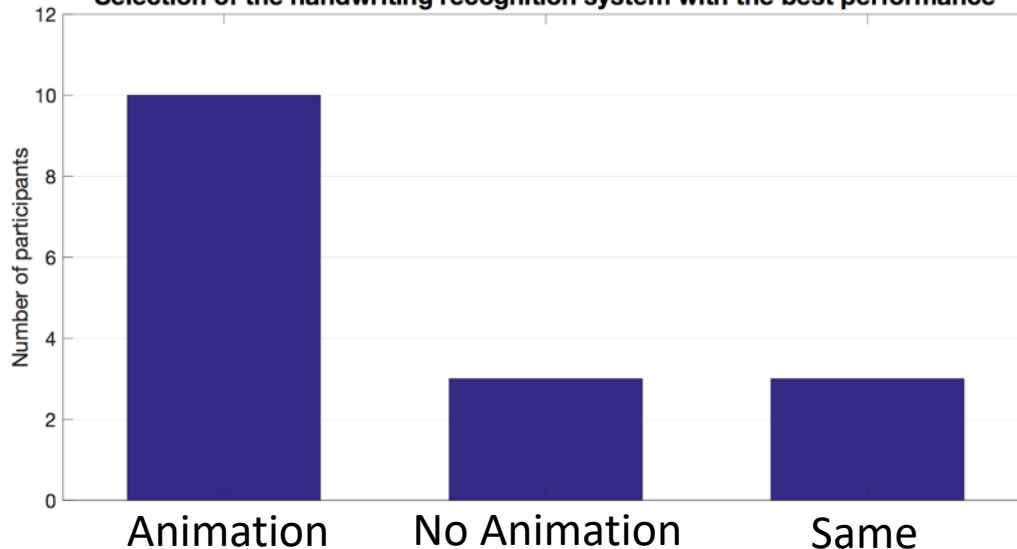
maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, naging gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaan sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

Selection of the handwriting recognition system with the best performance

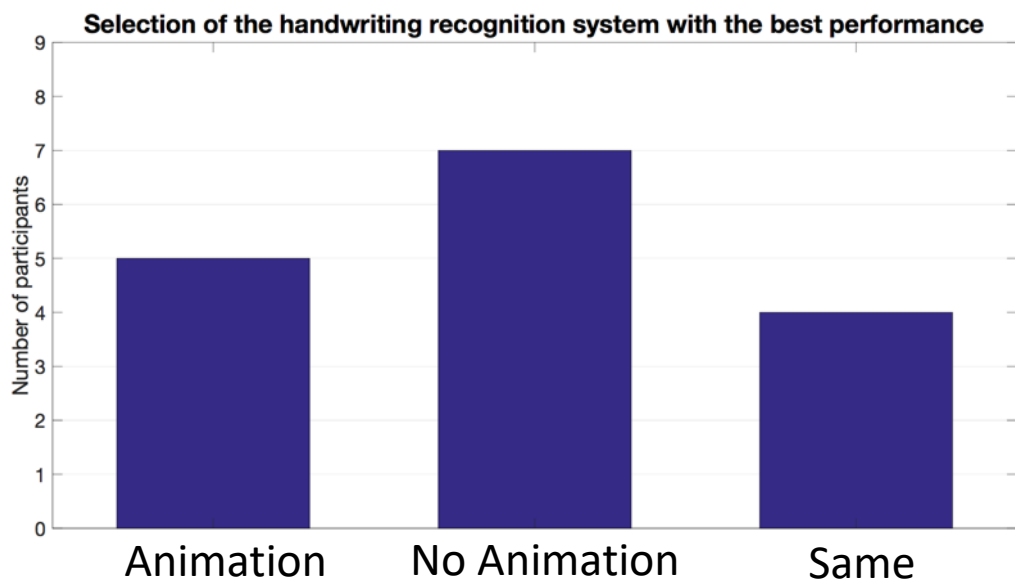


16 participants on m-turk

- Pre-study question eliminates the effect
 - “You are about to evaluate a hand-writing recognition system – how do you think this kind of systems work?”

Mental Model Influence?

- Pre-study question eliminates the effect
 - “You are about to evaluate a hand-writing recognition system – how do you think this kind of systems work?”



16 participants on m-turk

Alternative Animation

Original version

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, **naging** gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

Alternative version

Handwriting Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtuturo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang layunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pagiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga magsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging datu, at isa pa, si Jose Aseniero, **naging** gobernador ng Zamboanga. Nagkaroon ng misyon ang mga Heswita na pabalikin si Rizal

E-Text

nagturo din ng pagsasaka. Nagtayo si Rizal ng paaralan para sa mga batang lalaki. Sa paaralang ito, wikang Kastila ang ginagamit sa pagtutuyo, at nagtuturo din ito ng Ingles bilang wikang banyaga. Ang loyunin ng paaralang ito ay upang turuan ang mga mag-aaral ng pogiging

maparaon sa buhay. Ang ilan sa mga mag-aaral ay naging matagumpay bilang mga mogsasaka at tapat na opisyal ng pamahalaan. Isang Muslim ang naging dotu, at isa pa, si Jose Aseniero,

Original animation's explanation

Instructions

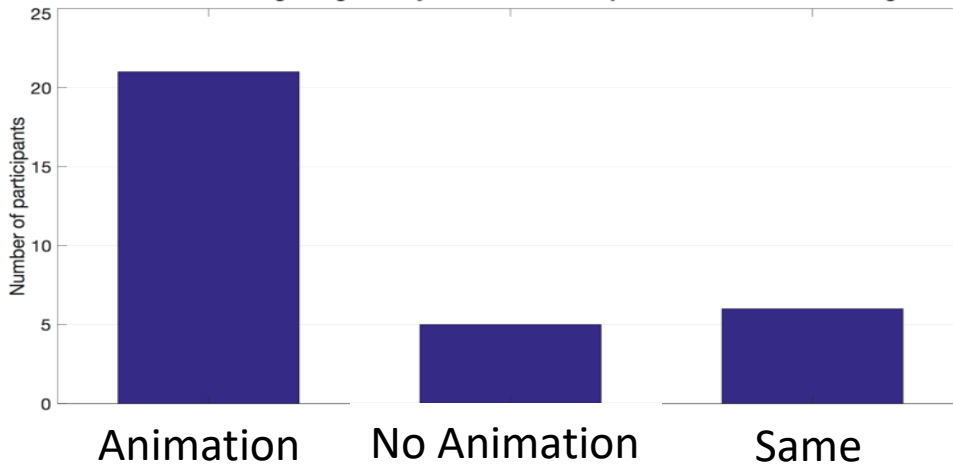
- * The computer programs that we use for this experiment will convert handwritten text to e-text, like the one above.
- * **How such a system works:** First a program needs to identify where a word is and then highlight the contour of the word. Once the program had highlighted the word, is possible for it to identify the characters of the word and then write into the e-text.

Alternative animation's explanation

Instructions

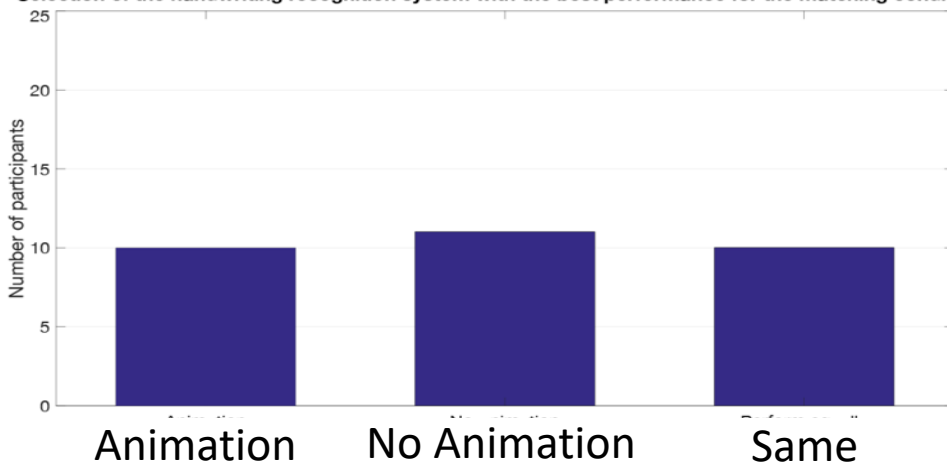
- * The computer programs that we use for this experiment will convert handwritten text to e-text, like the one above.
- * **How such a system works:** First a computer program needs to identify where a word is and then switch the colour of the ink with the colour of the paper. Once the program inverted the colours, it will be able to identify the characters and then write the e-text.

Selection of the handwriting recognition system with the best performance for the matching conditions



Matching conditions
32 participants on m-turk

Selection of the handwriting recognition system with the best performance for the matching conditions



Mismatching conditions
32 participants on m-turk

- Introduction
- Human Bias around AI
- **Comparing Computer Vision Feedback Strategies**
- Classification confidence information
- Evaluating CNN explanations
- Some work in progress

- How to guide user interaction with image recognition applications?
- Are keypoint markers helpful?

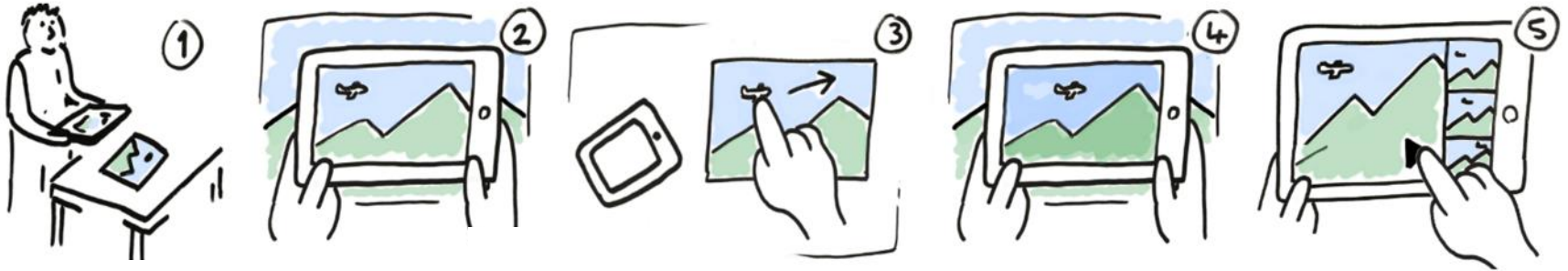


Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, Seb Stein, [Evaluating the Effect of Feedback from Different Computer Vision Processing Stages: a Comparative Lab Study](#), In Proceedings CHI 2019

- Feedback is especially useful when things go wrong
- Challenges of controlled, non-obvious failure

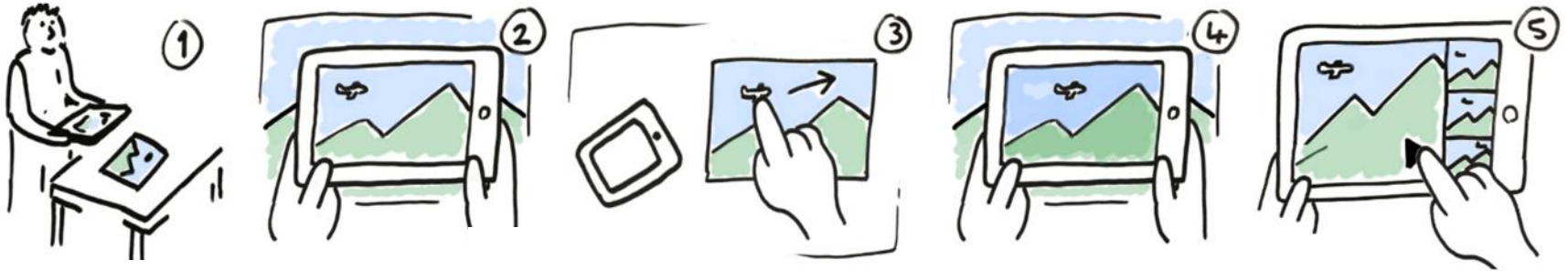
Setting Users up to Fail

- Feedback is especially useful when things go wrong
- Challenges of controlled, non-obvious failure
- An app for stop motion animations without tripod

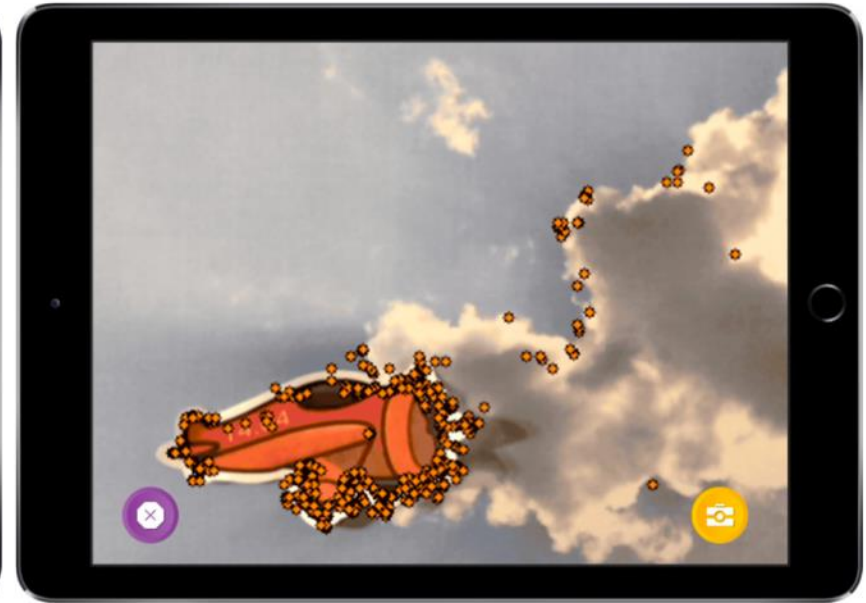


Setting Users up to Fail

- Feedback is especially useful when things go wrong
- Challenges of controlled, non-obvious failure
- An app for stop motion animations without tripod



Keypoint Markers vs No-feedback

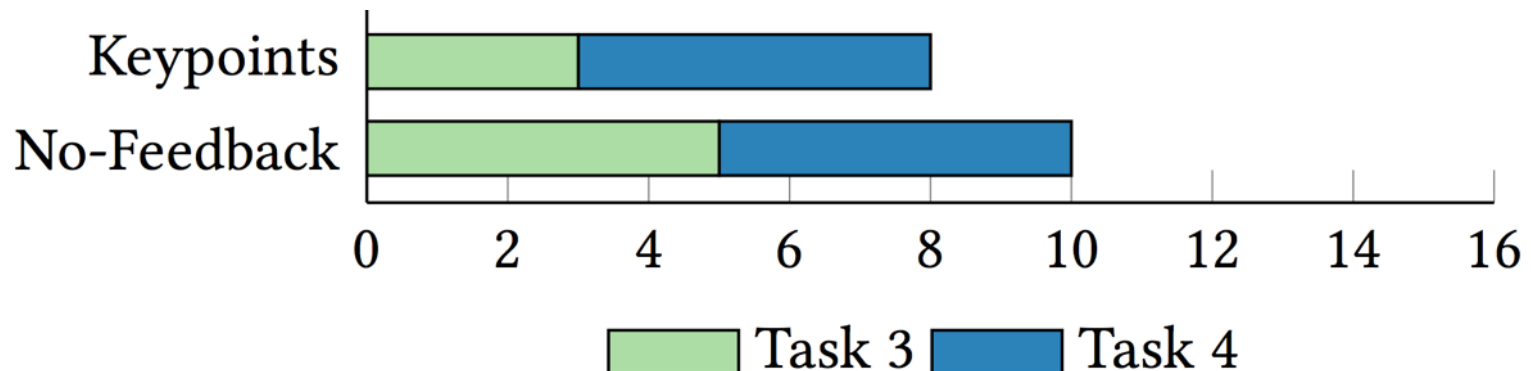


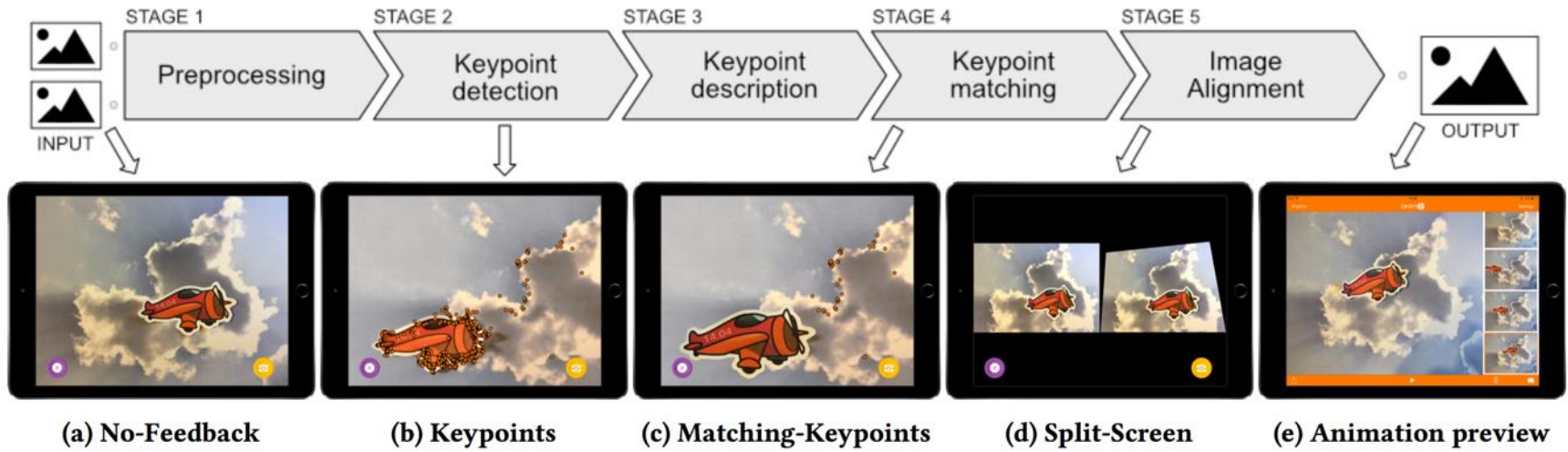
Keypoint Markers vs No-feedback



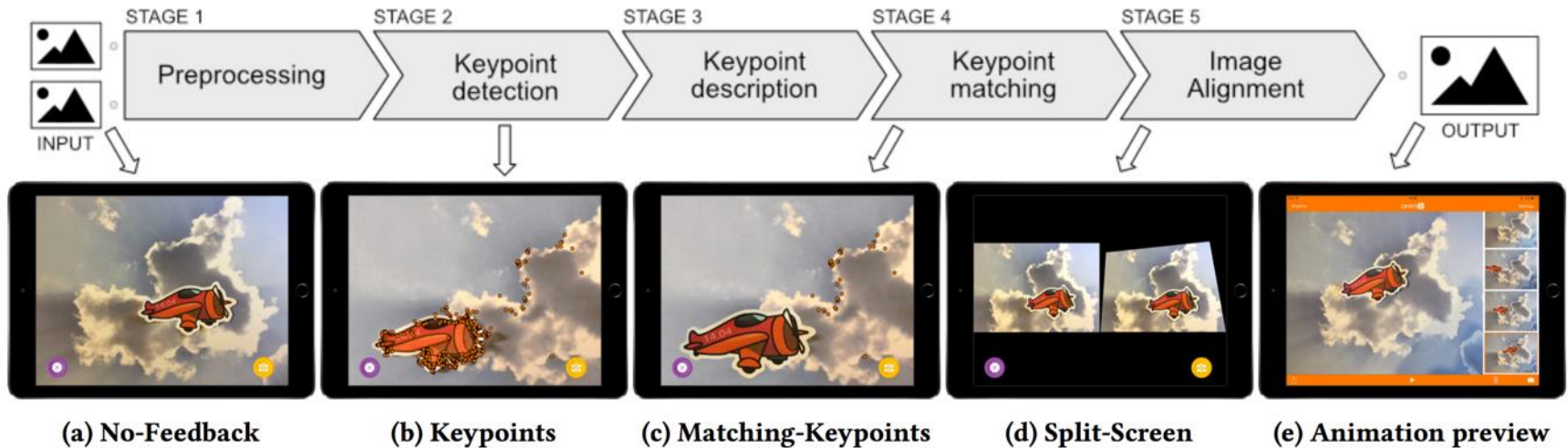
- Keypoint markers do not seem to help
- Users seem expect *higher-level* information

Number of participants demonstrating correct understanding of the system

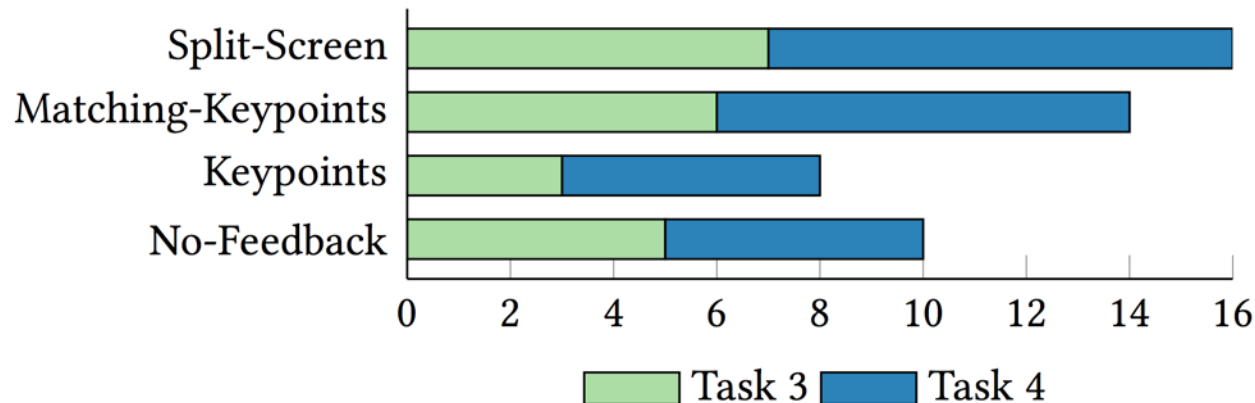




Higher-level Feedback Works Better



Number of participants demonstrating correct understanding of the system



- Introduction
- Human Bias around AI
- Comparing Computer Vision Feedback Strategies
- **Classification confidence information**
- Evaluating CNN explanations
- Some work in progress

- Some ML / pattern recognition algorithms can provide *confidence information* about their inferences
- Can this information help users? Or is it just “noise”?

Jhim Verame, Enrico Costanza, Sarvapali Ramchurn, The effect of displaying system confident information on the usage of autonomous systems for non-specialist applications: a lab study, In Proceedings CHI 2016

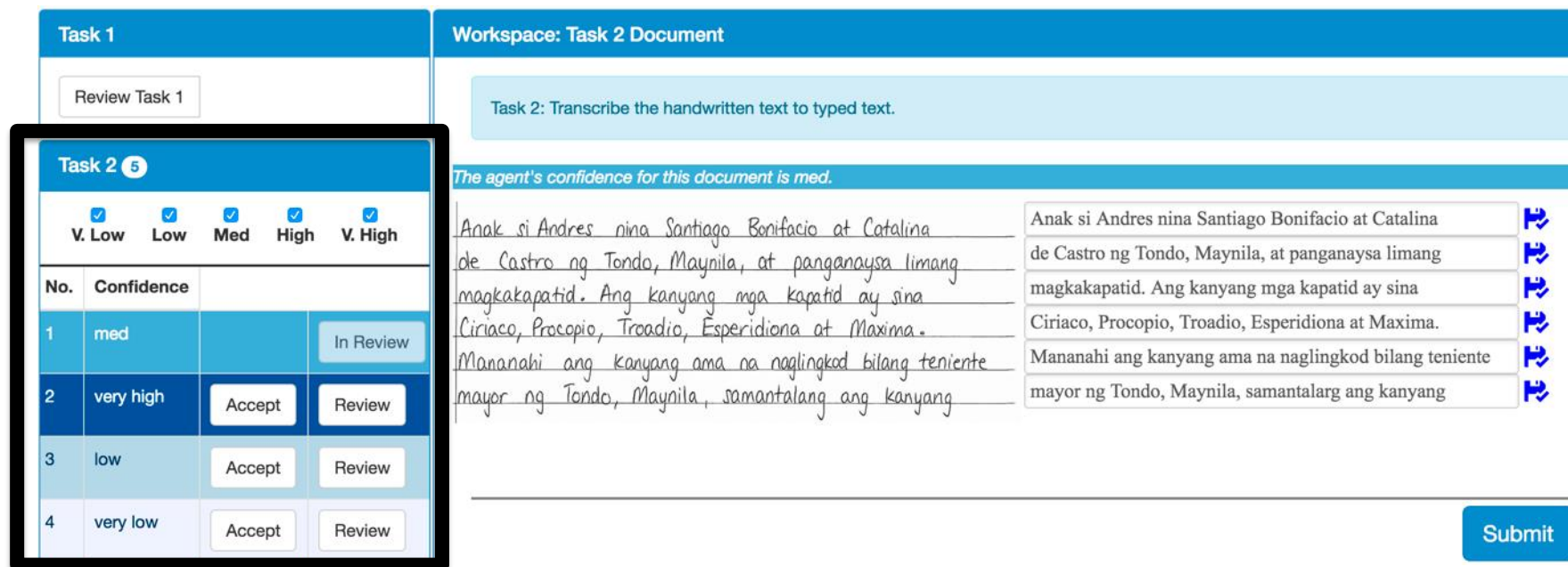
- Participants were offered to perform a number of micro-tasks for pay (somewhat similar to m-turk)
- Two types of tasks:
 - A. Fix grammatical mistakes in text
 - B. Check and fix output from *hand-writing recognition system*

- Participants were offered to perform a number of micro-tasks for pay (somewhat similar to m-turk)
- Two types of tasks:
 - A. Fix grammatical mistakes in text – **manual**
 - B. Check and fix output from *hand-writing recognition system* – **AI assisted**

- Participants were offered to perform a number of micro-tasks for pay (somewhat similar to m-turk)
- Two types of tasks:
 - A. Fix grammatical mistakes in text – **manual**
 - B. Check and fix output from *hand-writing recognition system* – **AI assisted**
- Which tasks will participants favour?
 - Does confidence information make a difference?

Confidence Condition

- Confidence information from the handwriting recognition algorithm shown
 - Five confidence levels: from very low to very high



The screenshot displays a workspace for 'Task 2 Document' with the instruction 'Task 2: Transcribe the handwritten text to typed text.' The interface shows a confidence condition table for 'Task 2' and a list of transcribed text segments with their corresponding confidence levels.

Task 2 Confidence Condition Table:

| No. | Confidence | Accept | Review |
|-----|------------|--------|-----------|
| 1 | med | | In Review |
| 2 | very high | Accept | Review |
| 3 | low | Accept | Review |
| 4 | very low | Accept | Review |

Transcribed Text and Confidence Levels:

The agent's confidence for this document is med.

| | |
|--|------|
| Anak si Andres nina Santiago Bonifacio at Catalina | High |
| de Castro ng Tondo, Maynila, at panganaysa limang | High |
| magkakapatid. Ang kanyang mga kapatid ay sina | High |
| Ciriaco, Procopio, Troadio, Esperidiona at Maxima. | High |
| Mananahi ang kanyang ama na naglingkod bilang teniente | High |
| mayor ng Tondo, Maynila, samantalarg ang kanyang | High |

Submit

No-confidence condition

Task 1

Review Task 1

Task 2 5

| No. | | |
|-----|--------|-----------|
| | | In Review |
| | Accept | Review |
| 3 | Accept | Review |
| 4 | Accept | Review |

Workspace: Task 2 Document

Task 2: Transcribe the handwritten text to typed text.

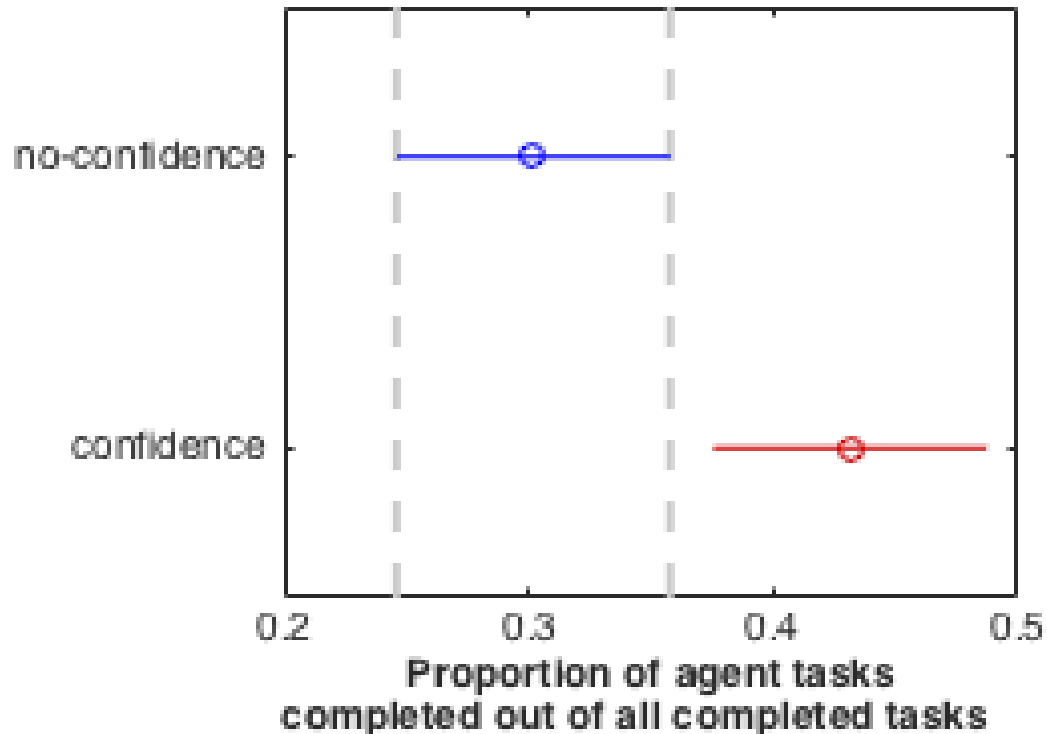
Anak si Andres nina Santiago Bonifacio at Catalina de Castro ng Tondo, Maynila, at panganaysa limang magkakapatid. Ang kanyang mga kapatid ay sina Ciriaco, Procopio, Troadio, Esperidiona at Maxima. Mananahi ang kanyang ama na naglingkod bilang teniente mayor ng Tondo, Maynila, samantalang ang kanyang

Anak si Andres nina Santiago Bonifacio at Catalina de Castro ng Tondo, Maynila, at panganaysa limang magkakapatid. Ang kanyang mga kapatid ay sina Ciriaco, Procopio, Troadio, Esperidiona at Maxima. Mananahi ang kanyang ama na naglingkod bilang teniente mayor ng Tondo, Maynila, samantalang ang kanyang



Submit

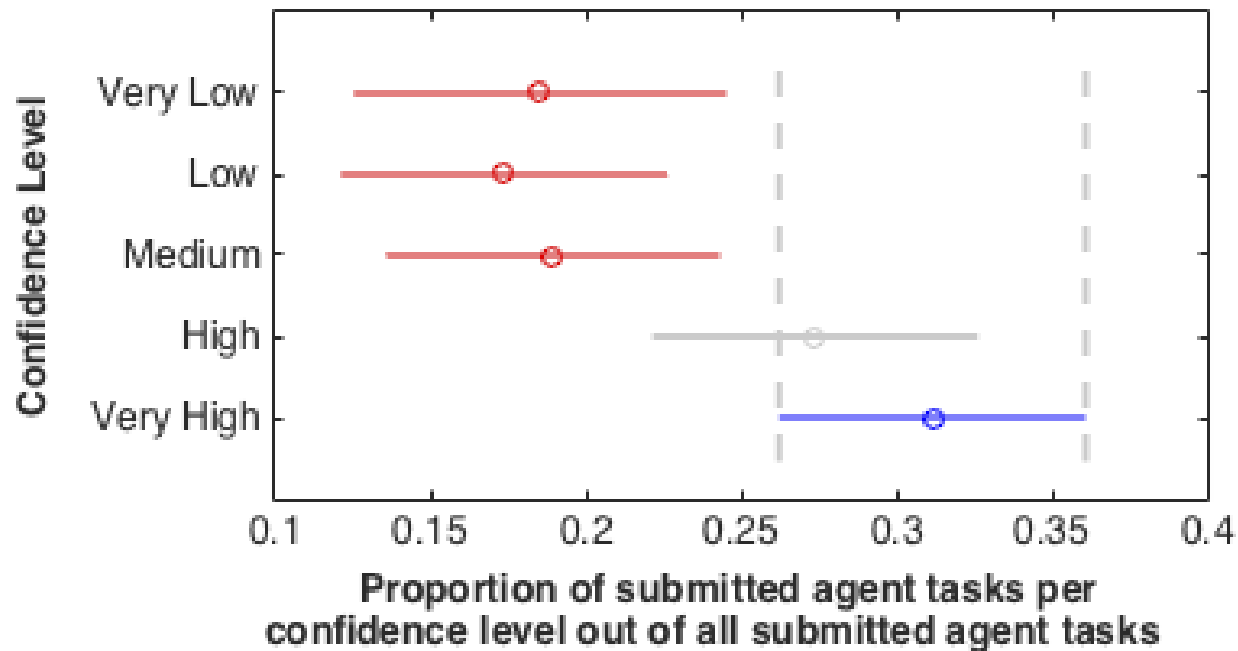
- More AI-related tasks were completed when confidence information was available than when it was not



Means comparison for agent tasks completed across different displays of confidence information with the 95% confidence bars (Tukey-HSD).

How Confidence Guides Choice

- Tasks with *very high* confidence level were completed more often than those with lower confidence levels



Means comparison for completed agent tasks per confidence level across all confidence levels

- Introduction
- Human Bias around AI
- Comparing Computer Vision Feedback Strategies
- Classification confidence information
- **Evaluating CNN explanations**
- Some work in progress

- Saliency Maps have been proposed as an explanation technique for CNNs
 - Red indicates pixels most responsible for the classification
- How well do they work?



Measuring Saliency Maps' Performance

- Can Saliency Maps help participants predict the behaviour of a CNN?

TP True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**

TP True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**

FP False Positive: Examples for mistakes because there is no **horse** in the images, but the system incorrectly predicted the label **horse**

FN False Negatives: Examples for mistakes because there is a **horse** in each image, but the system failed to predict the label **horse**

Examples

Task :
What features do you think this system is sensitive to? What features do you think this system ignores?

Please name 2-3 features you think the system is **sensitive to**:

Please name 2-3 features you think the system **ignores**:

Note, In the heatmap images:
Red: highlights the parts that look like **horse**
Blue: highlights the parts that do NOT look like **horse**

Next

Based on what you have learned from the examples shown to you do you think the system is going to recognize a **horse** this image?

Yes No

How confident are you in your answer?

extremely unconfident slightly unconfident slightly confident extremely confident

Questions

Four Experimental Conditions

- Different groups of participants experienced different conditions:
 - Saliency Map & Classification Scores
 - Saliency Map & no Classification Scores
 - No Saliency Map & Classification Scores
 - No Saliency Map & no Classification Scores

TP

True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**



Four Experimental Conditions

- Different groups of participants experienced different conditions:
 - Saliency Map & Classification Scores
 - **Saliency Map & no Classification Scores**
 - No Saliency Map & Classification Scores
 - No Saliency Map & no Classification Scores

TP

True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**



Four Experimental Conditions

- Different groups of participants experienced different conditions:
 - Saliency Map & Classification Scores
 - Saliency Map & no Classification Scores
 - No Saliency Map & Classification Scores
 - No Saliency Map & no Classification Scores

TP

True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**



Four Experimental Conditions

- Different groups of participants experienced different conditions:
 - Saliency Map & Classification Scores
 - Saliency Map & no Classification Scores
 - No Saliency Map & Classification Scores
 - No Saliency Map & no Classification Scores

TP

True Positives: Examples of correct decisions, because there is a **horse** in each image and the system correctly predicted the label **horse**

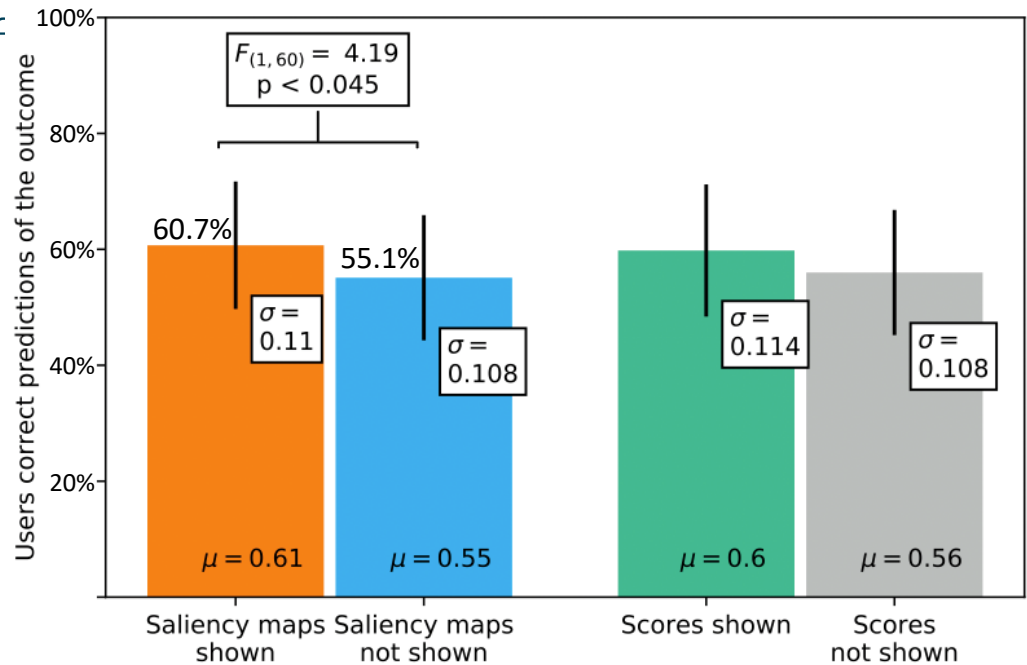


Running the Study + Results

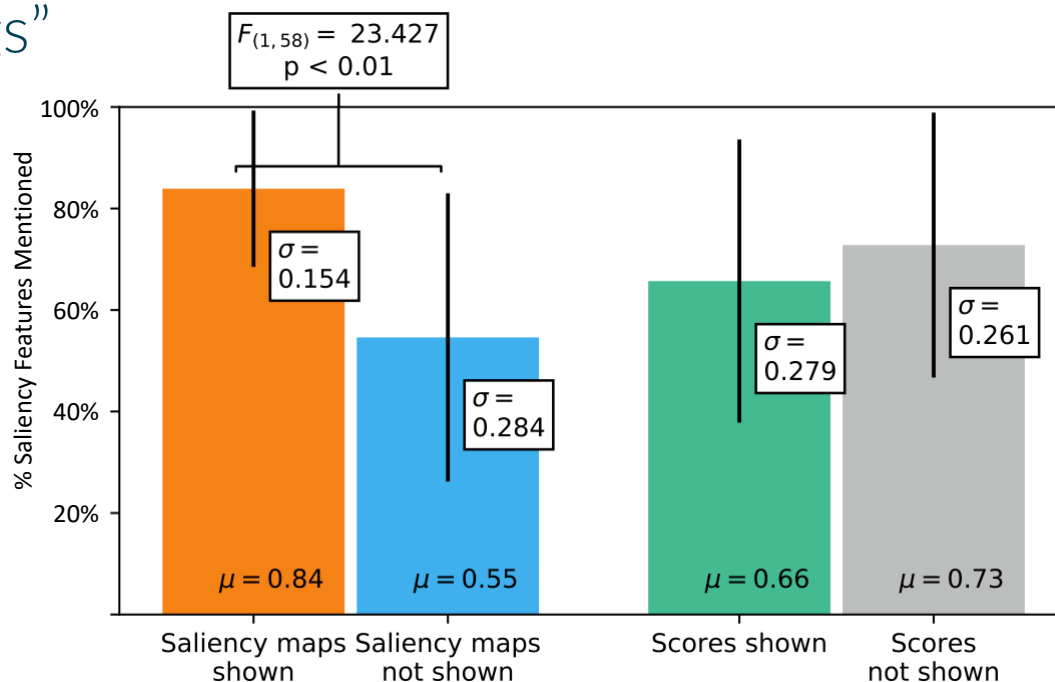
- Study implemented as interactive Web application
- Recruited 64 participants on the “Prolific” platform

– Can recruit tens of

- Main result:
When saliency maps were shown, participants were significantly more accurate in predicting the CNN's outcome
-- *still not much better*

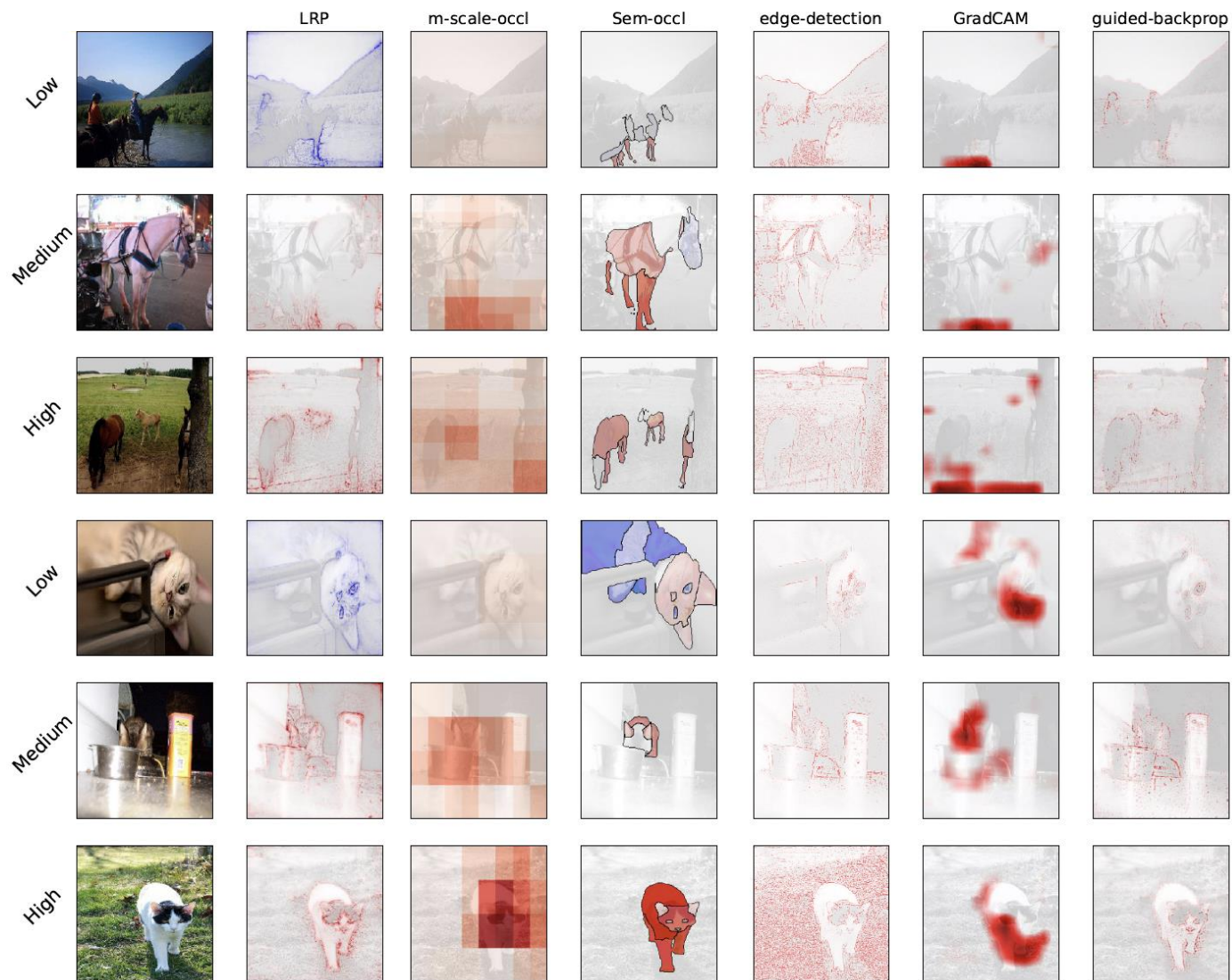


- When Saliency Maps were shown, participants referred more frequently to features localized to pixels around the objects of interest
 - E.g. “eyes” or “legs” rather than “size” or “image quality”



- Introduction
- Human Bias around AI
- Comparing Computer Vision Feedback Strategies
- Classification confidence information
- Evaluating CNN explanations
- **Some work in progress**

W-i-P: Saliency Map Follow-up Study



- Evaluating how specialist optometrist interpret deep learning results for ambiguous or challenging cases
 - Applying a model published by others

ARTICLES

<https://doi.org/10.1038/s41591-018-0107-6>

nature
medicine

Clinically applicable deep learning for diagnosis and referral in retinal disease

Jeffrey De Fauw¹, Joseph R. Ledsam¹, Bernardino Romera-Paredes¹, Stanislav Nikolov¹, Nenad Tomasev¹, Sam Blackwell¹, Harry Askham¹, Xavier Glorot¹, Brendan O'Donoghue¹, Daniel Visentin¹, George van den Driessche¹, Balaji Lakshminarayanan¹, Clemens Meyer¹, Faith Mackinder¹, Simon Bouton¹, Kareem Ayoub¹, Reena Chopra², Dominic King¹, Alan Karthikesalingam¹, Cian O. Hughes^{1,3}, Rosalind Raine³, Julian Hughes², Dawn A. Sim², Catherine Egan², Adnan Tufail², Hugh Montgomery², Demis Hassabis¹, Geraint Rees³, Trevor Back¹, Peng T. Khaw², Mustafa Suleyman¹, Julien Cornebise^{1,3,4}, Pearse A. Keane^{2,4*} and Olaf Ronneberger^{1,4*}

The volume and complexity of diagnostic imaging is increasing at a pace faster than the availability of human expertise to interpret it. Artificial intelligence has shown great promise in classifying two-dimensional photographs of some common diseases and typically relies on databases of millions of annotated images. Until now, the challenge of reaching the performance of expert clinicians in a real-world clinical pathway with three-dimensional diagnostic scans has remained unsolved. Here, we apply a novel deep learning architecture to a clinically heterogeneous set of three-dimensional optical coherence tomography scans from patients referred to a major eye hospital. We demonstrate performance in making a referral recommendation that reaches or exceeds that of experts on a range of sight-threatening retinal diseases after training on only 14,884 scans. Moreover, we demonstrate that the tissue segmentations produced by our architecture act as a device-independent representation; referral accuracy is maintained when using tissue segmentations from a different type of device. Our work removes previous barriers to wider clinical use without prohibitive training data requirements across multiple pathologies in a real-world setting.

ARTICLES

NATURE MEDICINE

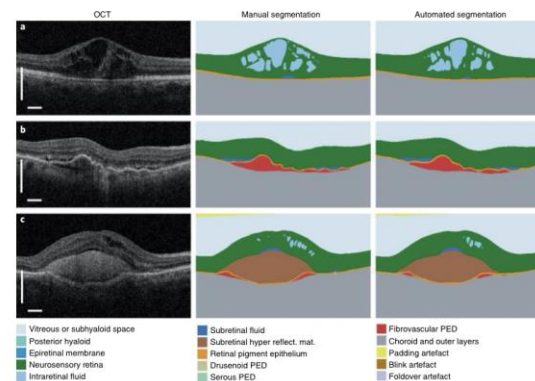


Fig. 2 | Results of the segmentation network. Three selected two-dimensional slices from the $n=224$ OCT scans in the segmentation test set (left) with manual segmentation (middle) and automated segmentation (right; detailed color legend in Supplementary Table 2). **a**, A patient with diabetic macular edema. **b**, A patient with choroidal neovascularization resulting from age-related macular degeneration (AMD), demonstrating extensive fibrovascular pigment epithelium detachment and associated subretinal fluid. **c**, A patient with neovascular AMD with extensive subretinal hyperreflective material. Further examples of the variation of pathology with model segmentation and diagnostic performance can be found in Supplementary Videos 1–9. In all examples the classification network predicted the correct diagnosis. Scale bars, 0.5 mm.

- Small range of studies around user interaction with AI/ML
- Users' perception and understanding of these systems is not always straightforward
- Making the systems operation visible *can* help (but not anything helps!)
- Important to take into account user interfaces and user interactions around AI/ML systems for responsible ML in healthcare

e.costanza@ucl.ac.uk

<https://ucl.ac.uk/people/enrico-costanza>