NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Towards Responsible AI

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University

Emerging Human-AI Collectives

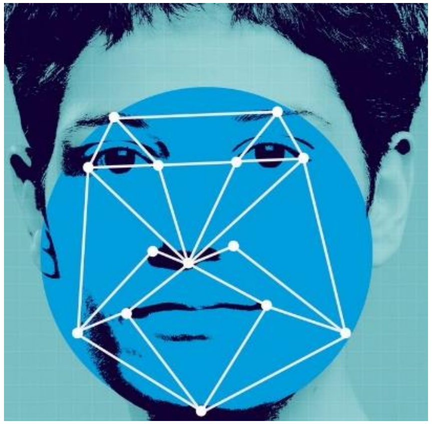
- **Flexible Autonomy**
  - Agents may take actions autonomously without reference to their owners
  - *Human-in-the-loop (a.k.a. man-on-the-loop):* humans approval required for every AI recommendation
  - *Human-over-the-loop:* humans give overall directives for AI to act autonomously
  - *Human-out-of-the-loop:* humans who are affected by AI but unable to influence AI

- **Agile Teaming**
  - Multiple agents and humans join together and disband dynamically

- **Incentive Engineering**
  - Signal the multitude of agents/humans with rewards to encourage socially desirable outcomes

Emerging Human-AI Collectives

AI is Ready. Are we?



Computer Vision

Darker skinned individuals most misclassified

Minority likely under-represented in dataset

Online Records and Ads

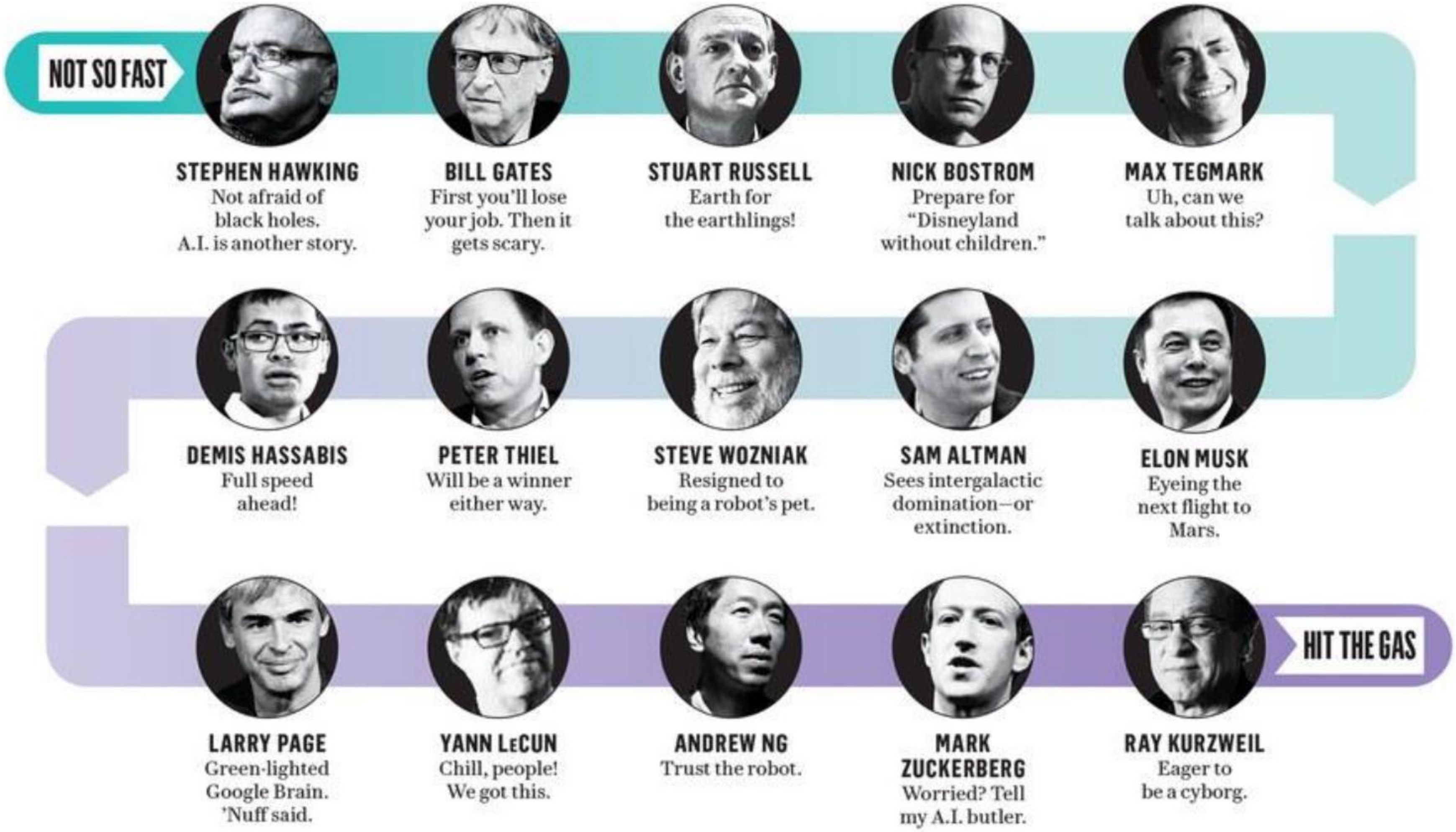
Male job seekers were more likely to be shown high paying jobs

Black-sounding names was 25% more likely to get hits suggestive of criminal record

Policing and Criminal Justice

Recidivism algorithm deems black defendants are more likely to reoffend

AI Ethics drawing Public Attention



NOT SO FAST

STEPHEN HAWKING
Not afraid of black holes. A.I. is another story.

BILL GATES
First you'll lose your job. Then it gets scary.

STUART RUSSELL
Earth for the earthlings!

NICK BOSTROM
Prepare for "Disneyland without children."

MAX TEGMARK
Uh, can we talk about this?

DEMIS HASSABIS
Full speed ahead!

PETER THIEL
Will be a winner either way.

STEVE WOZNIAK
Resigned to being a robot's pet.

SAM ALTMAN
Sees intergalactic domination—or extinction.

ELON MUSK
Eyeing the next flight to Mars.

LARRY PAGE
Green-lighted Google Brain. 'Nuff said.

YANN LECUN
Chill, people! We got this.

ANDREW NG
Trust the robot.

MARK ZUCKERBERG
Worried? Tell my A.I. butler.

RAY KURZWEIL
Eager to be a cyborg.

HIT THE GAS

AI Ethics drawing Public Attention

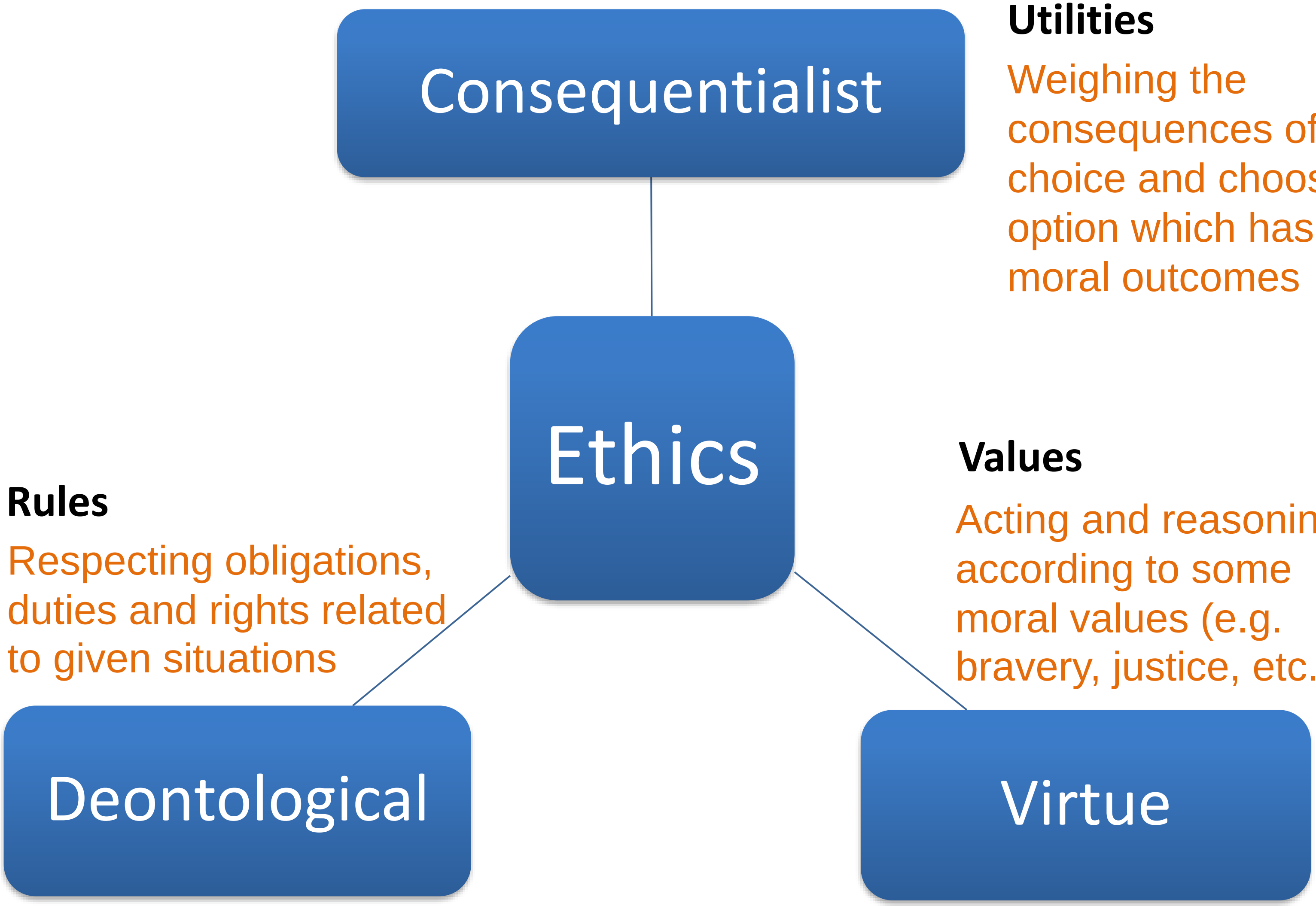
- In the real world, humans often constrain their actions according to a number of priorities:
  - Business values
  - Social norms
  - Morality
  - Religious values
- Overriding concern:
  - AI systems may not obey such values when they try to maximize their objective functions

How to design AI systems that act in line with our ethical values while achieving their design objectives?

Effort for Building Ethics in AI

H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser & Q. Yang, "Building Ethics into Artificial Intelligence," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pp. 5527–5533, 2018.

The 3 Dimensions of Ethics



Consequentialist

Utilities

Weighing the consequences of each choice and choosing the option which has the most moral outcomes

Ethics

Rules

Respecting obligations, duties and rights related to given situations

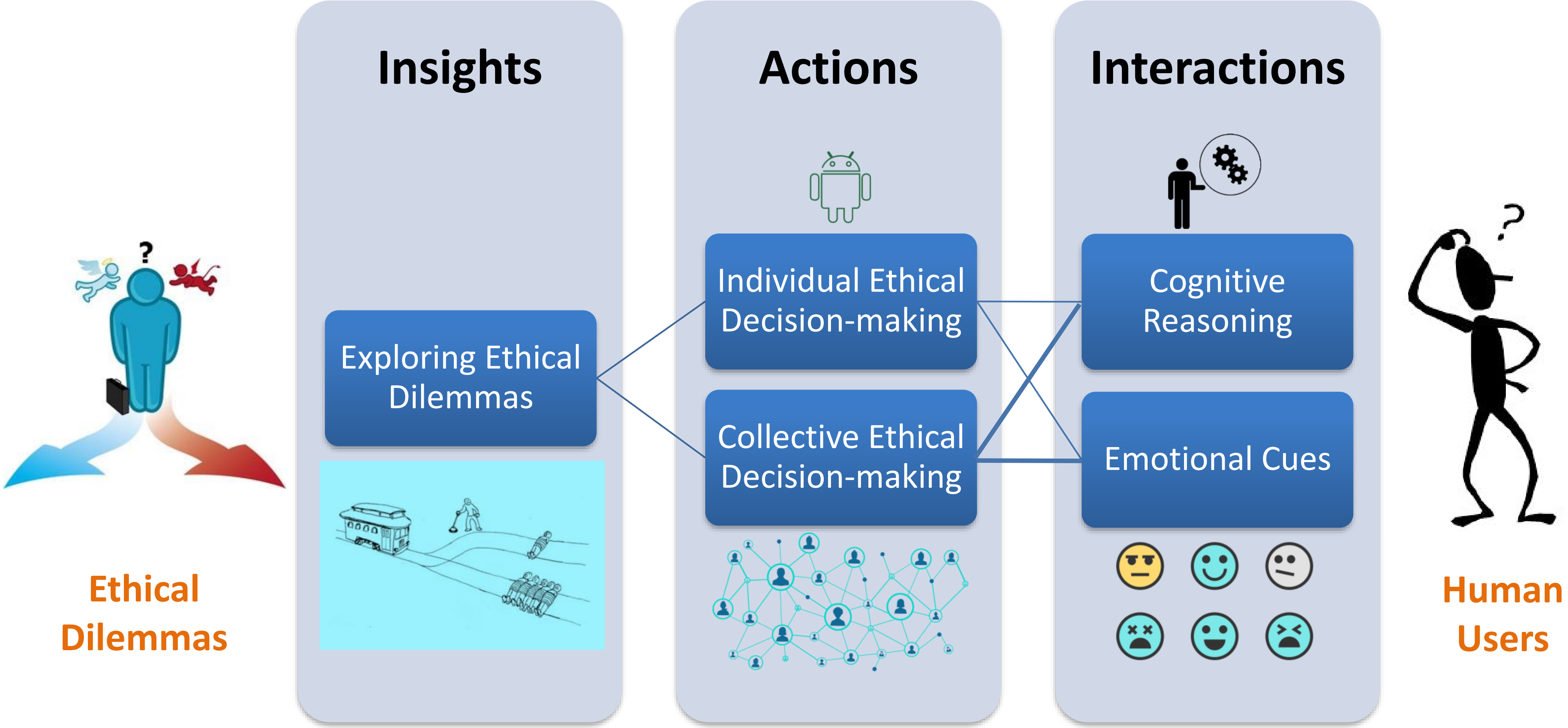
Deontological

Values

Acting and reasoning according to some moral values (e.g. bravery, justice, etc.)

Virtue

Incorporating Ethics into AI



Ethical Dilemmas

Insights

Exploring Ethical Dilemmas

Actions

Individual Ethical Decision-making

Collective Ethical Decision-making

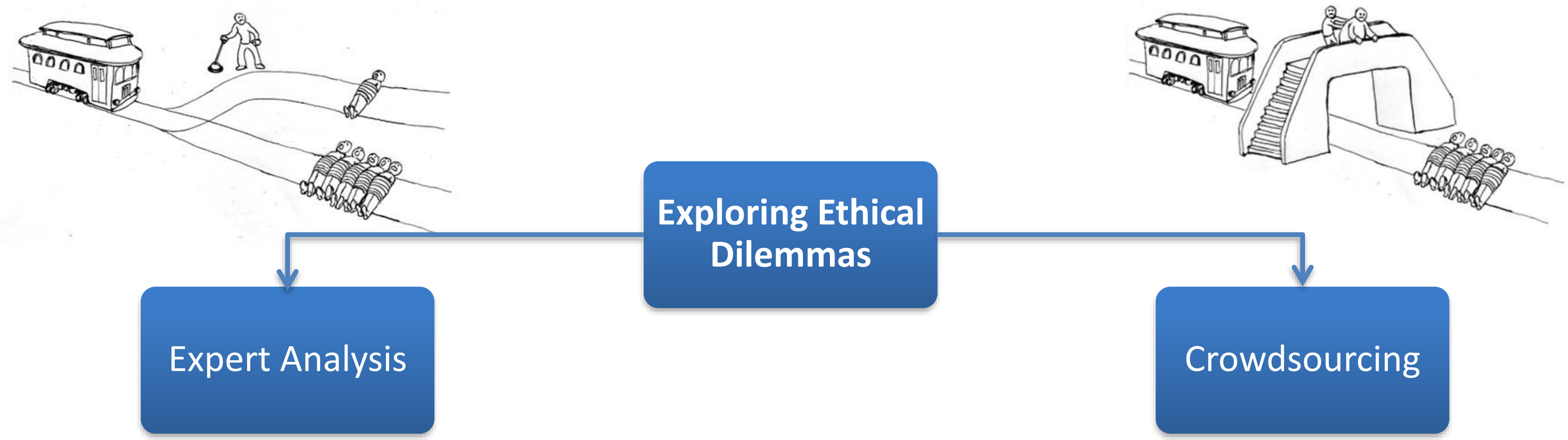
Interactions

Cognitive Reasoning

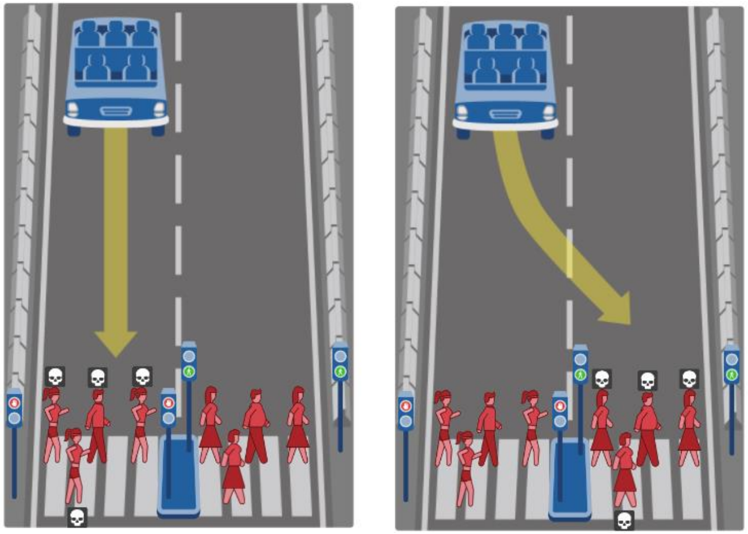
Emotional Cues

Human Users

Exploring Ethical Dilemmas

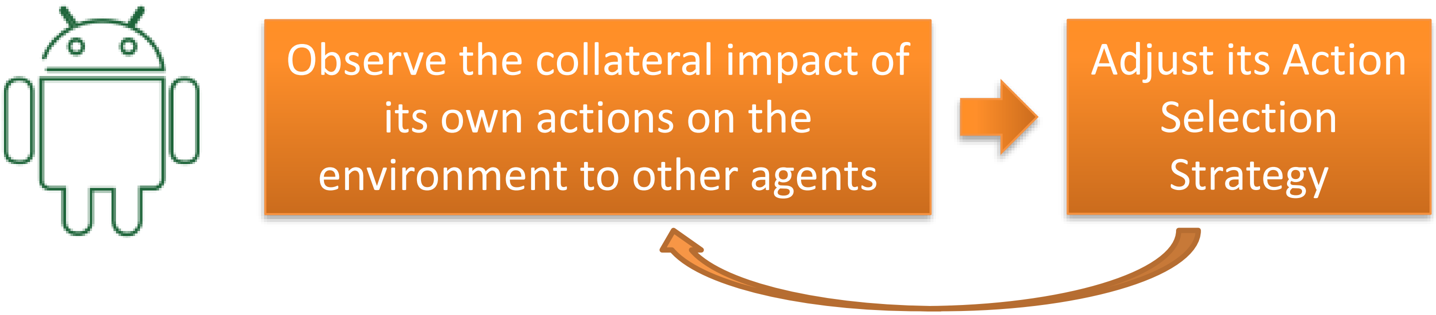


- Designing knowledge representation schemas for discussion of ethical issues (e.g., features, duties, actions, cases, and principles)
- Accounting for cultural differences, application domain specificity, and the framing effect
- Making decisions rather than declaring preferences



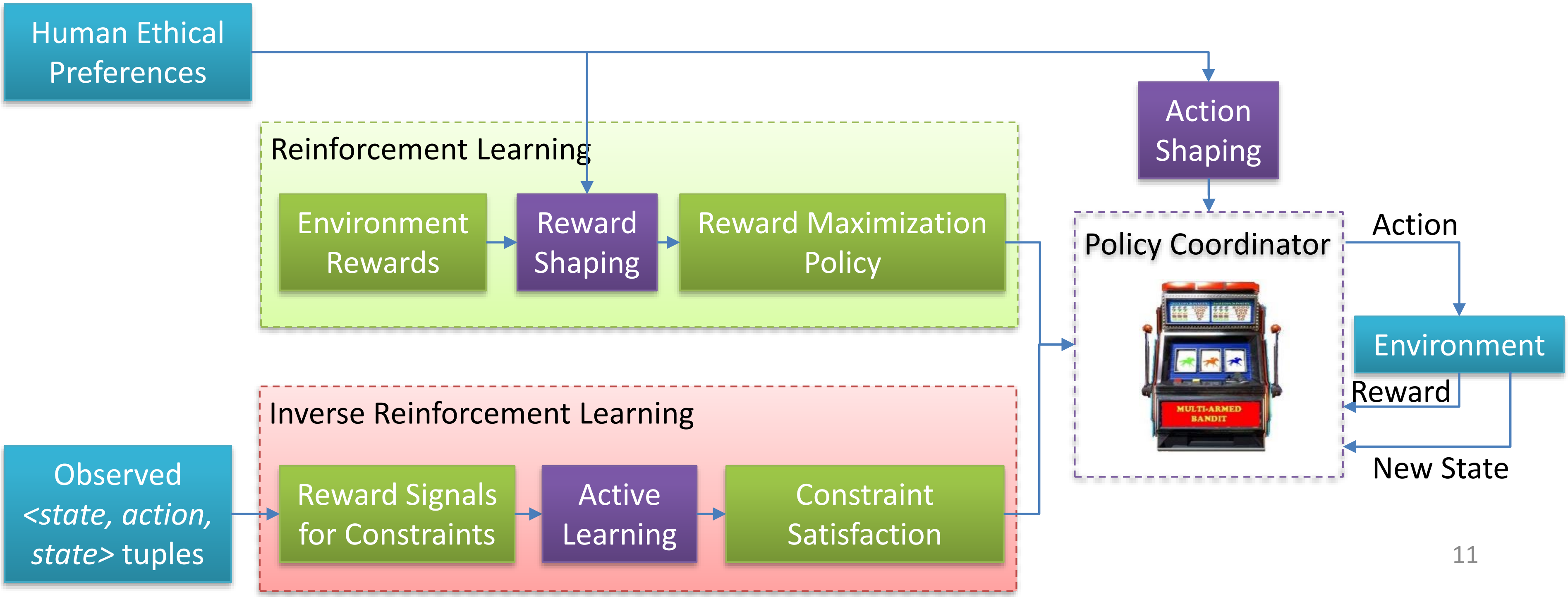
Source: http://moralmachine.mit.edu/

Individual Ethical Decision-making



Observe the collateral impact of its own actions on the environment to other agents

Adjust its Action Selection Strategy



Human Ethical Preferences

Reinforcement Learning

Environment Rewards

Reward Shaping

Reward Maximization Policy

Inverse Reinforcement Learning

Observed *<state, action, state>* tuples

Reward Signals for Constraints

Active Learning

Constraint Satisfaction

Action Shaping

Policy Coordinator

Action

Environment

Reward

New State

Collective Ethical Decision-making

Judging the Ethics of Others' Actions

Multi-agent Reinforcement Learning

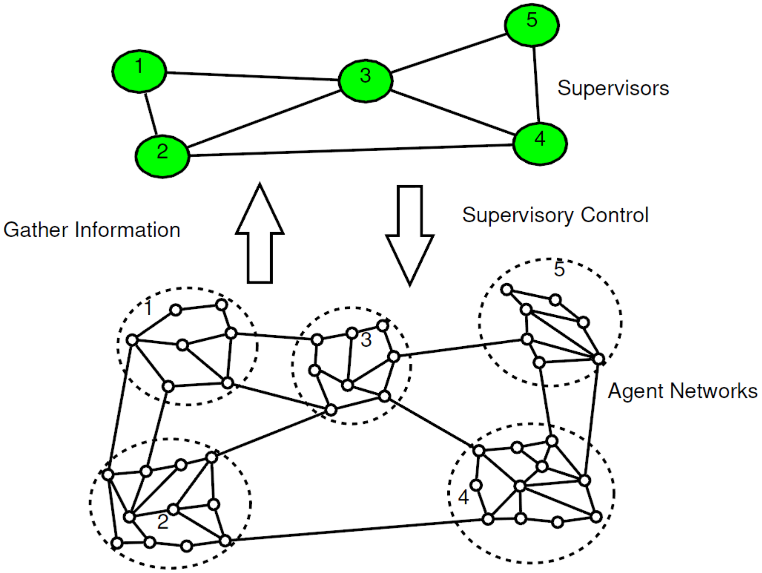
Distributed Evaluation of the 3 Dimensions of Ethics + Decision Fusion

Evolve and Enforce Social Norms

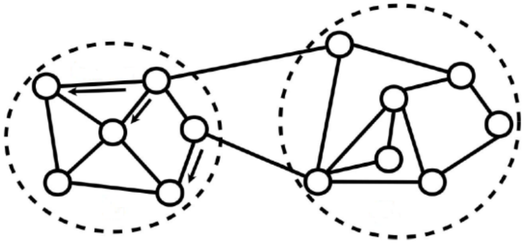
Reputation and Trust Modelling

Multi-Agent Voting

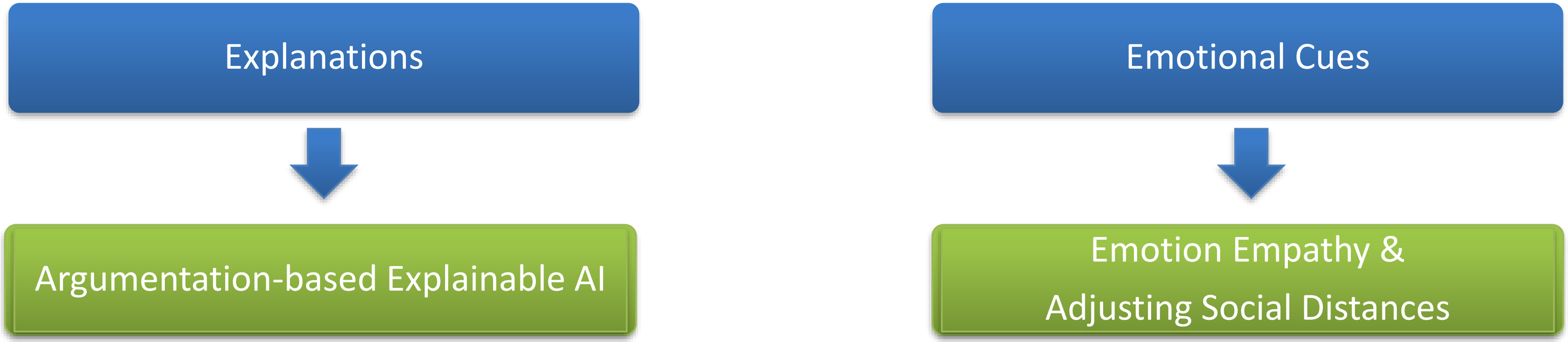
Ethics Shaping in Multi-agent RL



Ethics Shaping Propagation



Ethical Human-AI Interaction



Explanations

Argumentation-based Explainable AI

Emotional Cues

Emotion Empathy & Adjusting Social Distances

Emotions	Coping activation conditions	Coping effects
Distress	An agent's goal has failed	Lower the desire importance of failure
Fear	The current plan has a low probability of success	Drop the plan
Shame	Self-caused: the active plan puts at stake a value v_i	Lower the value priority, continue performing the current plan. Ignore the threat to the moral dimension
Reproach	Other-caused: a value is put at stake by an action performed by another agent$_i$	Create goal increaseSocialDistance(agent$_i$)
Anger	A value v_i is at stake and one of adopted goal g_j is unachievable	Create goal (increaseSocialDistance(agent$_i$) $\wedge$ reEstablish(g_j))
Remorse	A value v_i is at stake and a goal g_j is unachievable	Create goal (reEstablish(v_i) $\wedge$ reEstablish(g_j))

AI Governance Framework

Google's Responsible AI Practices:
https://ai.google/responsibilities/responsible-ai-practices/

Responsible AI Practices

Google AI

“The development of AI is creating new opportunities to improve the lives of people around the world, from business to healthcare to education. It is also raising new questions about the best way to build fairness, interpretability, privacy, and security into these systems.”

Responsible AI Practices

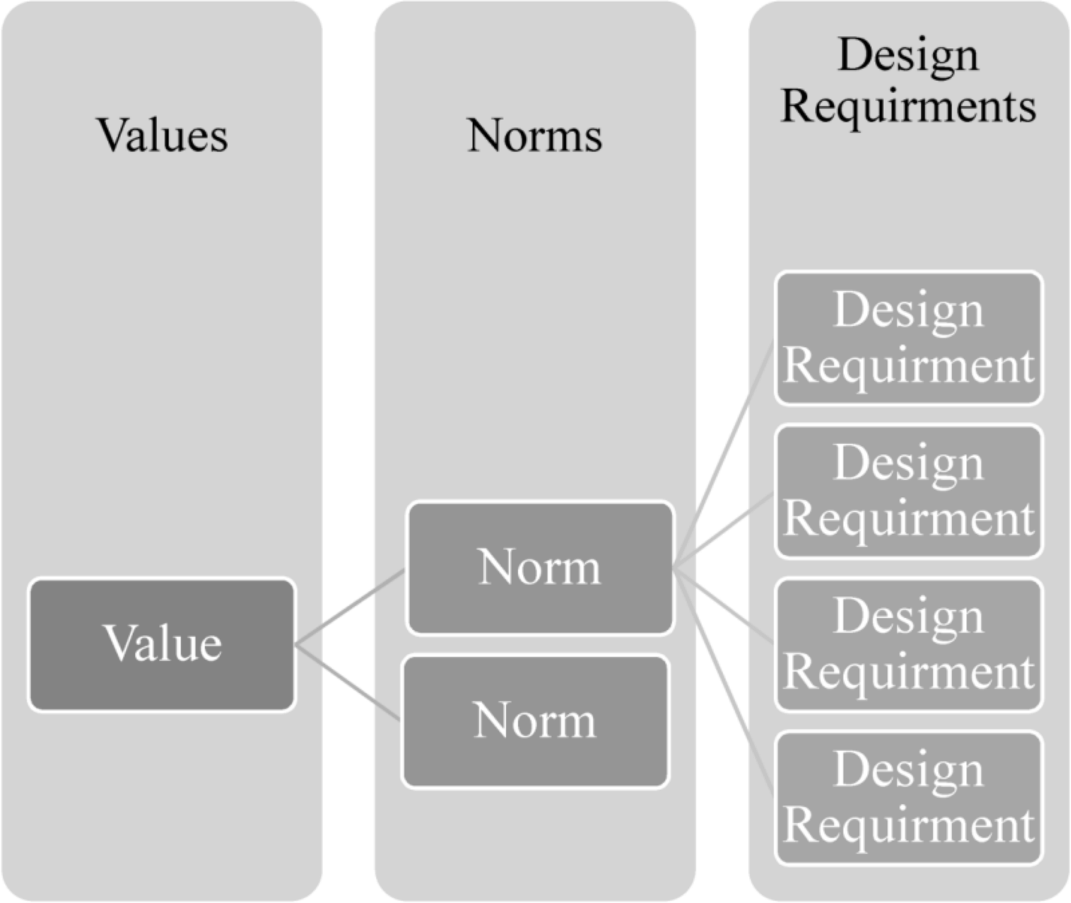
- Human-Centered Design (Week 10)
- Fairness (Week 11)
- Interpretability (Week 11)
- Privacy Preservation (Week 12)
- *Coursework Mini-Project Presentation* (Week 13)

Human-Centered Design

The way actual users experience your system is essential to assessing the true impact of its predictions, recommendations, and decisions.

- Design features with **appropriate disclosures** built-in: clarity and control is crucial to a good user experience.
- Consider **augmentation and assistance**: producing a single answer can be appropriate where there is a high probability that the answer satisfies a diversity of users and use cases. In other cases, it may be optimal for your system to suggest a few options to the user.
- **Model potential adverse feedback** early in the design process, followed by specific live testing and iteration for a small fraction of traffic before full deployment.
- **Engage with a diverse set of users and use-case scenarios**, and incorporate feedback before and throughout project development.

Human-Centered Design



Fairness

- AI systems can be used for many critical tasks, such as predicting the presence and severity of a medical condition, matching people to jobs and partners, or identifying if a person is crossing the street.
- Such computerized assistive or decision-making systems have the potential to be fairer and more inclusive at a broader scale than decision-making processes based on ad hoc rules or human judgments.
- The risk is that any unfairness in such systems can also have a wide-scale impact.
- Thus, as the impact of AI increases across sectors and societies, it is critical to work towards systems that are fair and inclusive for all.

Fairness

Design Concrete Goals of Fairness into AI Systems:

- **Engage** with social scientists, humanists, and other relevant experts for your product to understand and account for various perspectives.
- Consider how the technology and its development over time will **impact** different use cases: Whose views are represented? What types of data are represented? What's being left out? What outcomes does this technology enable and how do these compare for different users and communities? What biases, negative experiences, or discriminatory outcomes might occur?
- Set **goals** for your system to work fairly across anticipated use cases: for example, in X different languages, or to Y different age groups. Monitor these goals over time and expand as appropriate.
- Design your algorithms and objective function to **reflect** fairness goals.
- **Update** your training and testing data frequently based on who uses your technology and how they use it.

Fairness

- AI systems can be used for many critical tasks, such as predicting the presence and severity of a medical condition, matching people to jobs and partners, or identifying if a person is crossing the street.
- Such computerized assistive or decision-making systems have the potential to be fairer and more inclusive at a broader scale than decision-making processes based on ad hoc rules or human judgments.
- The risk is that any unfairness in such systems can also have a wide-scale impact.
- Thus, as the impact of AI increases across sectors and societies, it is critical to work towards systems that are fair and inclusive for all.

Fairness

- Fairness in AI decision support through joint objective optimization

Yu, H., Miao, C., Chen, Y., Fauvel, S., Li, X. & Lesser, V. R. Algorithmic management for improving collective productivity in crowdsourcing. *Scientific Reports*, vol. 7, no. 12541, Nature Publishing Group (2017).

Interpretability

- **Explicable** AI:
  - Making AI decisions obvious to a human being (i.e. a human being can understand the reason behind an AI decision without explanation)
  - Might not be the optimal solution!
- **Explainable** AI:
  - AI attempts to make the optimal decisions, the reason of which is not obvious to a human being, and requires explanation

Interpretability

Design your model to be interpretable:

- Use the **smallest** set of inputs necessary for your performance goals to make it clearer what factors are affecting the model.
- Use the **simplest** model that meets your performance goals.
- Learn **causal relationships not correlations** when possible (e.g., use height not age to predict if a kid is safe to ride a roller coaster).
- Constrain your model to produce input-output relationships that reflect **domain expert knowledge** (e.g., a coffee shop should be more likely to be recommended if it’s closer to the user, if everything else about it is the same).

Interpretability

- Explainable AI through Argumentation

- Model Interpretability with LIME

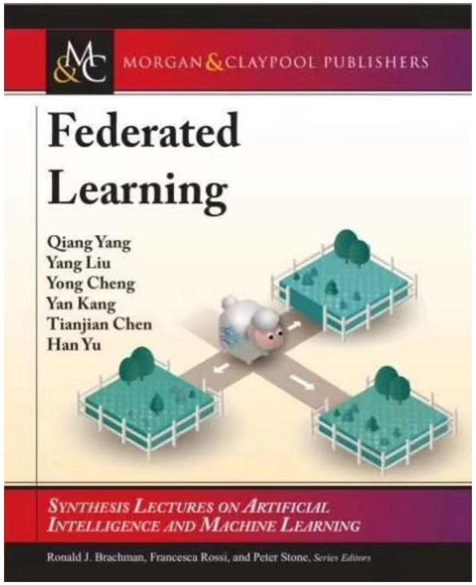
Privacy Preservation

- AI models learn from training data and make predictions on input data. Sometimes the training data, input data, or both can be quite sensitive.
- It is essential to not only respect the legal and regulatory requirements, but also consider social norms and typical individual expectations.
- Fortunately, the possibility that AI models reveal underlying data can be minimized by appropriately applying various techniques in a precise, principled fashion.
- This is an ongoing area of research in the AI community with the Federated Learning paradigm showing promise

Privacy Preservation

- Identify whether your model can be trained without the use of sensitive data, e.g., by utilizing non-sensitive data collection or an existing public data source.
- If your goal is to learn statistics of individual interactions, consider collecting only statistics that have been computed locally, on-device, rather than raw interaction data.
- Consider whether techniques like federated learning, where a fleet of devices coordinates to train a shared global model from locally-stored training data, can improve privacy in your system.

Privacy Preservation


MORGAN & CLAYPOOL PUBLISHERS
Federated Learning
Qiang Yang
Yang Liu
Yong Cheng
Yan Kang
Tianjian Chen
Han Yu
Synthesis Lectures on Artificial Intelligence and Machine Learning
Ronald J. Brachman, Francesca Rossi, and Peter Stone, Series Editors

Yang, Q., Liu, Y., Cheng, Y., Kang, Y., Chen, T. & Yu, H. (2019) *Federated Learning*. Morgan & Claypool Publishers, San Rafael, CA, USA, p. 207.

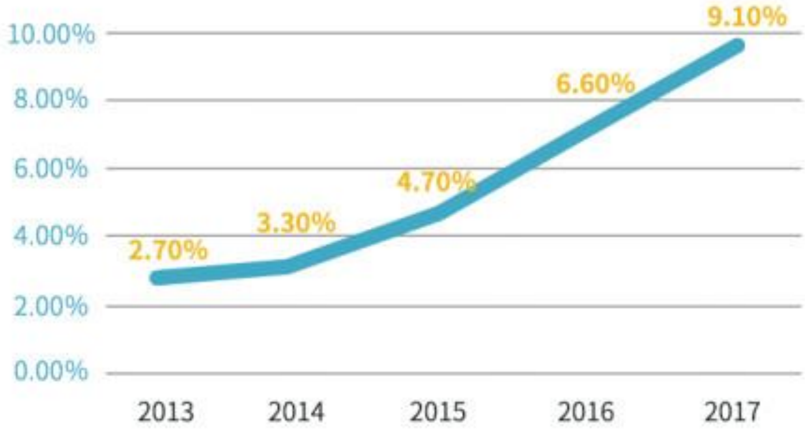
P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. G. L. D'Oliveira, S. E. Rouayheb, D. Evans, J. Gardner, Z. Garrett, A. Gascón, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, J. Konečný, A. Korolova, F. Koushanfar, S. Koyejo, T. Lepoint, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Özgür, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. T. Suresh, F. Tramèr, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. X. Yu, H. Yu & S. Zhao, "Advances and Open Problems in Federated Learning," *CoRR*, arXiv:1912.04977, 2019.

Case Study

Overview

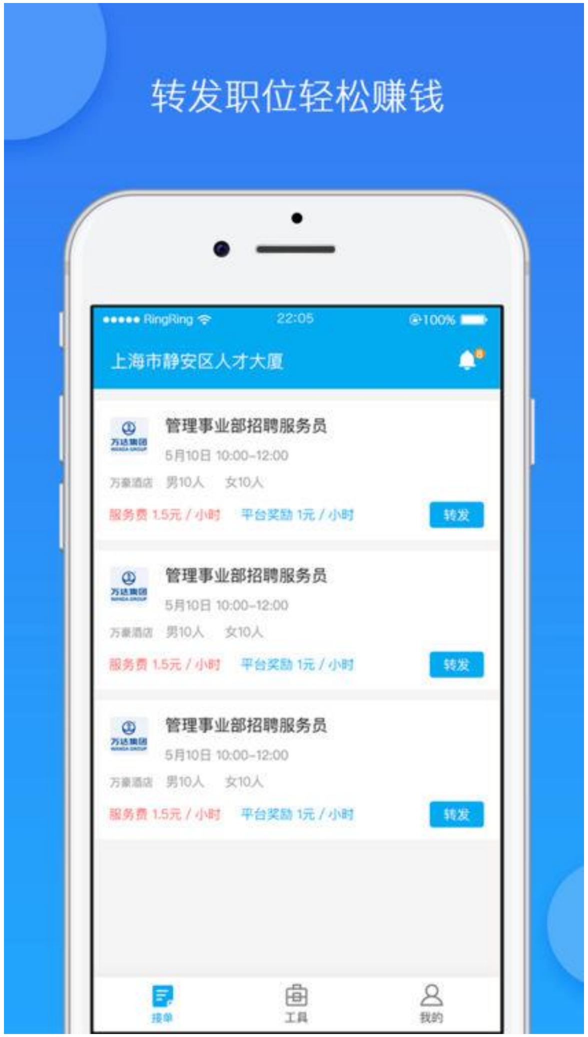
KTV
2718 6888
新景源批发
贝美
鸿发百货
佳和参茸
龙华人才市场

深圳市华辉大人力资源
人才市场管理指引
求职应聘流程


10.00%
8.00%
6.00%
4.00%
2.00%
0.00%
2.70%
3.30%
4.70%
6.60%
9.10%
2013
2014
2015
2016
2017

新信人力资源市场
深圳领域人力资源市场

Personal Work Management Apps



转发职位轻松赚钱

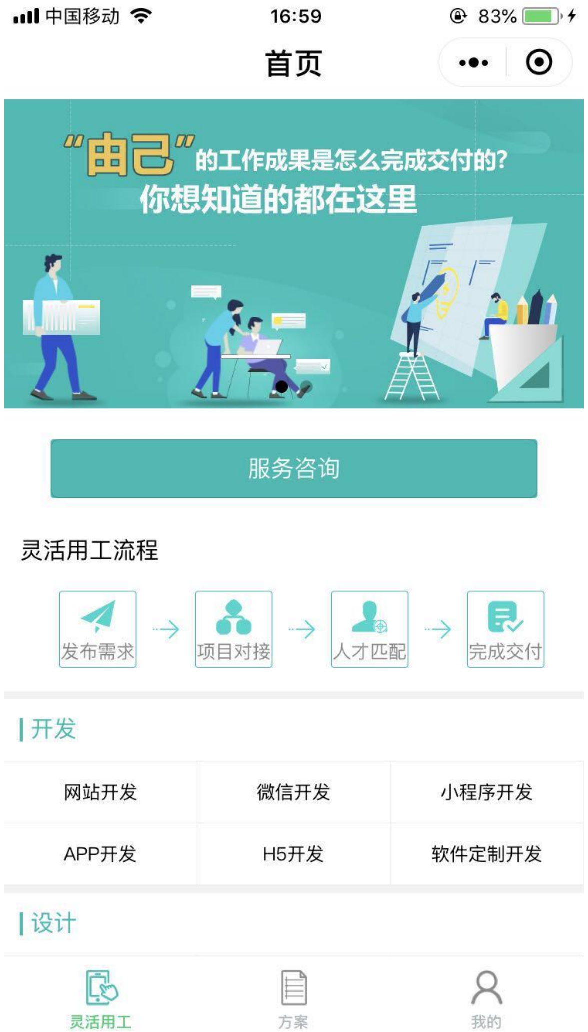
●●●●● RingRing 22:05 100%

上海市静安区人才大厦

管理事业部招聘服务员
5月10日 10:00-12:00
万豪酒店 男10人 女10人
服务费 1.5元 / 小时 平台奖励 1元 / 小时 转发

管理事业部招聘服务员
5月10日 10:00-12:00
万豪酒店 男10人 女10人
服务费 1.5元 / 小时 平台奖励 1元 / 小时 转发

管理事业部招聘服务员
5月10日 10:00-12:00
万豪酒店 男10人 女10人
服务费 1.5元 / 小时 平台奖励 1元 / 小时 转发



中国移动 16:59 83%

首页

“由己”的工作成果是怎么完成交付的?
你想知道的都在这里

服务咨询

灵活用工流程

发布需求 → 项目对接 → 人才匹配 → 完成交付

| 开发

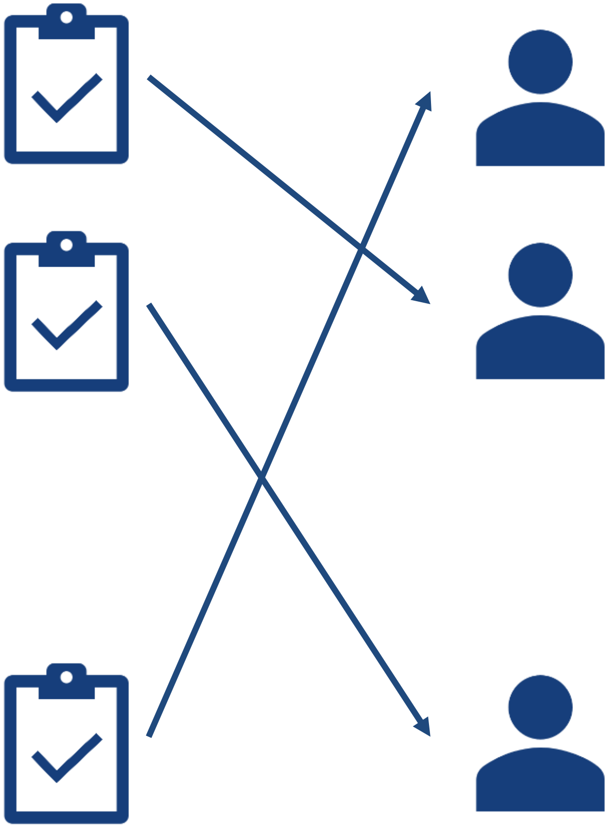
网站开发	微信开发	小程序开发
APP开发	H5开发	软件定制开发

| 设计

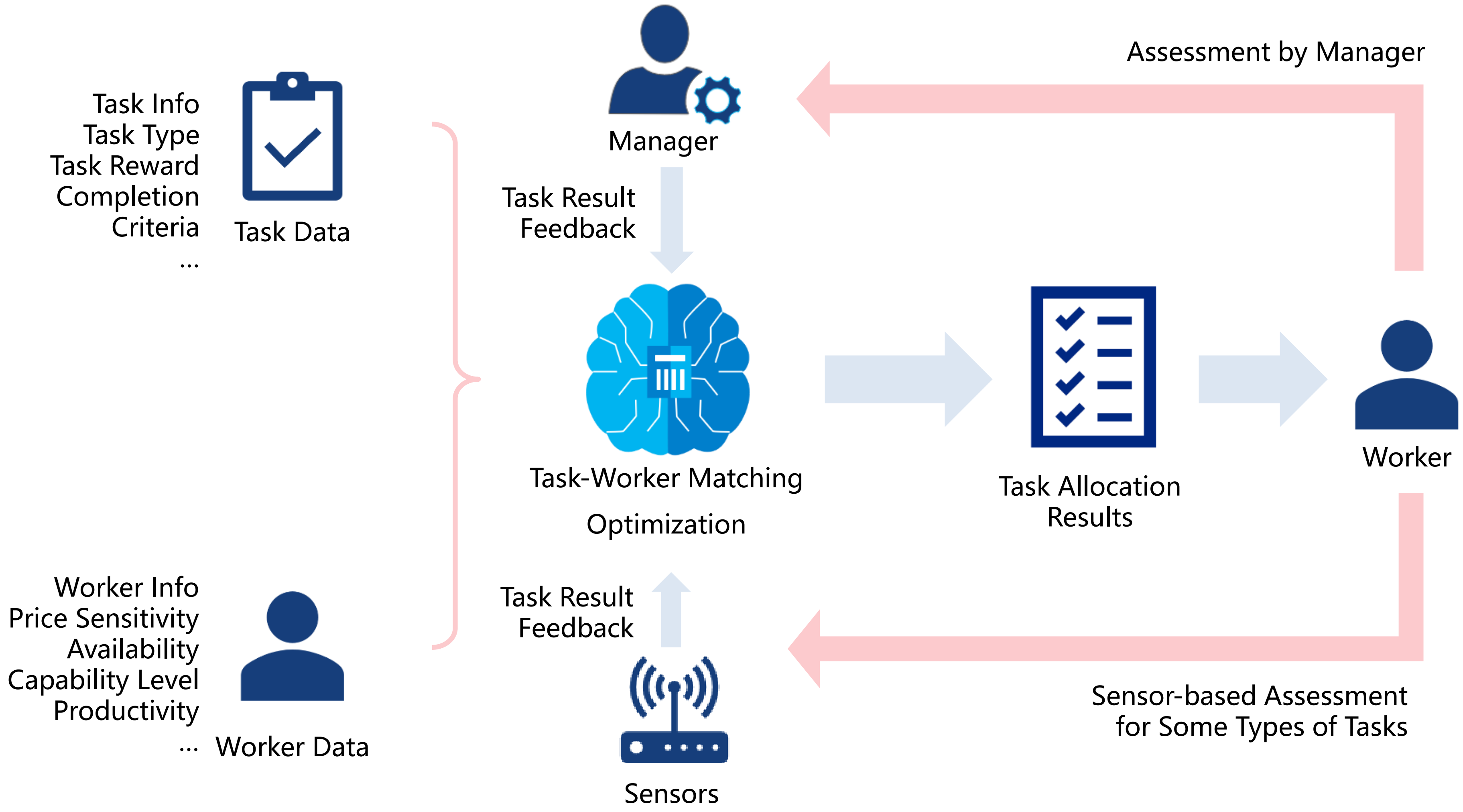
Key Challenge

*How to optimize the **dynamic** matching of tasks to the **most suitable workers** in **real time,** in order to **satisfy business objectives** while **protecting workers' wellbeing**?*

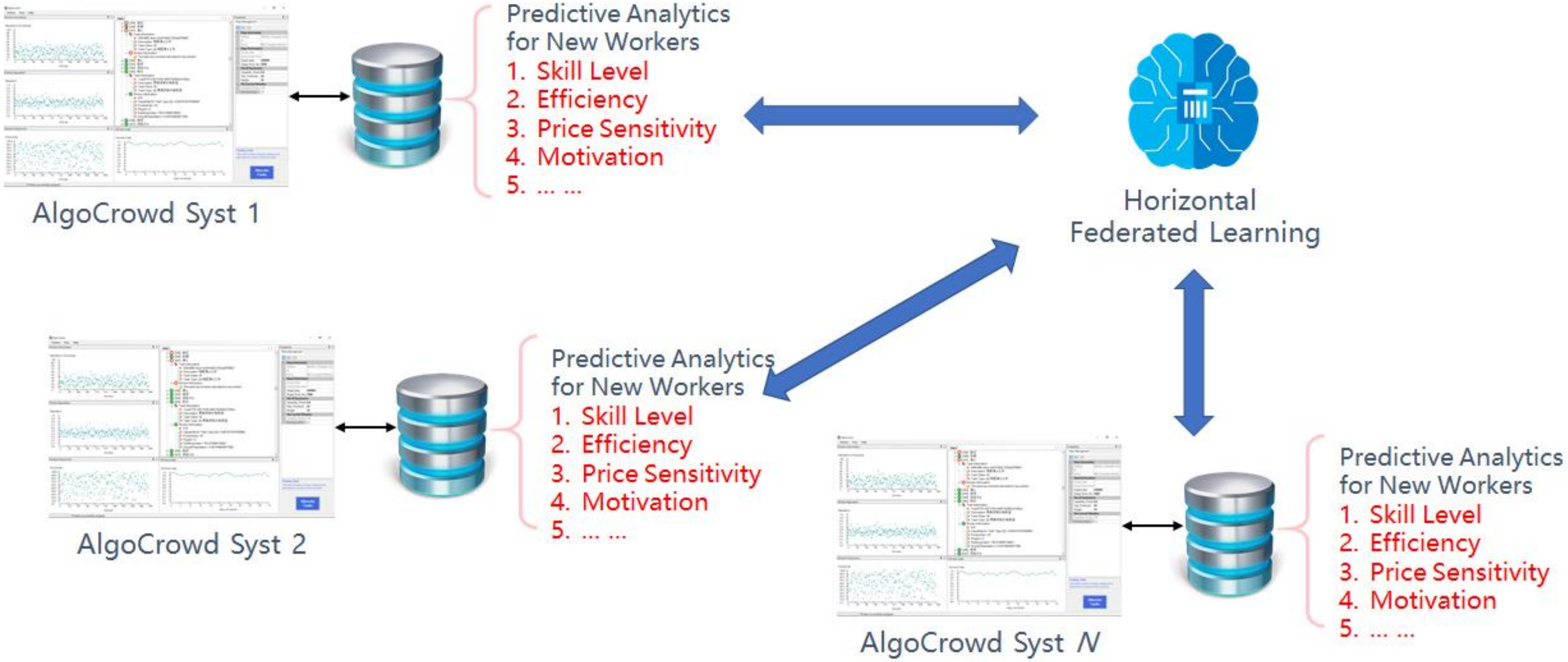
The Role of AI - Task-Worker Matching



System Architecture



Privacy Preservation by Design



Predictive Analytics for New Workers

1. Skill Level
2. Efficiency
3. Price Sensitivity
4. Motivation
5. … …

AlgoCrowd Syst 1

Horizontal Federated Learning

Predictive Analytics for New Workers

1. Skill Level
2. Efficiency
3. Price Sensitivity
4. Motivation
5. … …

AlgoCrowd Syst 2

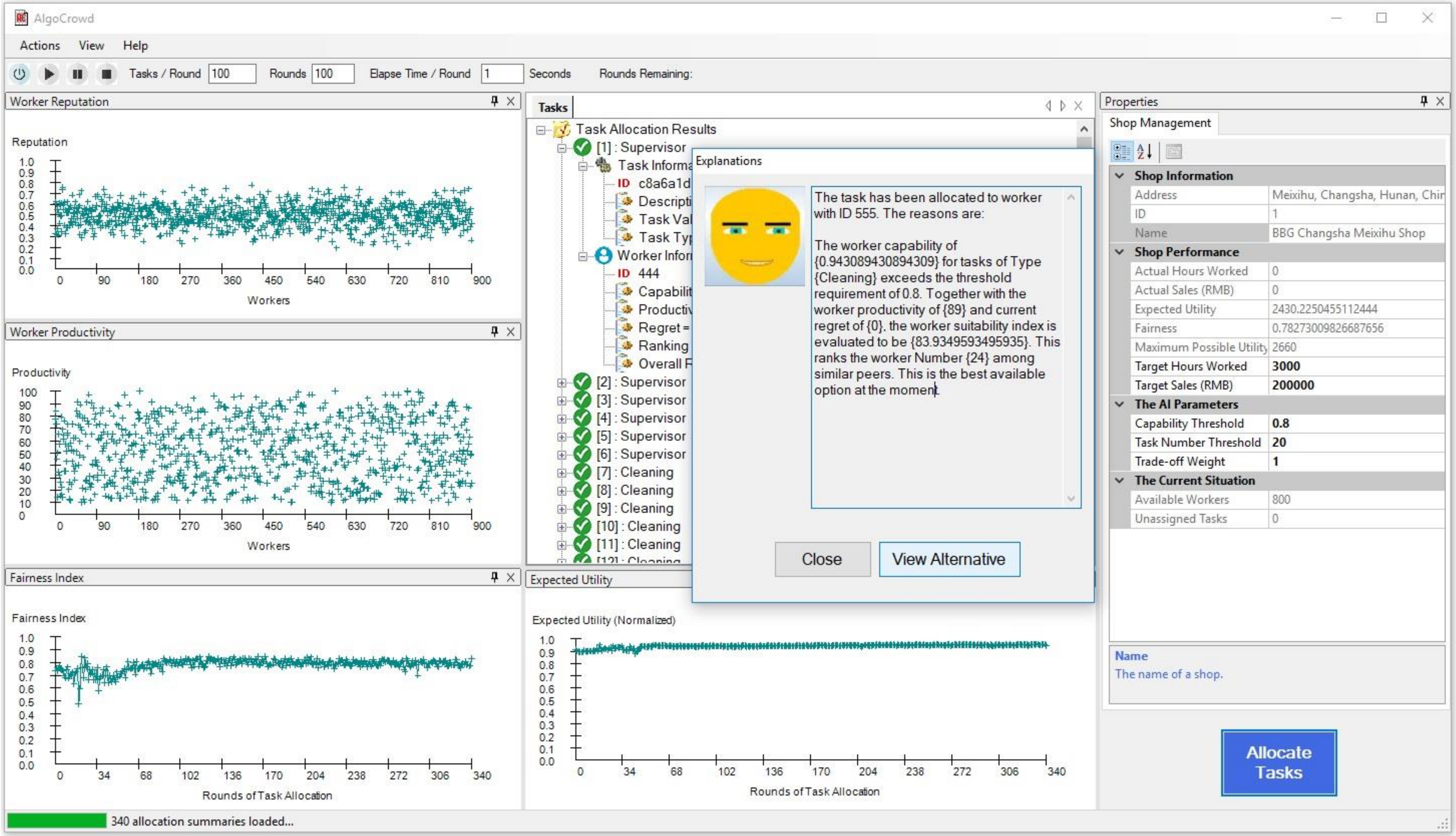
Predictive Analytics for New Workers

1. Skill Level
2. Efficiency
3. Price Sensitivity
4. Motivation
5. … …

AlgoCrowd Syst *N*

Responsible AI Practices Covered

- ✓ **Fairness**
- ✓ **Interpretability**
- ✓ **Privacy Preservation**



Tasks / Round 100 Rounds 100 Elapse Time / Round 1 Seconds Rounds Remaining:

Task Allocation Results
- [1] : Supervisor
  - Task Informa
    - ID c8a6a1d
    - Descripti
    - Task Val
    - Task Typ
  - Worker Infor
    - ID 444
    - Capabilit
    - Productiv
    - Regret =
    - Ranking
    - Overall F
- [2] : Supervisor
- [3] : Supervisor
- [4] : Supervisor
- [5] : Supervisor
- [6] : Supervisor
- [7] : Cleaning
- [8] : Cleaning
- [9] : Cleaning
- [10] : Cleaning
- [11] : Cleaning
- [12] : Cleaning

The task has been allocated to worker with ID 555. The reasons are:

The worker capability of {0.943089430894309} for tasks of Type {Cleaning} exceeds the threshold requirement of 0.8. Together with the worker productivity of {89} and current regret of {0}, the worker suitability index is evaluated to be {83.9349593495935}. This ranks the worker Number {24} among similar peers. This is the best available option at the moment.

Shop Information	
Address	Meixihu, Changsha, Hunan, Chir
ID	1
Name	BBG Changsha Meixihu Shop
Shop Performance	
Actual Hours Worked	0
Actual Sales (RMB)	0
Expected Utility	2430.2250455112444
Fairness	0.78273009826687656
Maximum Possible Utility	2660
Target Hours Worked	**3000**
Target Sales (RMB)	**200000**
The AI Parameters	
Capability Threshold	**0.8**
Task Number Threshold	**20**
Trade-off Weight	**1**
The Current Situation	
Available Workers	800
Unassigned Tasks	0

Name
The name of a shop.

340 allocation summaries loaded...

View demo video: https://youtu.be/vV3RsdCCETw

Towards Responsible AI

Yu Han

han.yu@ntu.edu.sg

Nanyang Assistant Professor
School of Computer Science and Engineering
Nanyang Technological University