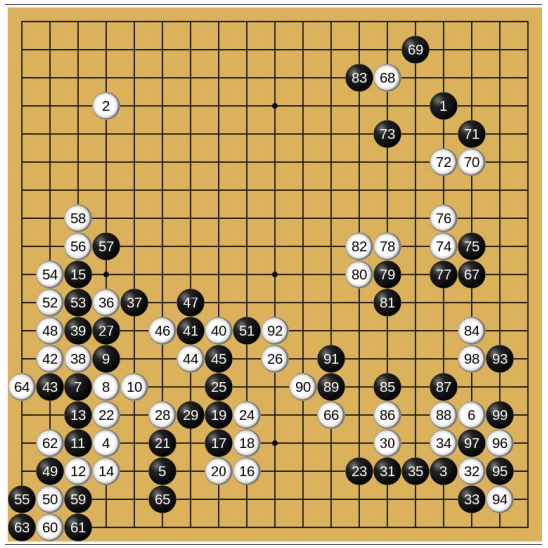
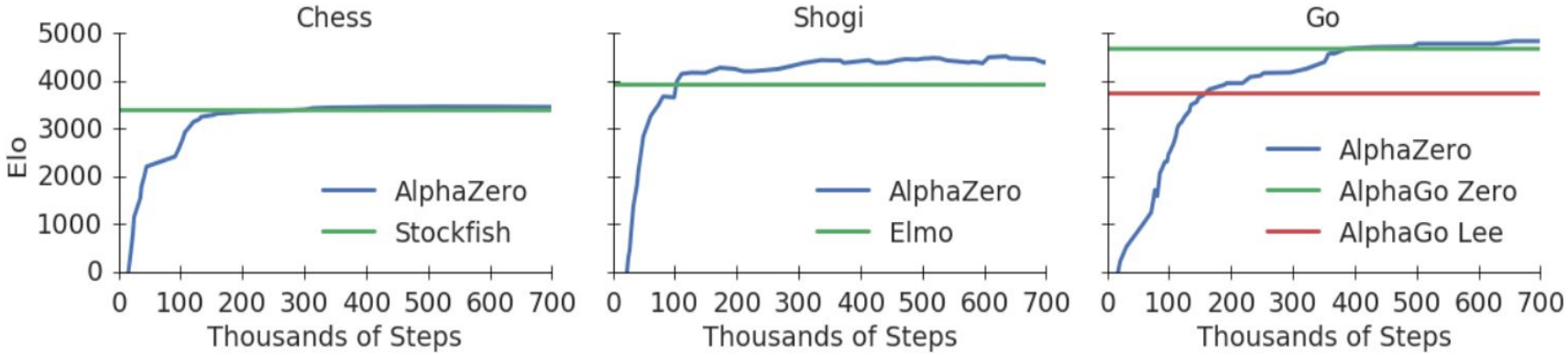
7 Habits to Build Ethical AI Systems

Karthik Bharadwaj Thirumalai
Data Council July 2019

Would YOU Trust AI?

Achievements of AI





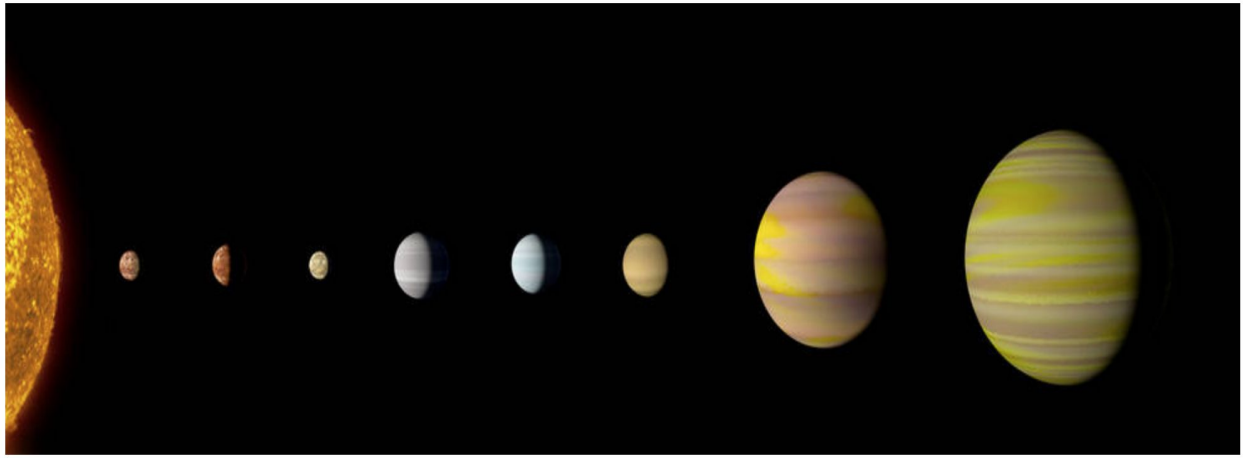
Equals Stock fish with 200K steps of training

Beats Shogi Lee with 4-1 performance in 3 days of training

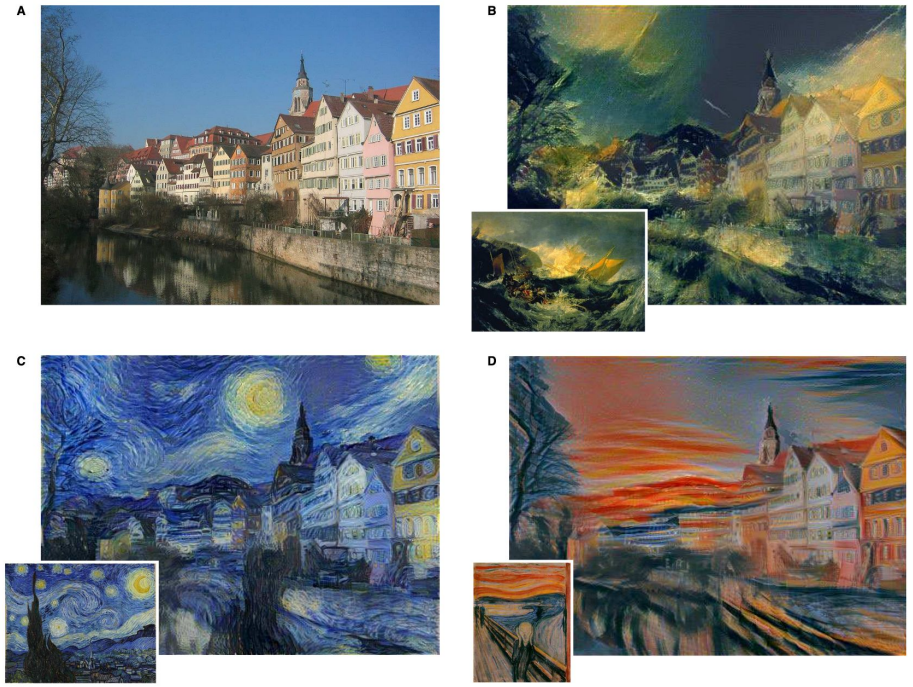
Beats 64 Professional Go players with 21 days of training.

Achievements of AI

Artificial Intelligence, NASA Data Used to Discover Eighth Planet Circling Distant Star



With the discovery of an eighth planet, the Kepler-90 system is the first to tie with our solar system in number of planets.
Credits: NASA/Wendy Stenzel



Neural Style Transfer[1]

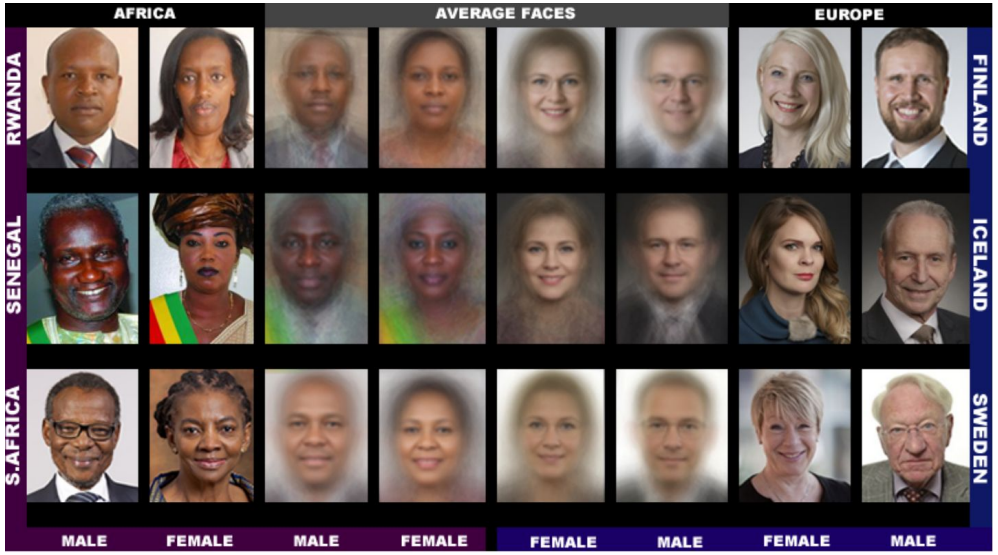
http://openaccess.thecvf.com/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf

Bias in AI

Sexism in AI[1]

Man = King ; Woman = Queen
Man = Computer Programmer; Woman = Homemaker

Gender Shades[2]



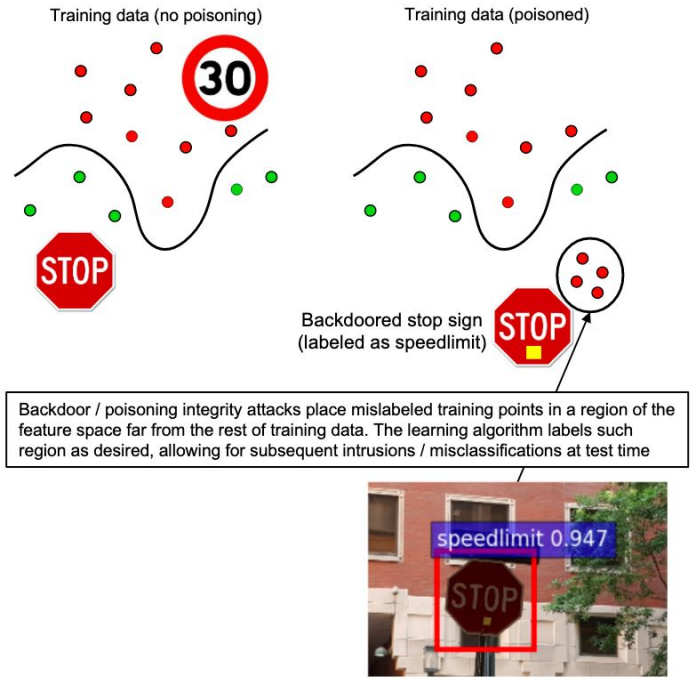
All classifiers perform better on male faces than female faces

All classifiers perform worst on darker female faces (20.8%−34.7% error rate)

[1] https://papers.nips.cc/paper/6228-man-is-to-computer-programmer-as-woman-is-to-homemaker-debiasing-word-embeddings.pdf
[2] http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
[3] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Adversarial Attacks on AI

Will you board a self-driving Car?



Impersonation Attacks - Who can take your place?

Impersonating Milla Jovovich
Impersonating Carson Daly

https://arxiv.org/pdf/1707.08945.pdf
https://arxiv.org/pdf/1712.03141.pdf
Sbhagava et al

7 Habits to Build Trustable AI

Habit #1 Fairness

Habit #2 Accountability

Habit #3 Robustness

Habit #4: Security

Habit #5: Privacy and Governance

Habit #6: Educate AI

Habit #7 : Empower Humans

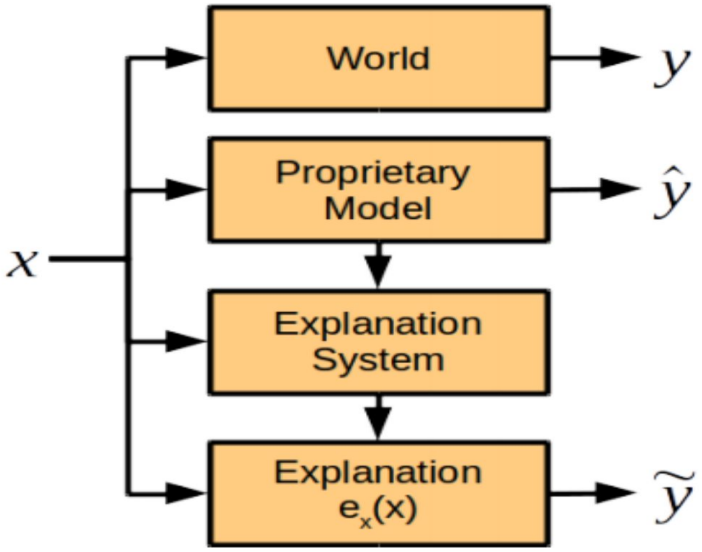
Habit #1 Fairness

1. Modify a pre-trained classifier to increase fairness
    - Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment

2. Equip for fairness during the training phase.
    - Empirical Risk Minimization Under Fairness Constraints

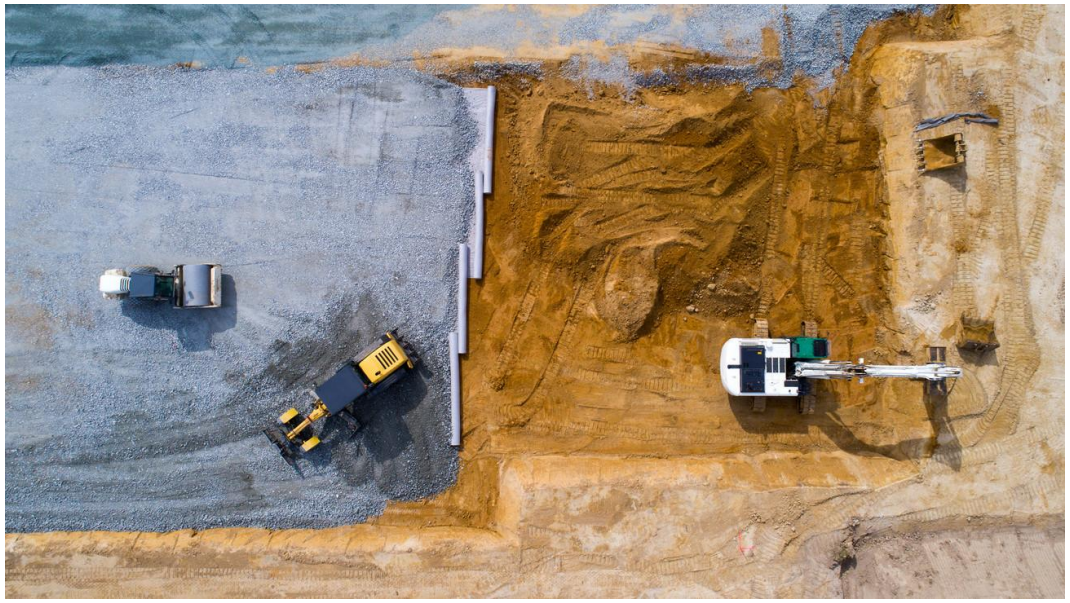
3. Modify data representation and apply algorithms.
    - Learning fair representations
    - Classification with No Discrimination by Preferential Sampling

Habit # 2 – Accountability and Governance



Explanation systems must be separate from AI systems, say the Harvard team

Framework for Explainable AI



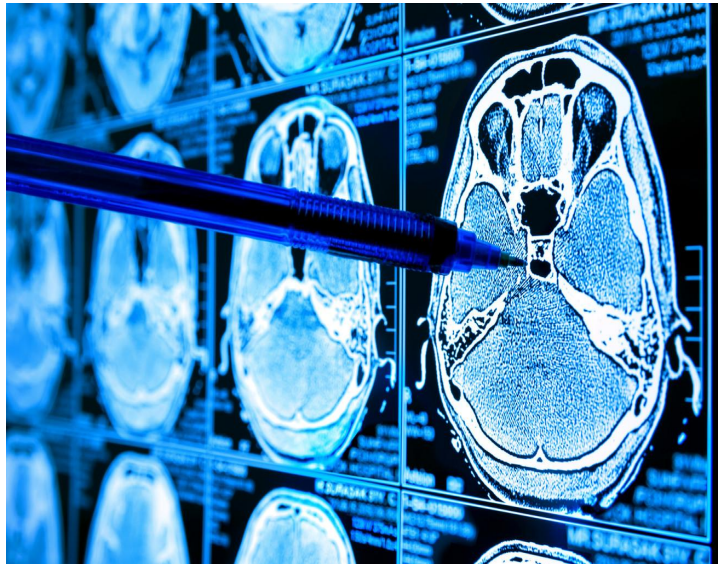
Traceability of AI systems

Model Framework, Singapore

https://arxiv.org/pdf/1711.01134.pdf

Habit #3 Robustness

Failsafe Designs



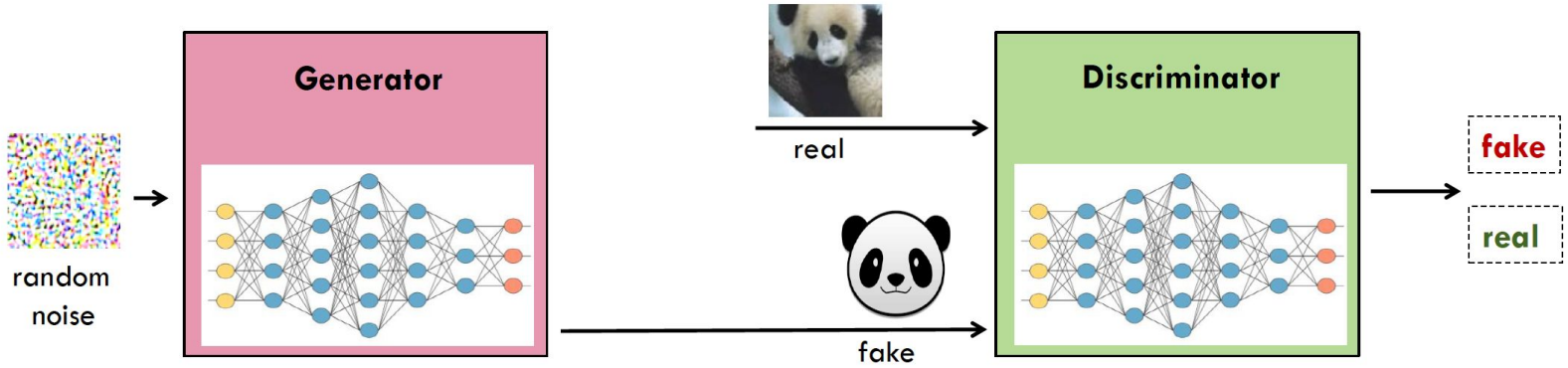
Reliable Performance Prediction

Understand the unknown

Steps towards robust AI
Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration

Habit #4 – Security

Enhance Robustness to Tampering - GENERATIVE ADVERSARIAL NETWORKS



Adversarial Training

Image blurring
Random Image resizing
Random image compression
Evaluate metrics using adversarial training.

Defensive Techniques

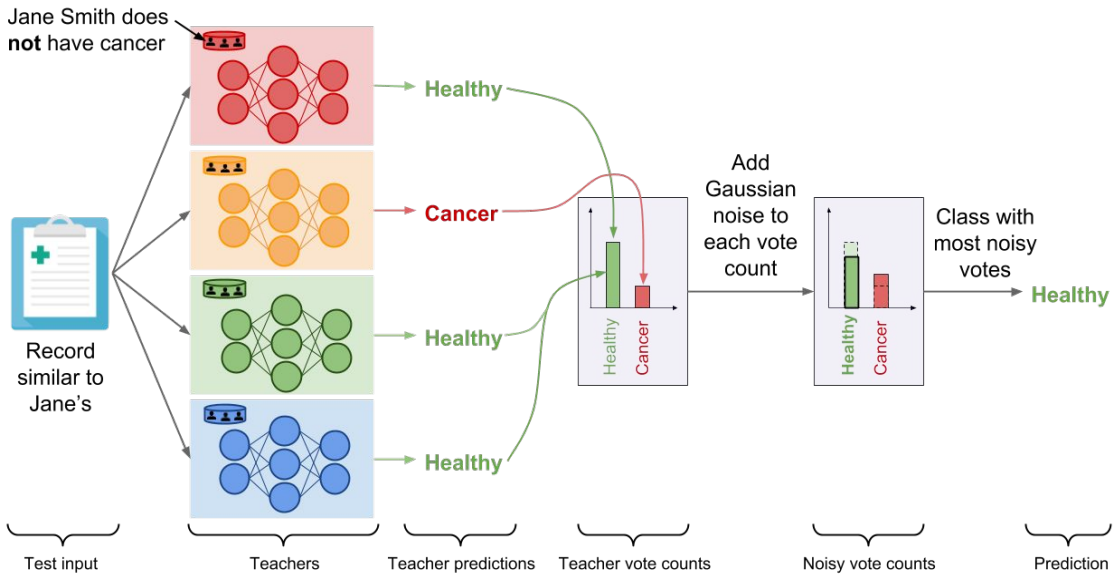
- Protect Model parameters by smoothing or hiding gradients
- Use of ensembles

Habit #5 – Privacy

Data Protection – AI Systems development ensure data protection at all stages of development.

Verified Consent – Develop systems by which people can give verified consent.

Privacy in AI: PATE Framework



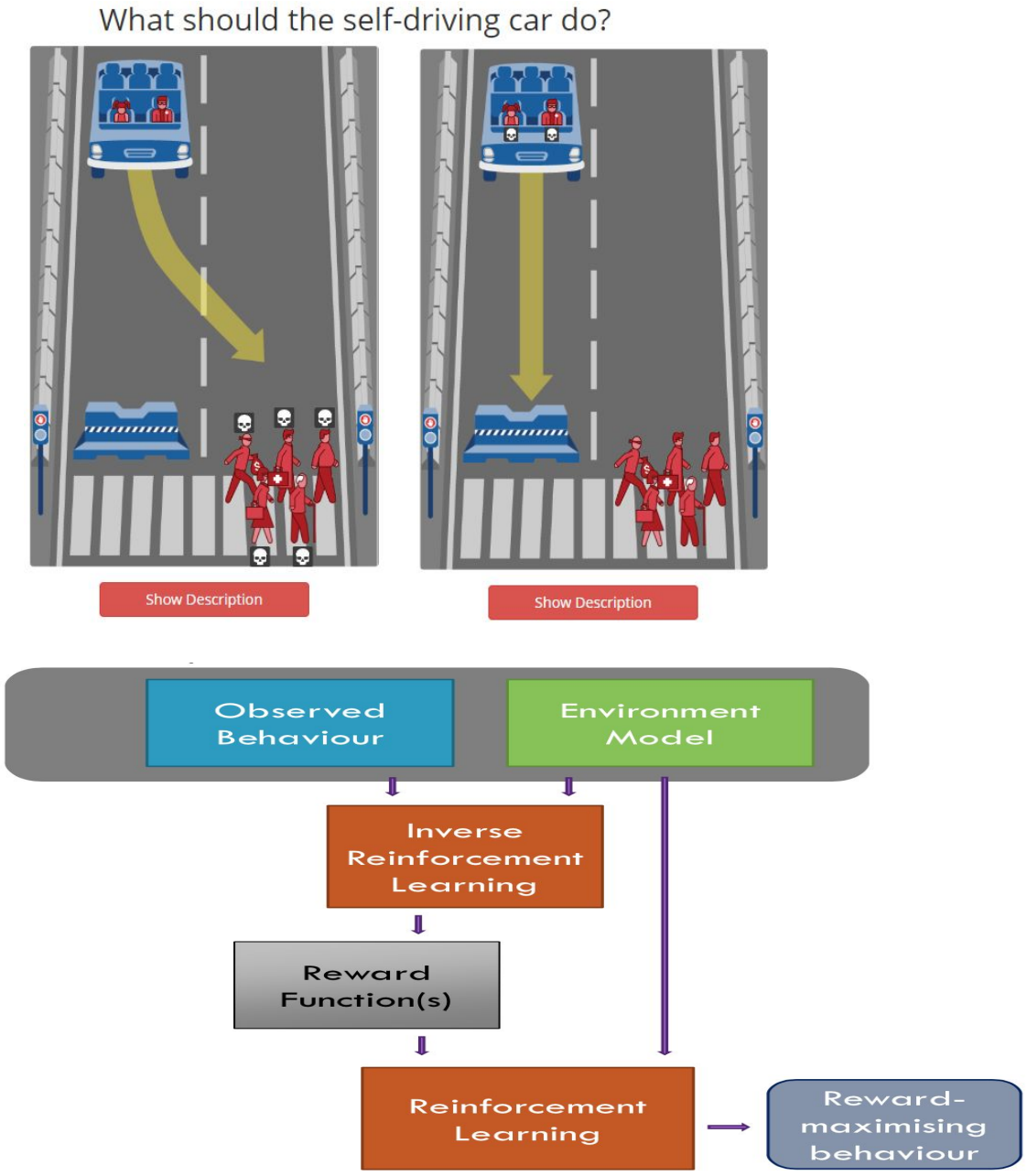
Habit #6 – Educate AI

Crowd Sourcing to teach AI to behave morally
http://moralmachine.mit.edu

Curriculum Learning
- The learning process is best when situations are not randomly presented by presented in an organized fashion
- Curriculum base learning to understand human values
and ethics

Curriculum Learning, Bengio

Inverse Reinforcement Learning
- What if RL could learn the reward function by imitating someone else.



Habit #7 – Empower Humans

GoodAI

For Social Good
Predicting Wildfires
Protecting Endangered Species
Prevent Diseases
www.goodai.com/school-for-ai

Do No Harm Prevent Harm from arising (intentional or unintentional)
Reduce compute capacities of AI

Preserve Human Agency Enable and help humans make better decisions and not take human control

Together, Make the World a Better Place