WHAT MACHINES SHOULDN'T DO

A Necessary Condition for Meaningful Human Control

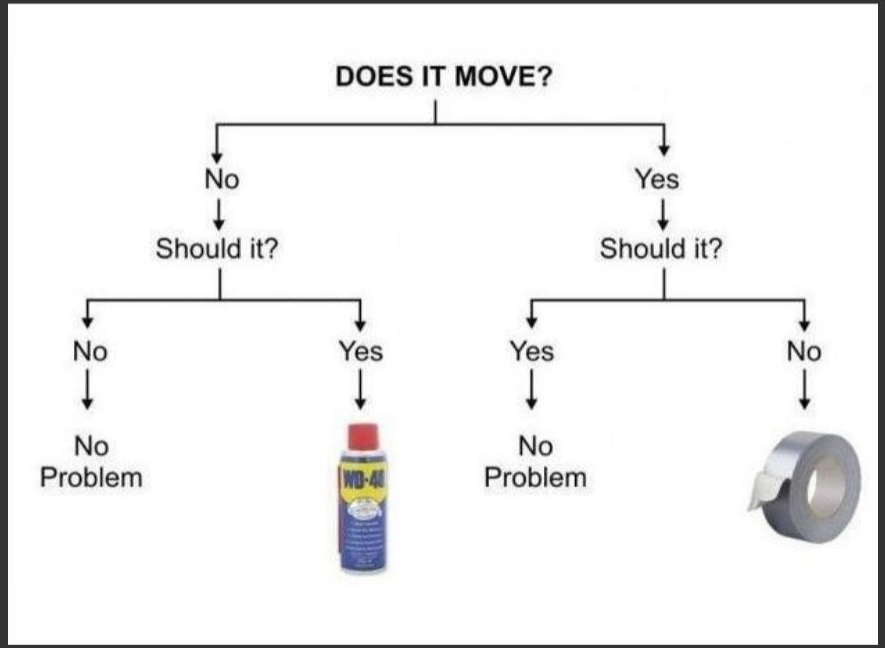
Scott Robbins | Delft University of Technology | AI Tech Agora | 9 September 2020


WHAT
MACHINES DO
DELEGATING DECISIONS TO MACHINES

WHEN HAVE WE DELEGATED A DECISION TO A MACHINE?

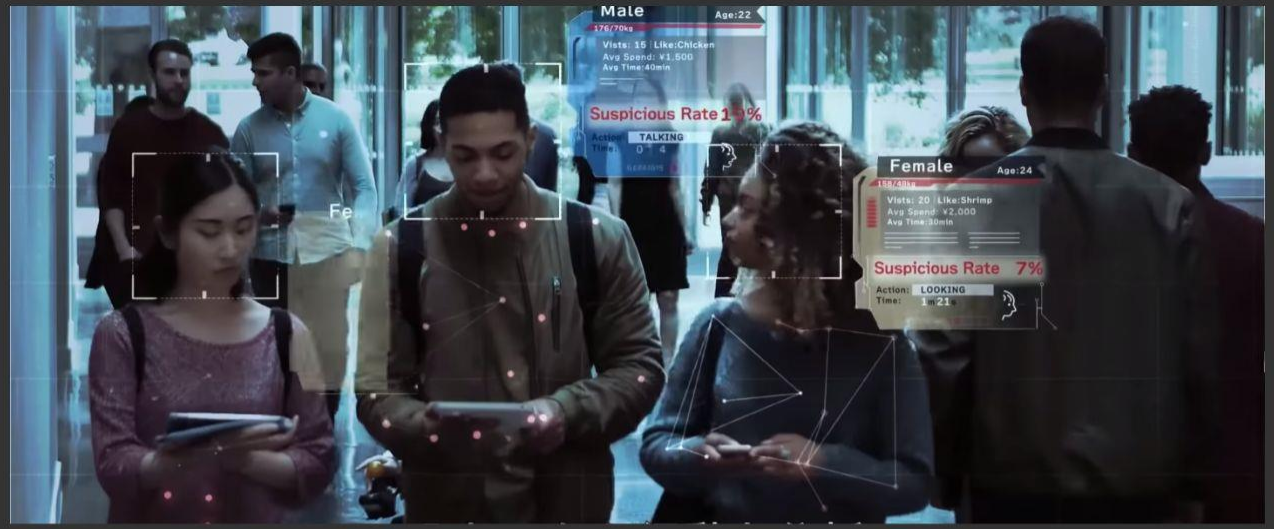
- Reach an output through considerations and weighting NOT given by humans

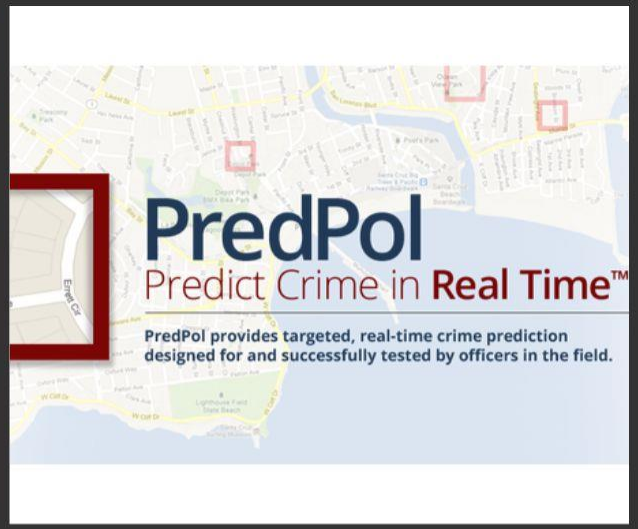
DELEGATED DECISIONS



Automated

Delegated


Male
Age:22
Suspicious Rate 19%
Female
Age:24
Suspicious Rate 7%

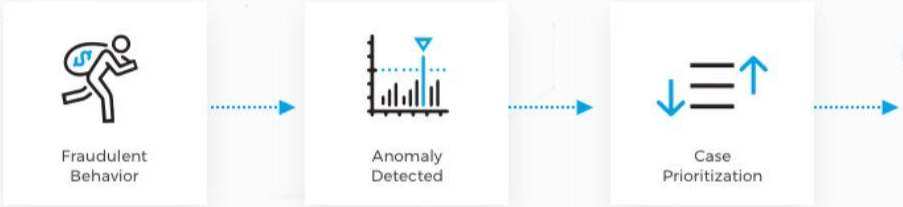

PredPol
Predict Crime in Real Time™
PredPol provides targeted, real-time crime prediction designed for and successfully tested by officers in the field.

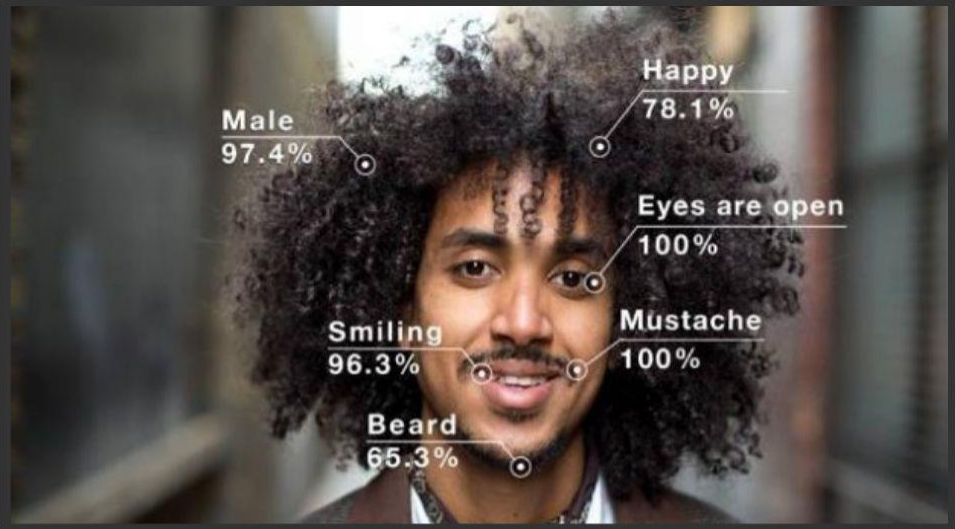

CLINICAL RECORDS
IMAGING DATA
(GEN)-OMICS INFO
SILO.AI

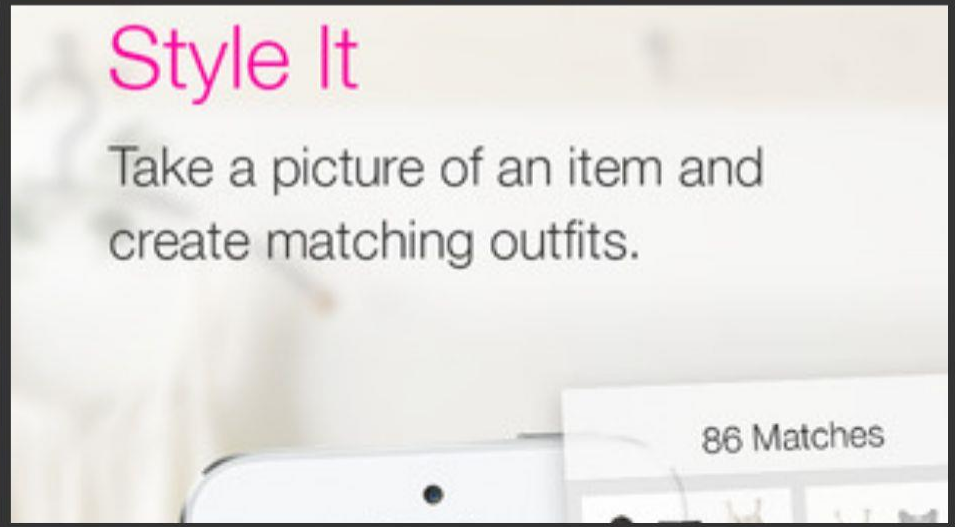

All rise.
Honorable AI
judge presiding.
entefy

AI-Powered Fraud Detection

In the US last year, $118 billion in credit card transactions were declined, but only $9 billion of this fraud. AI has been able to reduce fraud by 25% without affecting the rate of false positives.



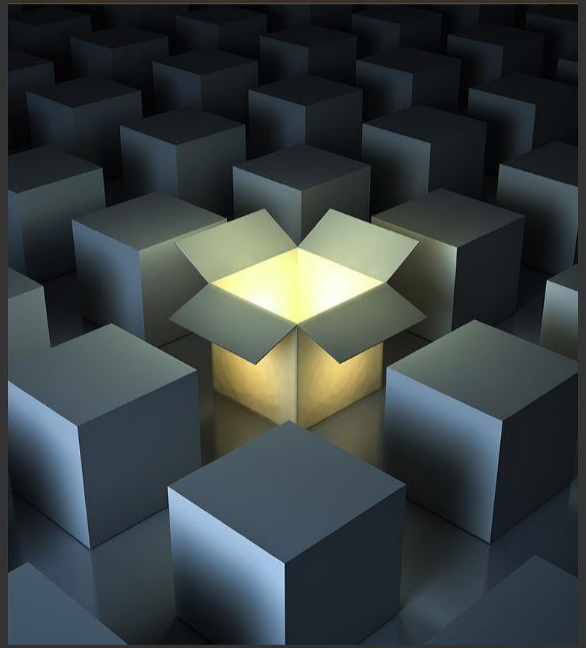




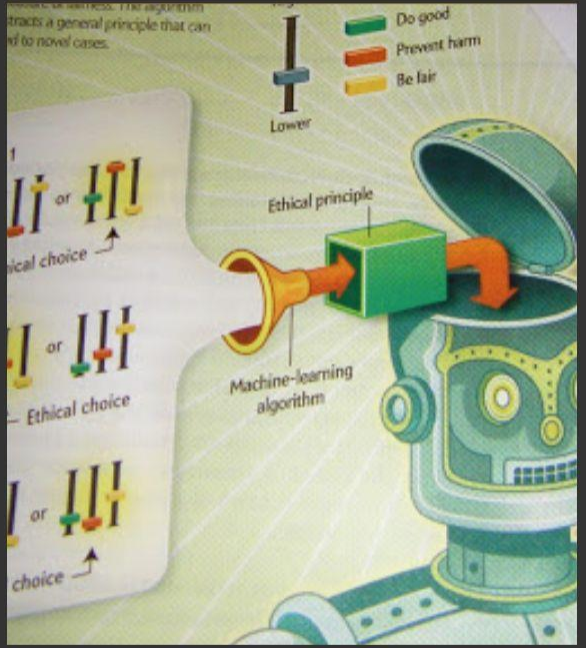
MEANINGFUL HUMAN CONTROL?

Tech & Human Centered

TECH CENTERED MHC



Explicability



Machine Ethics

Track & Trace

HUMAN CENTERED MHC

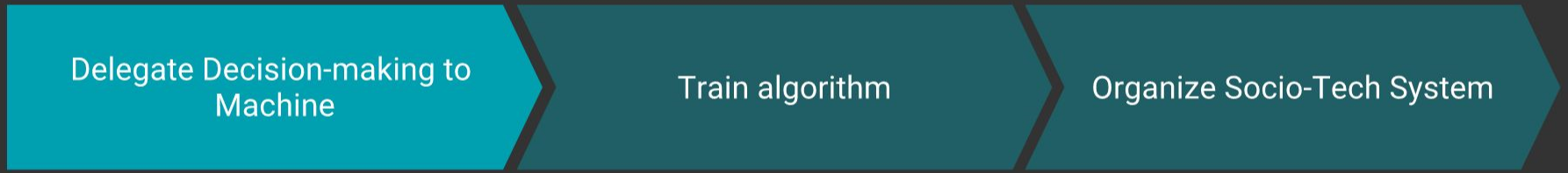
On the Loop

In the Loop

REARRANGING DECKCHAIRS

We've already hit the iceberg!

MHC IS ABOUT WHICH DECISIONS SHOULD BE DELEGATED TO MACHINES



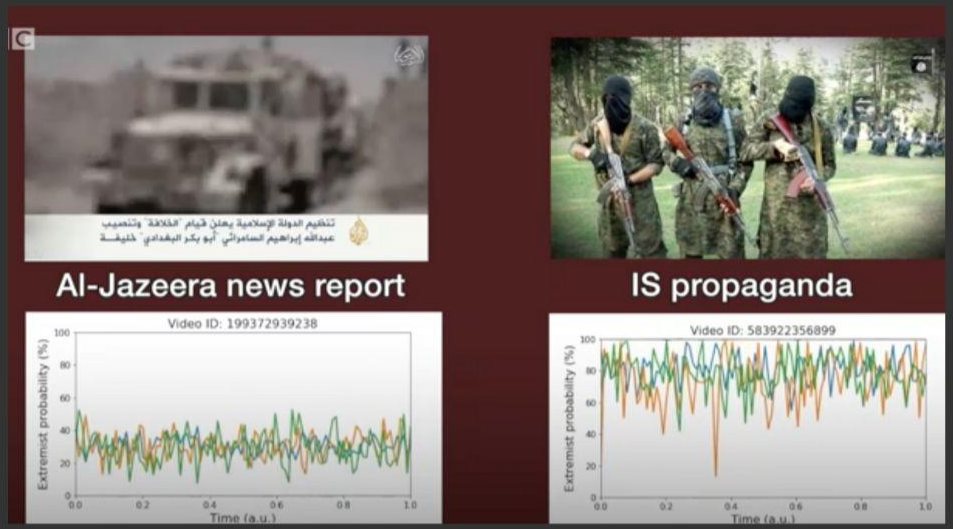

STOP
MACHINES SHOULD NOT HAVE EVALUATIVE OUTPUTS

EVALUATIVE OUTPUTS

(a) Three samples in criminal ID photo set S_c.

(b) Three samples in non-criminal ID photo set S_n

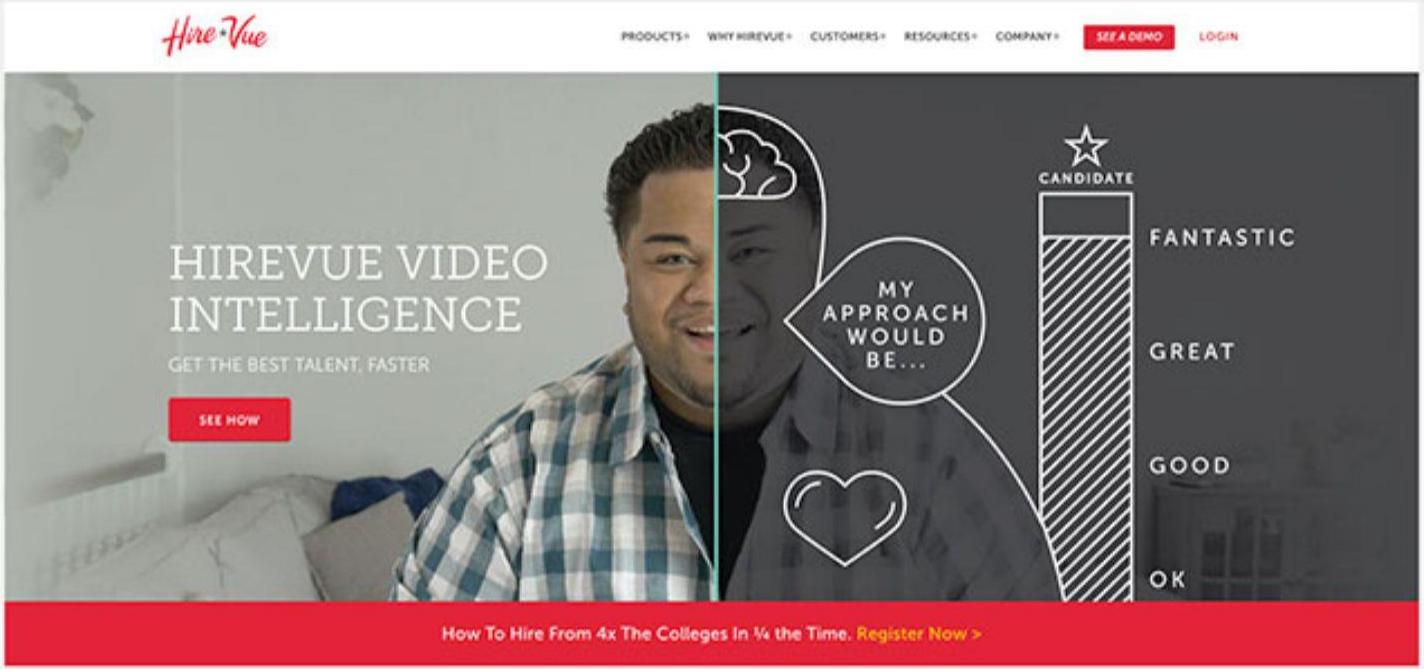
- “Criminal”
  person + convicted of crime + bad
- “Suspicious”
  person + loitering + bad
- “Beautiful”
  person + ... + good
- “Propaganda”
  picture + political message + bad
- “Fake News”
  news + false + bad
- “Bad”, “Good”, “Wrong”, “Right”


Al-Jazeera news report
IS propaganda
Video ID: 199372939238
Video ID: 583922356899
Extremist probability (%)
Time (a.u.)


(a) Three samples in criminal ID photo set S_c.


A.I. Beauty Recognition Technology


Microsoft
BUILDING AI
TO DETECT
FAKE NEWS



Reimagining Pre-hire Assessments

HireVue delivers custom assessments by blending the power of artificial intelligence with the science of I-O Psychology - all built within an easy to use video interview software platform. Watch the video to see how.


673
668
724
信用极好
753
信用极好

WHY?

Efficacy

Control

EFFICACY: UNKNOWN IN PRINCIPLE

- Every evaluative output is unverifiable
- Suspicious is a judgment
- Even if it turns out this man stole something, that does not mean he was 'suspicious'

GUN?

SUSPICIOUS?

- 1 year ago this kid would have justifiably been labeled 'suspicious'
- CONTEXTS CHANGE!
- Like, for example, a pandemic...

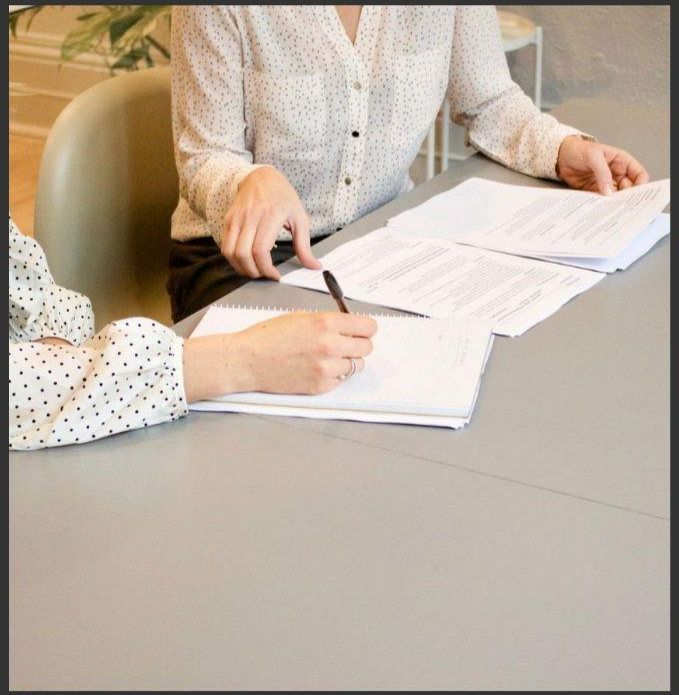
ETHICS: FUNDAMENTAL LACK OF CONTROL

"Good" or "Bad" Candidate

Evaluation

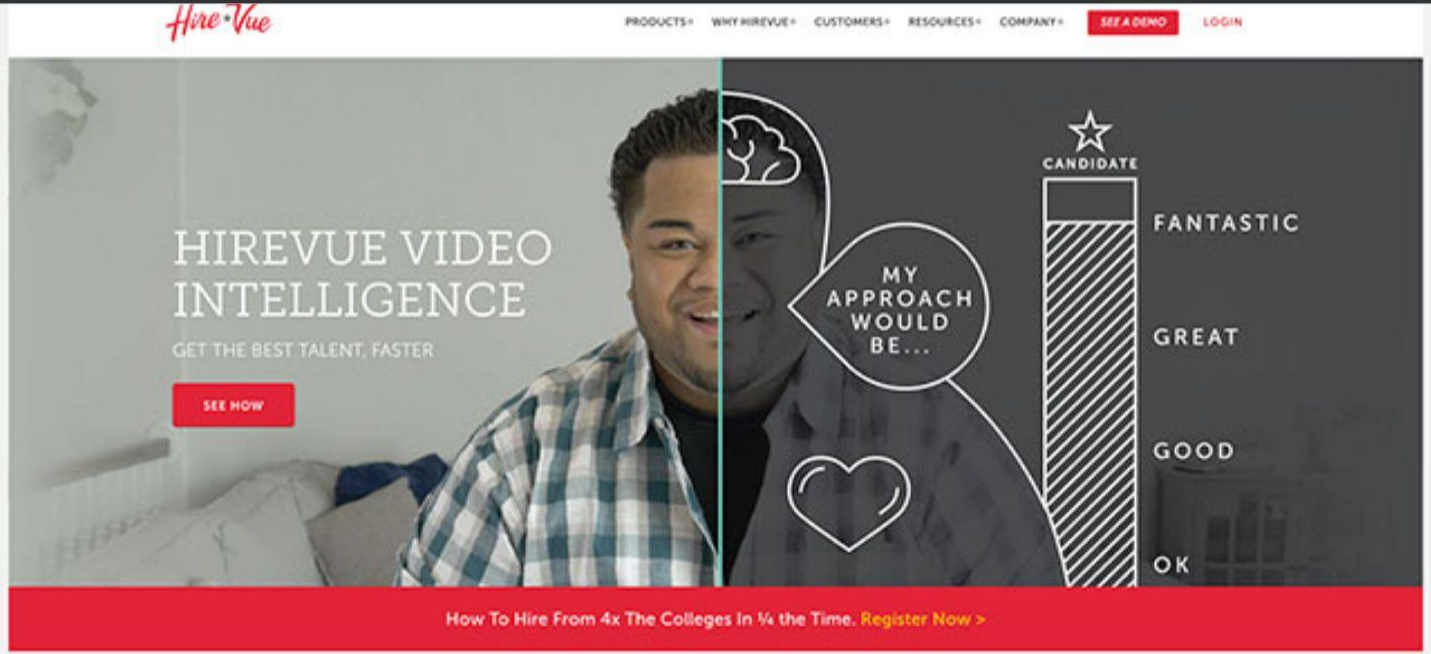
e.g. "Years of Experience", "Level of Education", "Number of Publications", etc.

Considerations

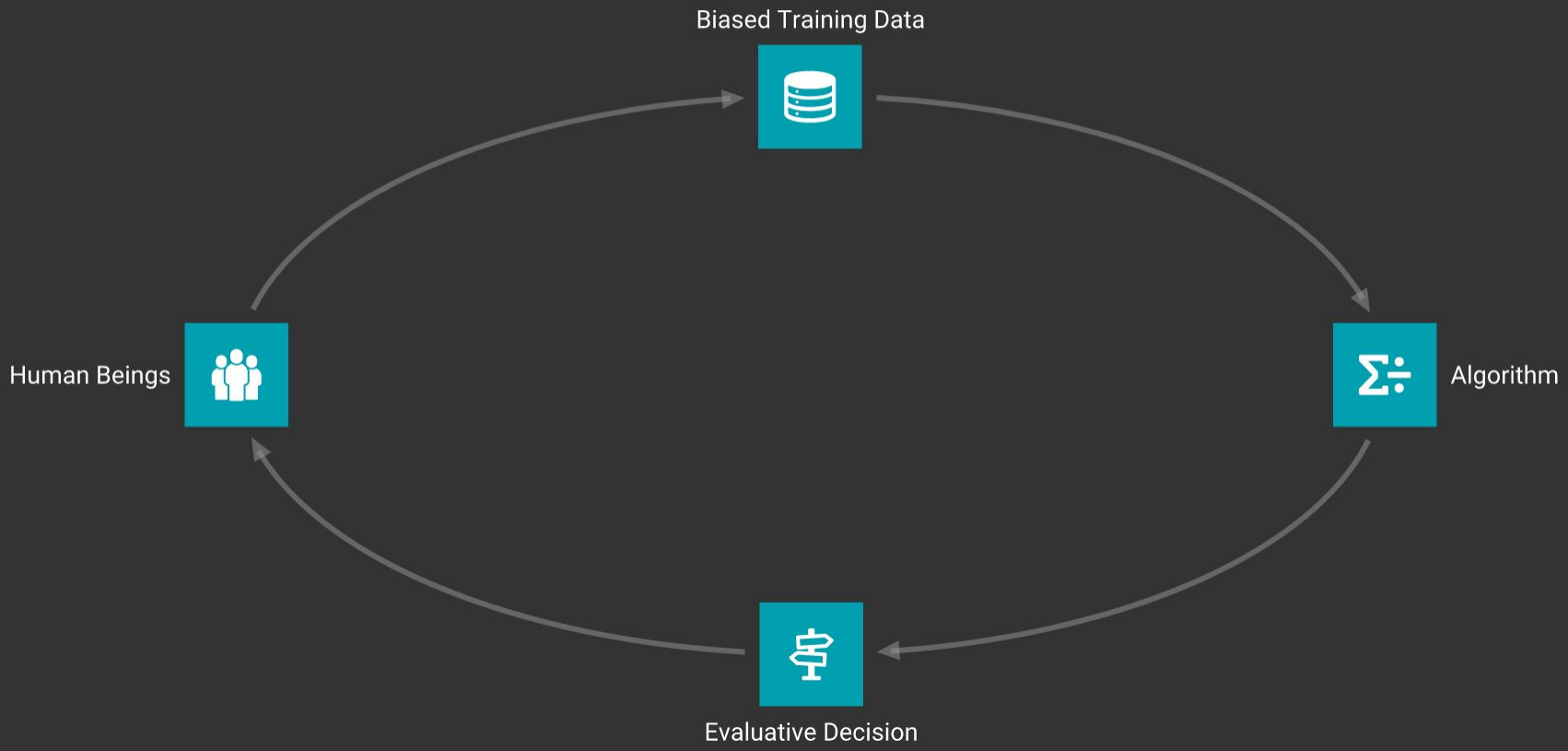


DYNAMIC: HOW WE GROUND VALUES

What makes a good candidate today may not be what makes a good candidate tomorrow



VICIOUS MACHINE FEEDBACK LOOP



DYNAMIC: PEOPLE'S BEHAVIOR

REPORT \ TECH \ ARTIFICIAL INTELLIGENCE

These students figured out their tests were graded by AI — and the easy way to cheat

"He's getting all 100s"

By Monica Chin | @mcsquared96 | Sep 2, 2020, 10:05pm EDT

EVALUATIVE OUTPUTS

Aesthetic

Ethical/Moral

Emotional

AI IS NOT A SILVER BULLET

IT CANNOT EVALUATE BETTER THAN WE CAN

KEEP AI BORING

AI should identify descriptive features that ground our evaluations

"You can't list your iPhone as your primary-care physician."

THANK YOU!

- sarobbins@protonmail.com
- https://scottrobbins.org

- #deletefacebook
- #deleteinstagram
- #deletelinkedin
- #deletegoogle
- #deletewhatsapp
- #deleteamazon
- @Boring_AI