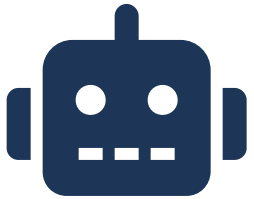
Harnessing explanations to bridge AI & humans

Vivian Lai, Samuel Carton, Chenhao Tan
@vivwylai | @SamHCarton | @ChenhaoTan
University of Colorado Boulder
CHI2020 Fair & Responsible AI Workshop

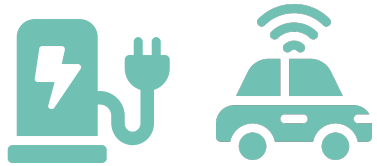
Ubiquitous Machine Learning



Medical diagnosis

Credit score prediction

Recidivism prediction



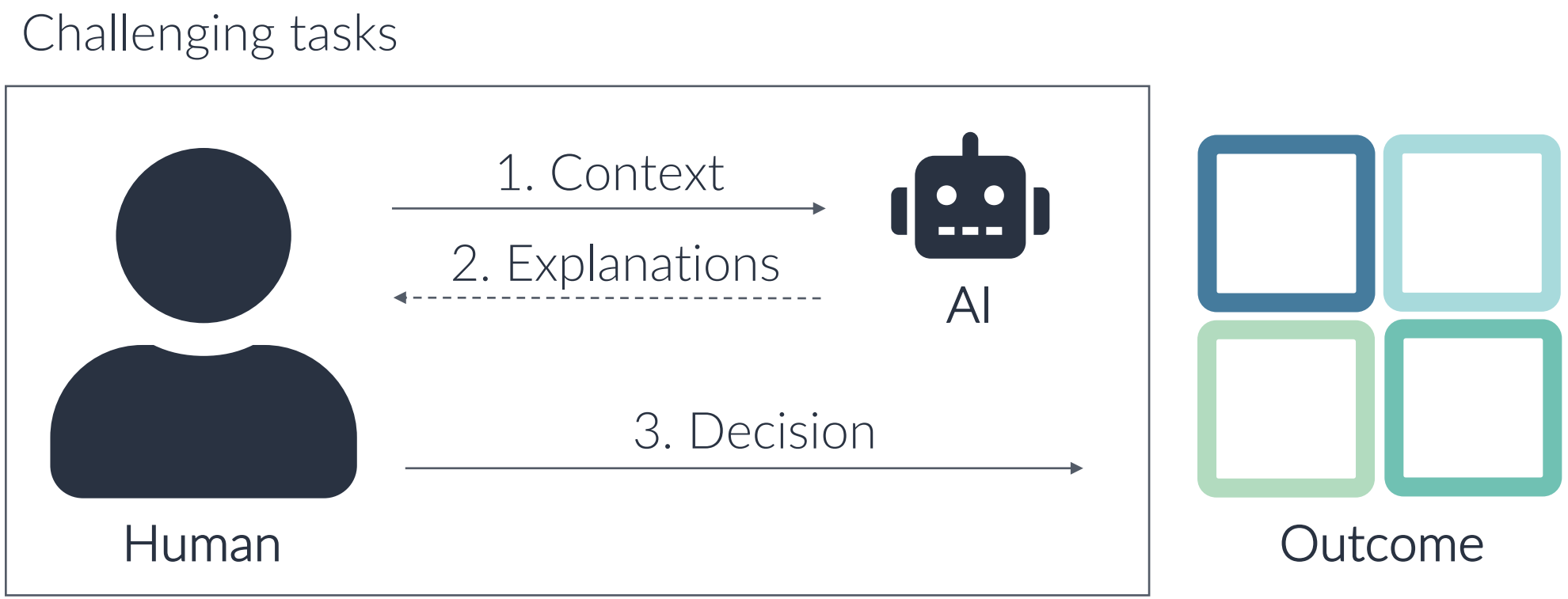
Autonomous driving

Lin et al. 2011; Geiger et al. 2012; Kleinberg et al. 2017; Ardila et al. 2019

Full automation is not desired

Supreme Court of Wisconsin 2016; Liptak 2017; Lai & Tan 2019

Machine-in-the-loop decision making



Kim et al. 2016; Lei et al. 2016; Ribeiro et al. 2016; Lundberg & Lee 2017; Tan 2018; Guidotti et al. 2019

Fail to improve performance

Poursabzi-Sangdeh et al. 2018; Green & Chen 2019; Lage et al. 2019; Lai & Tan 2019; Carton et al. 2020; Lai et al. 2020

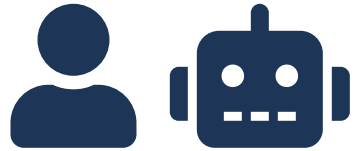
Fail to improve performance

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

You made Chicago a wonderful stay! The room was gorgeous! I came with very little on hand and my deluxe room supplied me with everything that I needed, I didn't even have to ask! Thank you so much, I will be back! Very tidy room as well!

AI can discover inconspicuous and counterintuitive patterns

Poursabzi-Sangdeh et al. 2018; Green & Chen 2019; Lage et al. 2019; Lai & Tan 2019; Carton et al. 2020; Lai et al. 2020



Misalignment between explanations and human mental model

Augment human mental model

Towards human-centered explanations

Augment human mental model

- Model-driven tutorials
- Interactive explanations
- Evaluating generalization

Cai et al. 2019; Heer 2019; Lai et al. 2020

Towards human-centered explanations

- Understanding human explanations
- Experimenting with alternative explanation types
- Explanations as model criticism

Zaidan et al. 2007; Kaushik et al. 2020

Vivian Lai, Samuel Carton, Chenhao Tan
@vivwylai | vivwylai@gmail.com
@SamHCarton | @ChenhaoTan
University of Colorado Boulder

workshop: https://tinyurl.com/harness-explanations
full paper: https://tinyurl.com/model-driven-tutorials