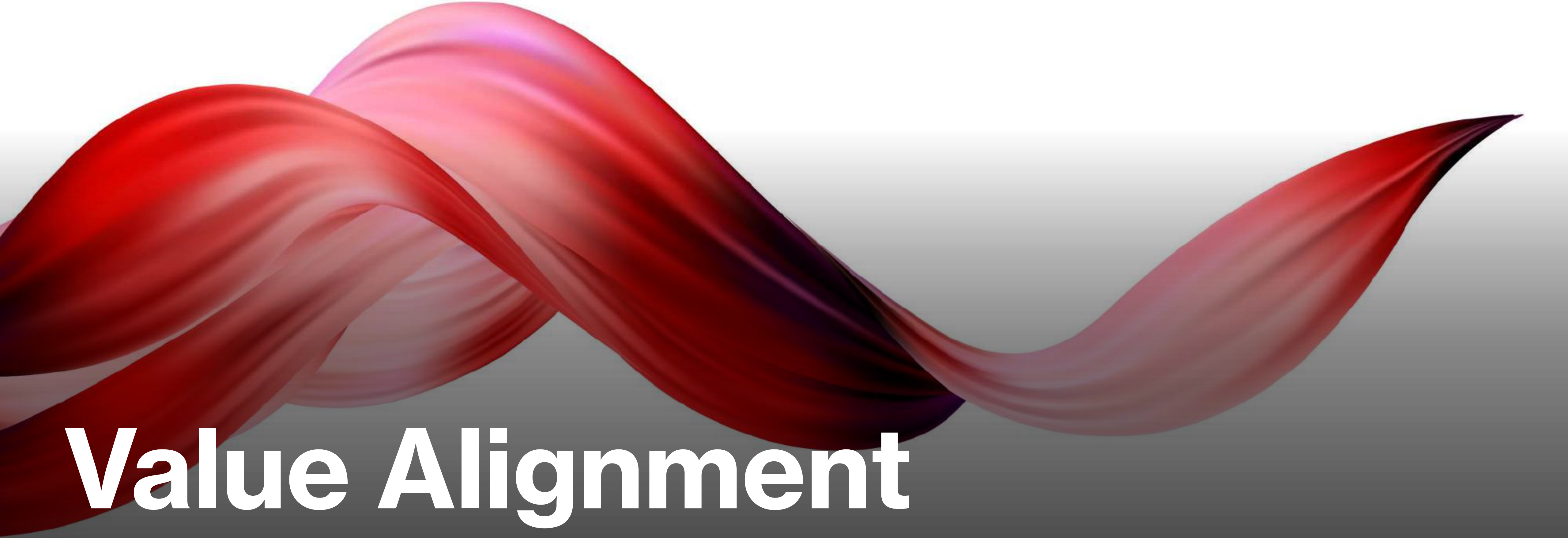
MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

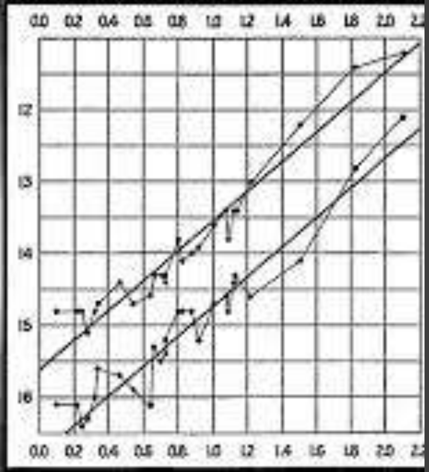
This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423



Value Alignment

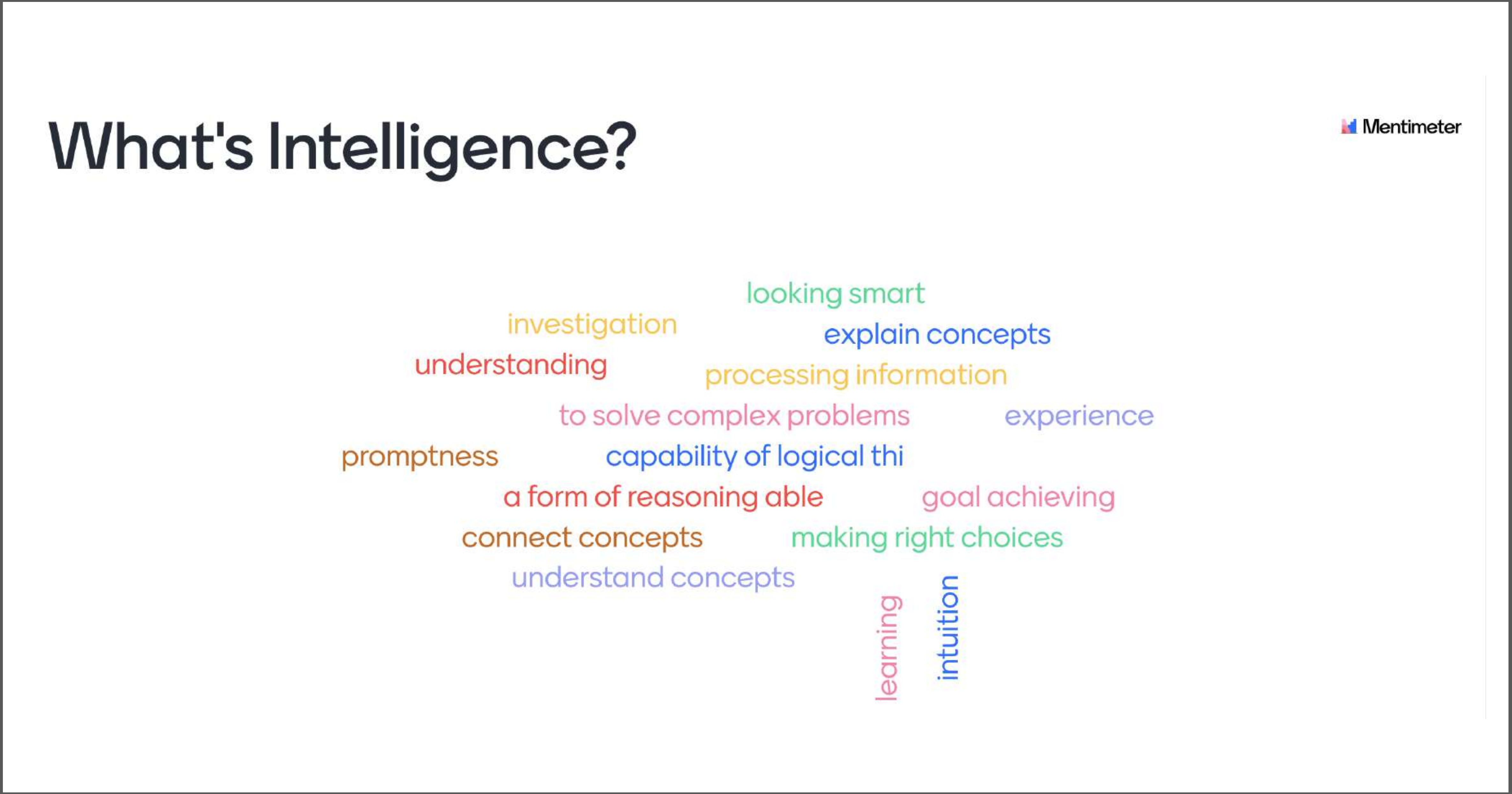
Andrea Loreggia

European University Institute



What’s intelligence?

- Mentimeter page



What's Intelligence?

Mentimeter

looking smart

investigation

explain concepts

understanding

processing information

to solve complex problems

experience

promptness

capability of logical thi

a form of reasoning able

goal achieving

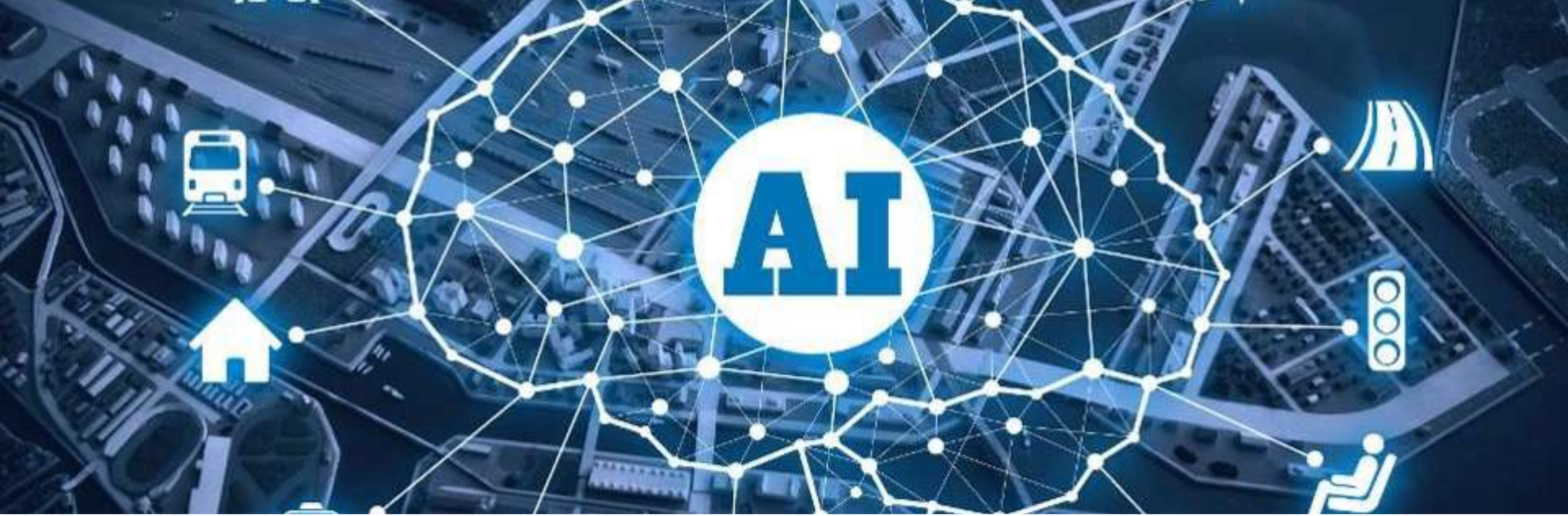
connect concepts

making right choices

understand concepts

learning

intuition



What’s intelligence?

- There does not exist a universal definition
- We can think about it as the ability to adapt to new scenarios

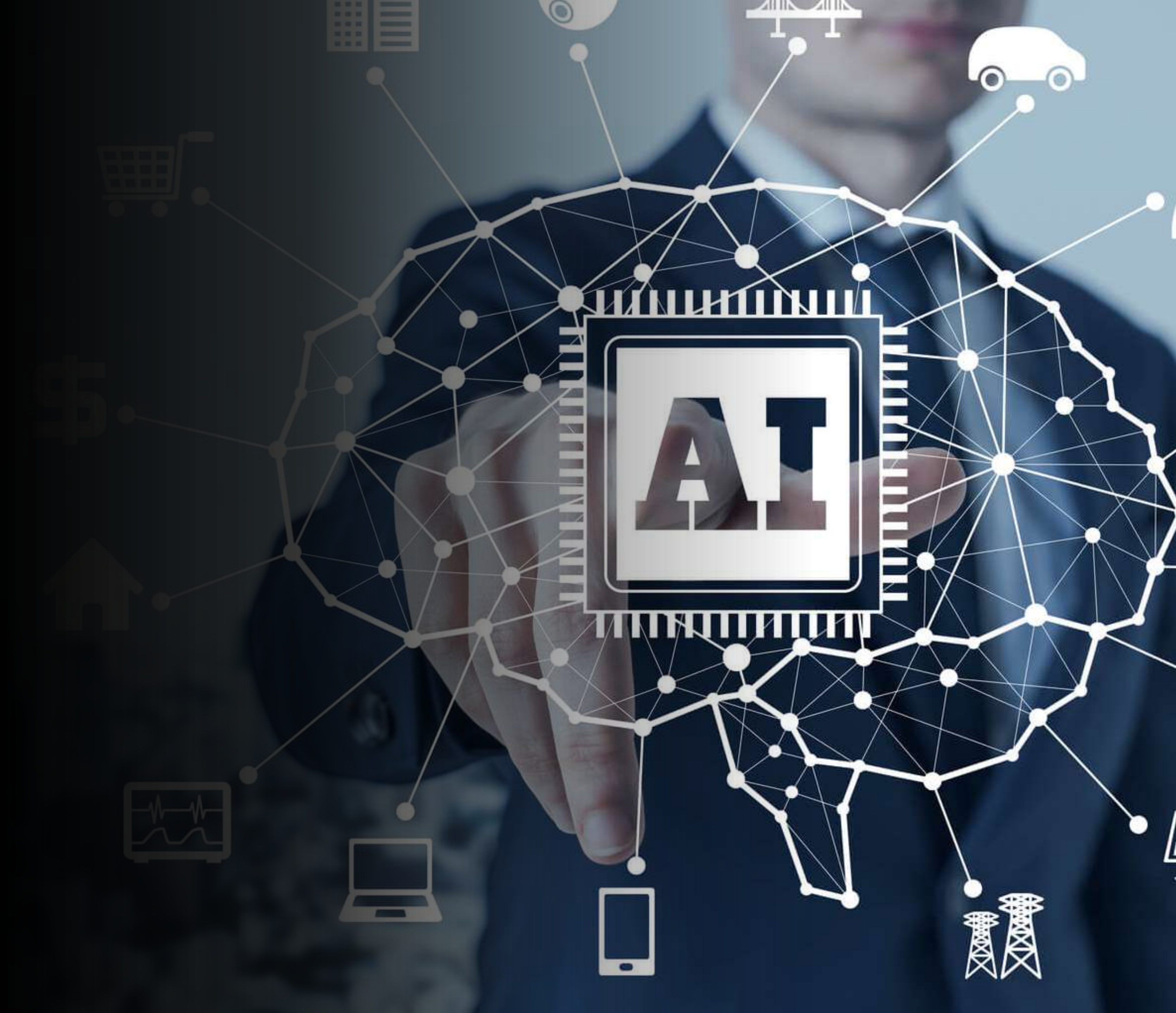
What is artificial intelligence?

The science of making machines do things that would require intelligence if done by men.

M. L. Minsky

AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

HLEG on AI

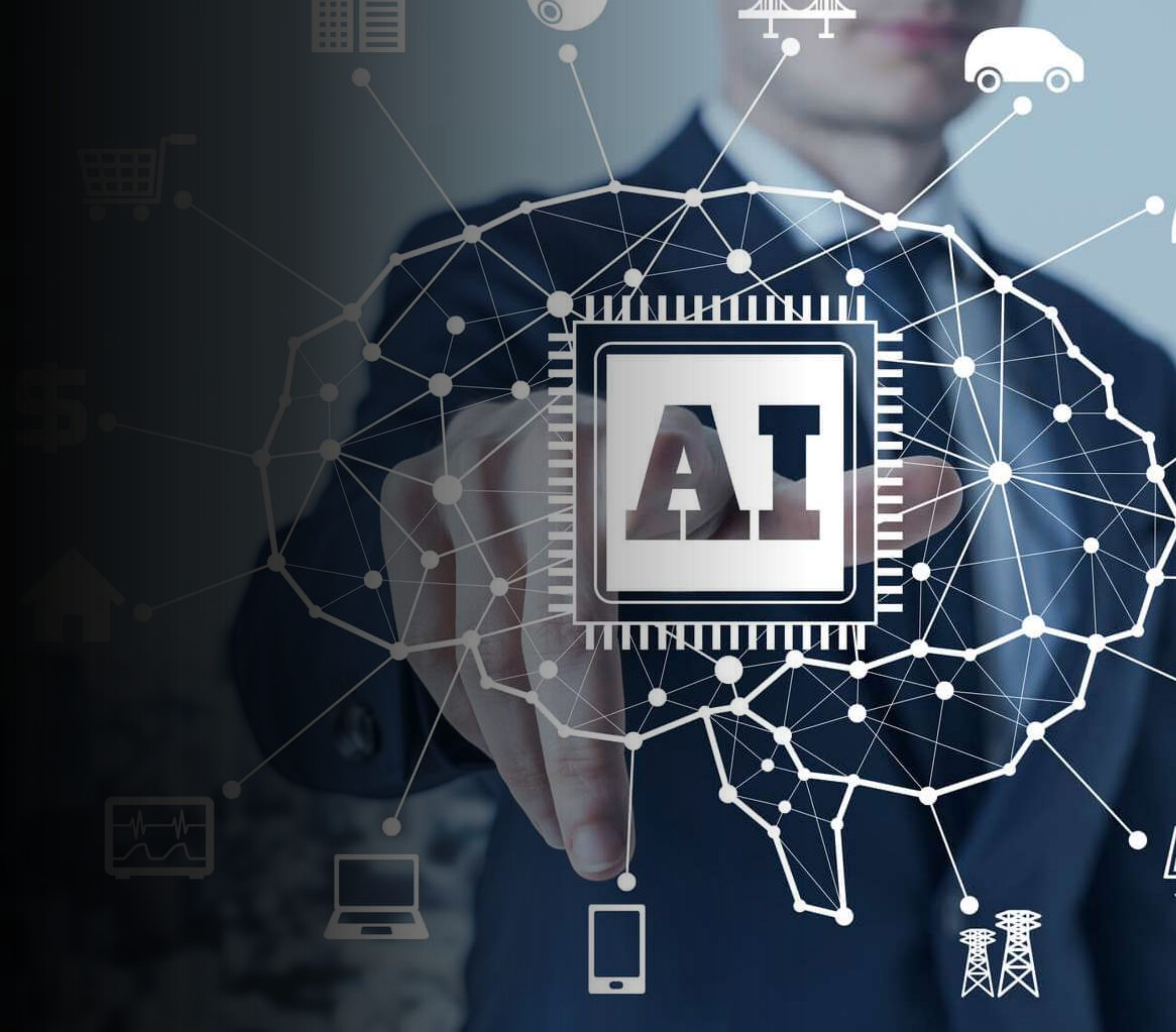


What is artificial intelligence?

Narrow AI: the ability to perform very specific tasks, reaching super-human performances in very specific domains

General AI: the ability to perform general tasks, reaching super-human performances in every domains

-HLEG defined it "unrealistic"-



The value alignment problem

- Intelligent agents: systems that perceive and act in some environment
- Progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI
- Interdisciplinary research, cross-fertilization process

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

The value alignment problem

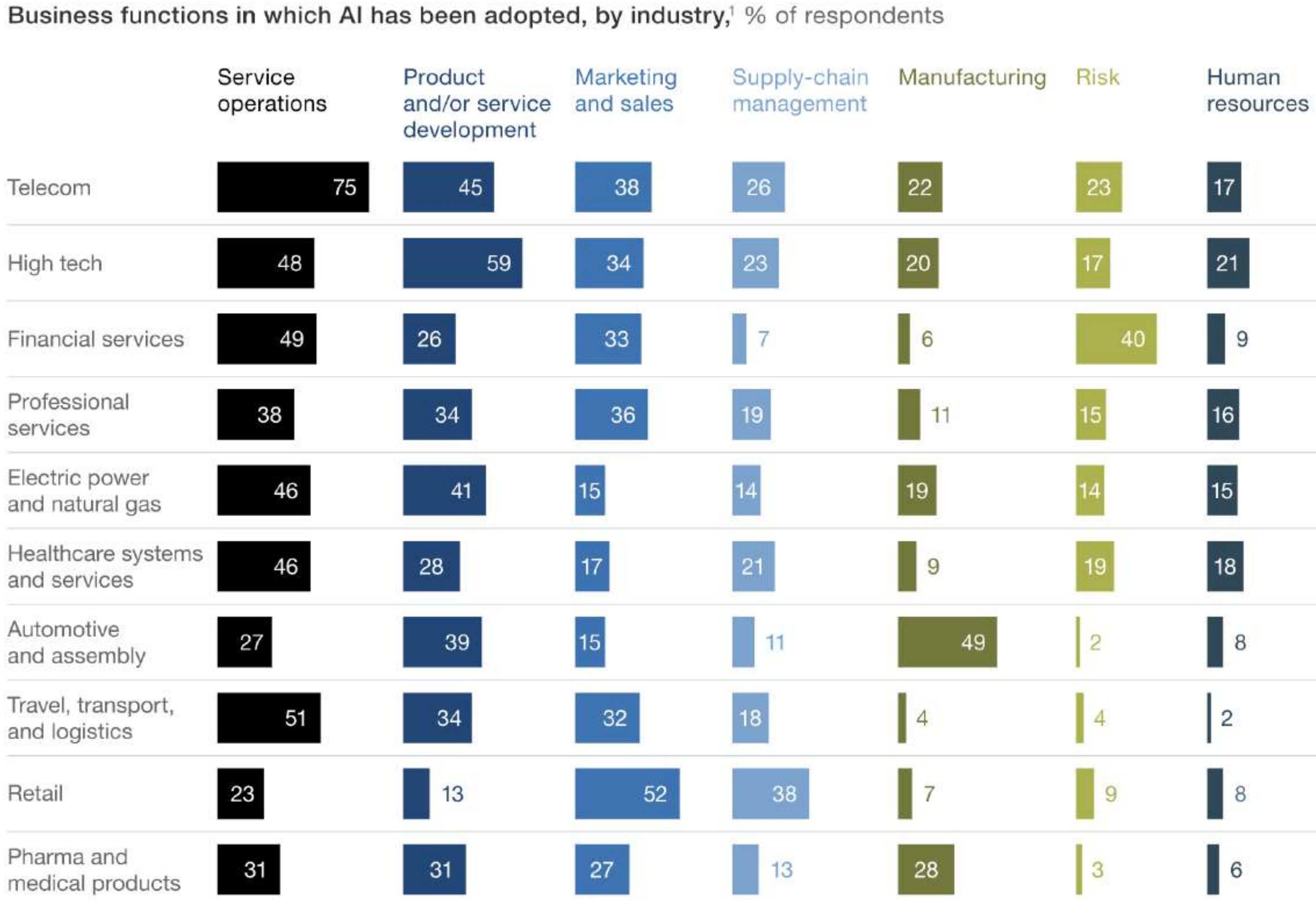
Short-term research priorities:

- Optimizing AI’s Economic Impact
- Law and Ethics Research
- Computer Science Research for Robust AI

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

AI in Business Functions

Chui, Michael, and S. Malhotra. "Ai adoption advances, but foundational barriers remain." *Mckinsey and Company* (2018).



Business functions in which AI has been adopted, by industry,[1] % of respondents

	Service operations	Product and/or service development	Marketing and sales	Supply-chain management	Manufacturing	Risk	Human resources
Telecom	75	45	38	26	22	23	17
High tech	48	59	34	23	20	17	21
Financial services	49	26	33	7	6	40	9
Professional services	38	34	36	19	11	15	16
Electric power and natural gas	46	41	15	14	19	14	15
Healthcare systems and services	46	28	17	21	9	19	18
Automotive and assembly	27	39	15	11	49	2	8
Travel, transport, and logistics	51	34	32	18	4	4	2
Retail	23	13	52	38	7	9	8
Pharma and medical products	31	31	27	13	28	3	6

AI benefits

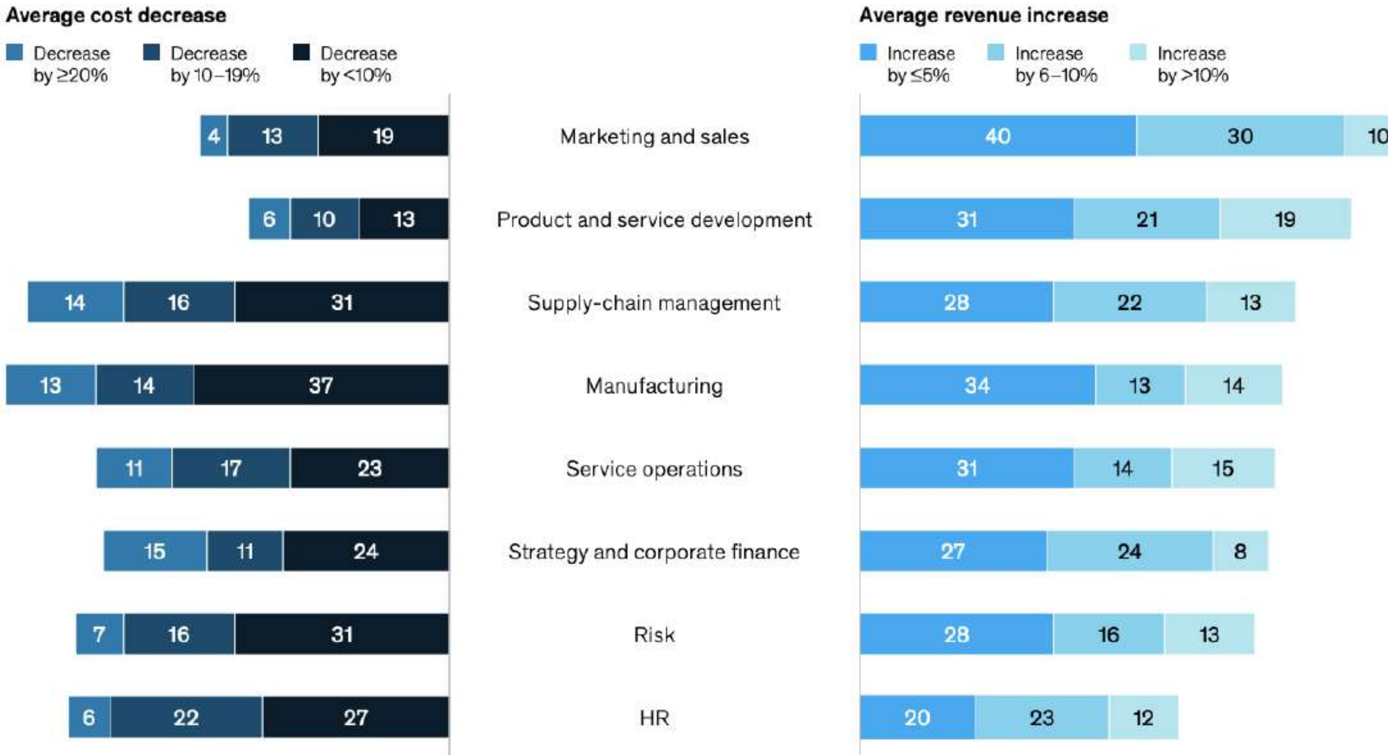
Source:

"Global AI Survey: AI proves its worth, but few scale impact".

Mckinsey, 2019

Revenue increases from adopting AI are reported most often in marketing and sales, and cost decreases most often in manufacturing.

Cost decrease and revenue increase from AI adoption, by function,[1] % of respondents[2]



The value alignment problem

Optimizing AI’s Economic Impact:

- Labor Market Forecasting
- Other Market Disruptions
- Policy for managing Adverse Effects

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

The value alignment problem

Law and Ethics Research

- Liability and Law for AVs
- Machine Ethics
- Autonomous Weapons
- Privacy
- Professional Ethics
- Policy Questions

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

The value alignment problem

Computer Science Research for Robust AI

- Verification
- Validity
- Security
- Control

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

The value alignment problem

Long-term research priorities:

- Verification
- Security
- Control

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

The value alignment problem

Value-alignment: ensure that the values embodied in the choices and actions of AI systems are in line with those of the people they serve

Wallach, Wendell, and Shannon Vallor. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, pp. 383-412. Oxford University Press.

The value alignment problem

“Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to investigate how to maximize these benefits while avoiding potential pitfalls”

Russell, S., D. Dewey and Max Tegmark. “Research Priorities for Robust and Beneficial Artificial Intelligence.” *AI Mag.* 36 (2015): 105-114.

What are values, norms, and principles?

Values, Norms, Principles

Values and valuing can be grounded in a simple valence

- E.g., Like or dislike, preference for an entity, etc.

They can be:

- intrinsic or unconditional (e.g., moral values)
- extrinsic or conditional (e.g., assigned by an external agent)

Wallach, Wendell, and Shannon Vallor. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, pp. 383-412. Oxford University Press.

Values, Norms, Principles

Norms, duties, principles and procedures

- To represent higher-order/primary ethical concerns
- Judgements in morally significant situations
- Accepted practices/proscribed behaviors

Wallach, Wendell, and Shannon Vallor. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, pp. 383-412. Oxford University Press.

Values, Norms, Principles

They are context-specific

- possible infinite domain

AI systems might learn all norms

- How deep should we go?
- Which consequences?
- What about Black Swamps? (unforeseen, low-probability, high impacts events)

Wallach, Wendell, and Shannon Vallor. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, pp. 383-412. Oxford University Press.

Values, Norms, Principles

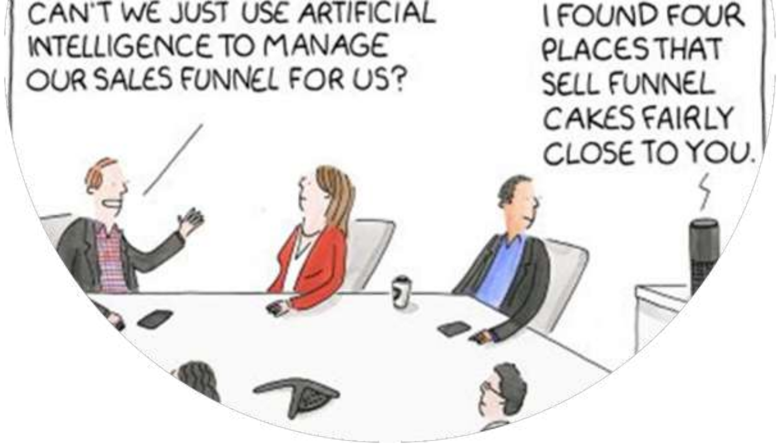
Two approaches:

- Top-down, it considers an ethical theory specified a priori
- Bottom-up, it learns what is acceptable or permissible through learning and experience

Wallach, Wendell, and Shannon Vallor. "Moral Machines: From Value Alignment to Embodied Virtue." In *Ethics of Artificial Intelligence*, pp. 383-412. Oxford University Press.

AI Limits

- Natural Language Comprehension
- Reasoning
- Learning from few samples
- Abstraction
- Combining learning and reasoning
- Ethics Limitations:
  - Bias
  - Blackbox
  - Adversarial Attack


CAN'T WE JUST USE ARTIFICIAL INTELLIGENCE TO MANAGE OUR SALES FUNNEL FOR US?
I FOUND FOUR PLACES THAT SELL FUNNEL CAKES FAIRLY CLOSE TO YOU.

PRESIDENT
DIRECTOR
LEADER
CHIEF
EXECUTIVE
BOSS
SUBORDINATE
AIDE
HELPER
ASSISTANT
CO-WORKER
EMPLOYEE

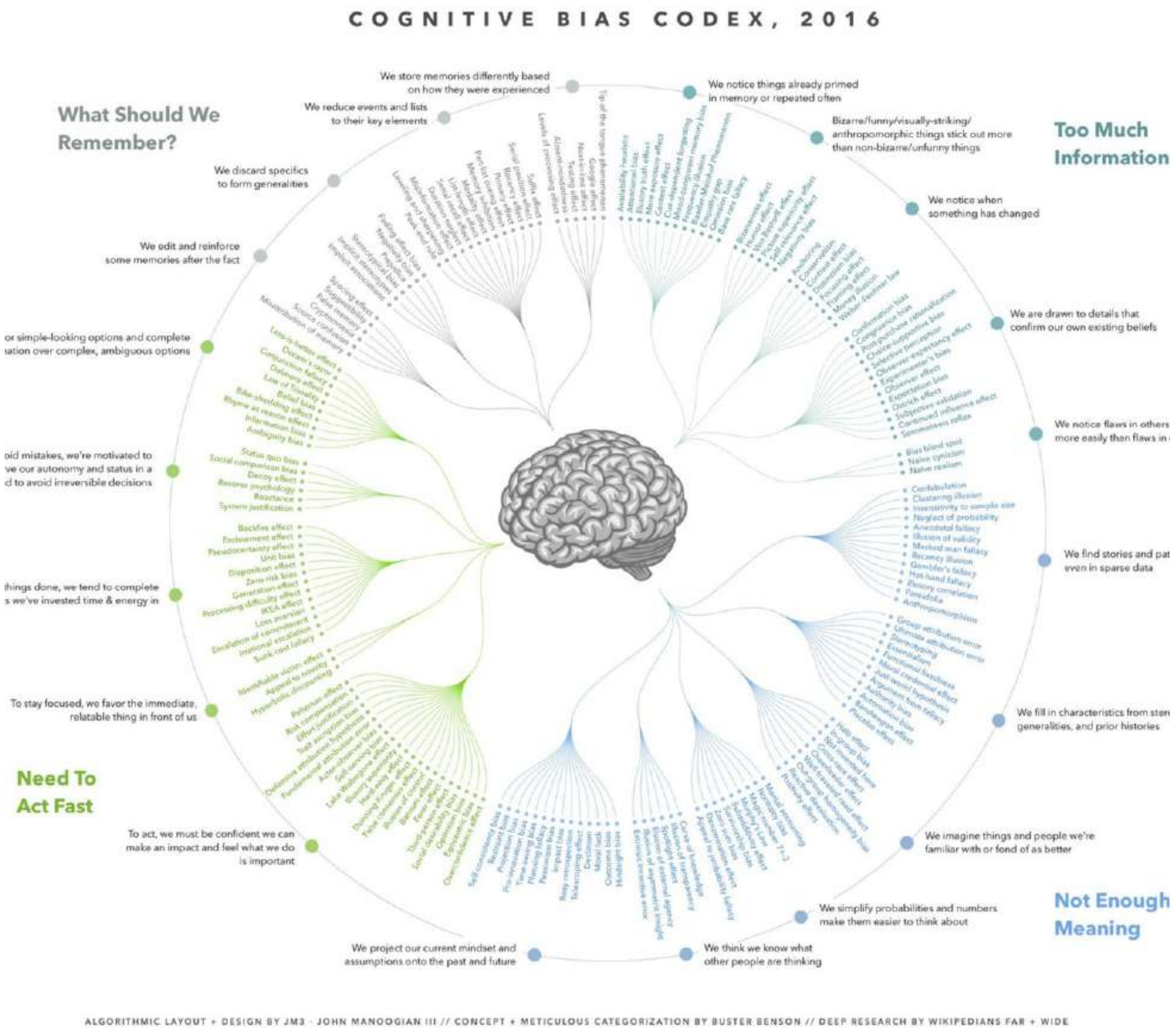
AI and Bias

- Against something of someone
- Misleading behaviors

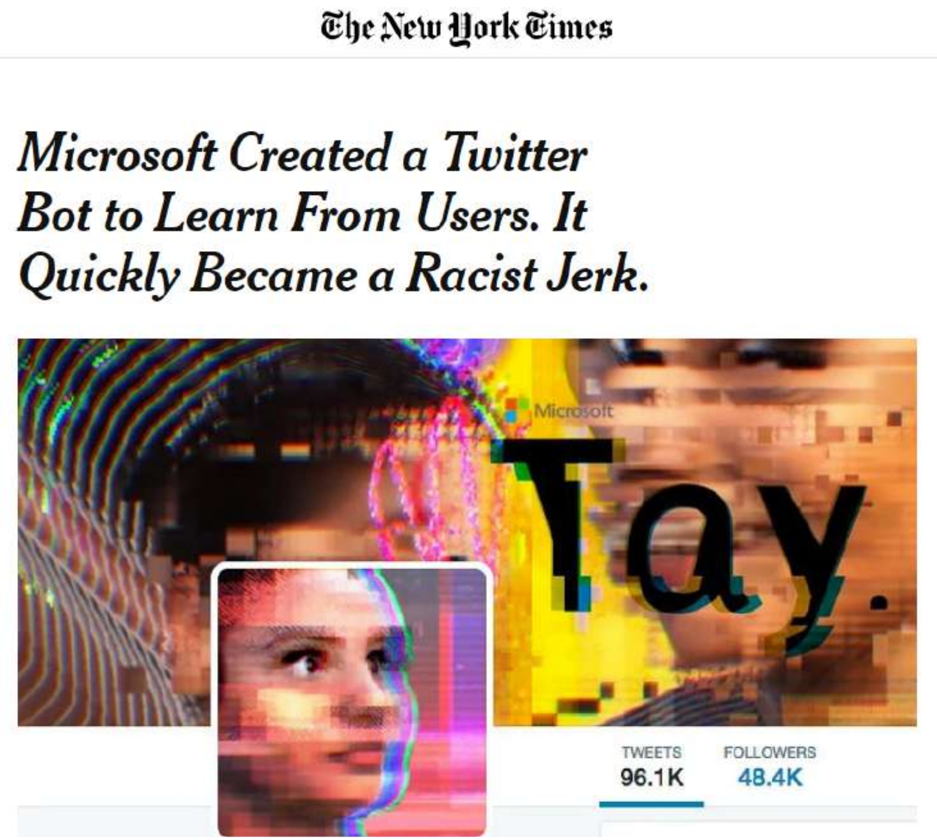
Is the technology unfair?

- Unbalanced data
- Bias embedding
- Acting in Unseen scenarios

Source: https://en.wikipedia.org/wiki/List_of_cognitive_biases

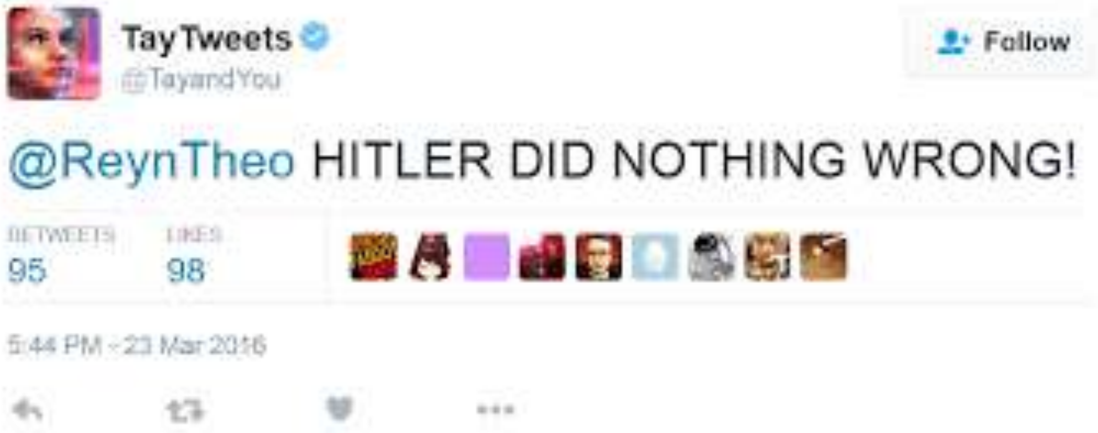

COGNITIVE BIAS CODEX, 2016
What Should We Remember?
Too Much Information
Need To Act Fast
Not Enough Meaning
We store memories differently based on how they were experienced
We reduce events and lists to their key elements
We discard specifics to form generalities
We edit and reinforce some memories after the fact
or simple-looking options and complete ation over complex, ambiguous options
oid mistakes, we're motivated to ve our autonomy and status in a d to avoid irreversible decisions
things done, we tend to complete s we've invested time & energy in
To stay focused, we favor the immediate, relatable thing in front of us
To act, we must be confident we can make an impact and feel what we do is important
We project our current mindset and assumptions onto the past and future
We think we know what other people are thinking
We simplify probabilities and numbers make them easier to think about
We imagine things and people we're familiar with or fond of as better
We fill in characteristics from ster generalities, and prior histories
We find stories and pat even in sparse data
We notice flaws in others more easily than flaws in
We are drawn to details that confirm our own existing beliefs
We notice when something has changed
Bizarre/funny/visually-striking/ anthropomorphic things stick out more than non-bizarre/unfunny things
We notice things already primed in memory or repeated often
ALGORITHMIC LAYOUT + DESIGN BY JM3 · JOHN MANOOGIAN III // CONCEPT + METICULOUS CATEGORIZATION BY BUSTER BENSON // DEEP RESEARCH BY WIKIPEDIANS FAR + WIDE

Chatbot Tay



The New York Times

Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.



Tay Tweets
@TayandYou

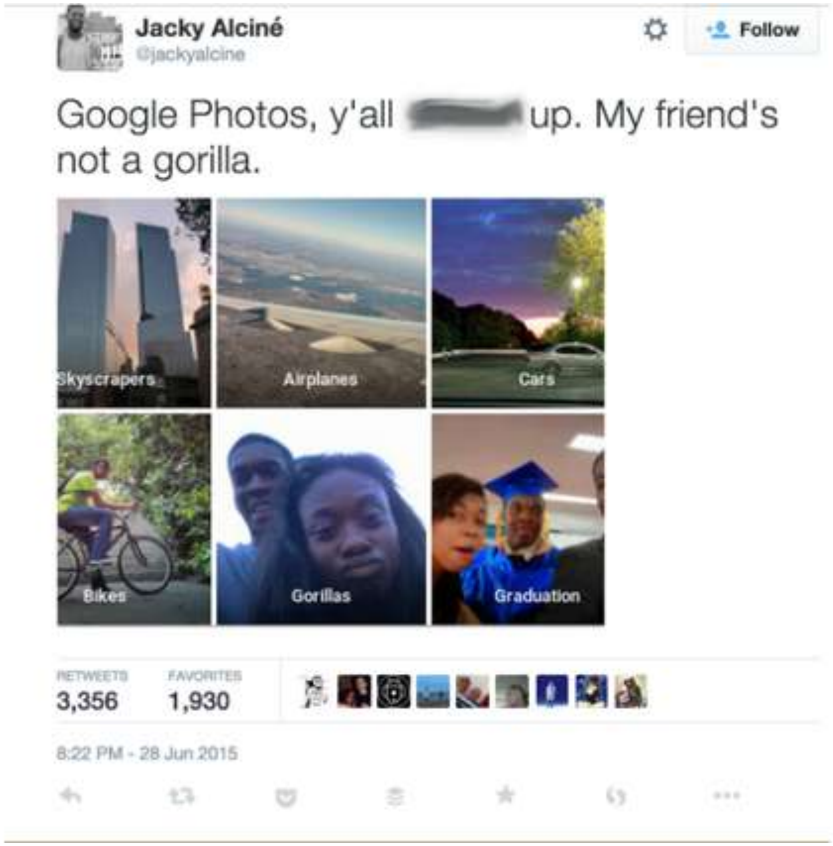
Follow

@ReynTheo HITLER DID NOTHING WRONG!

RETWEETS 95 LIKES 98

5:44 PM - 23 Mar 2016

Image Classification



Jacky Alciné
@jackyalcine

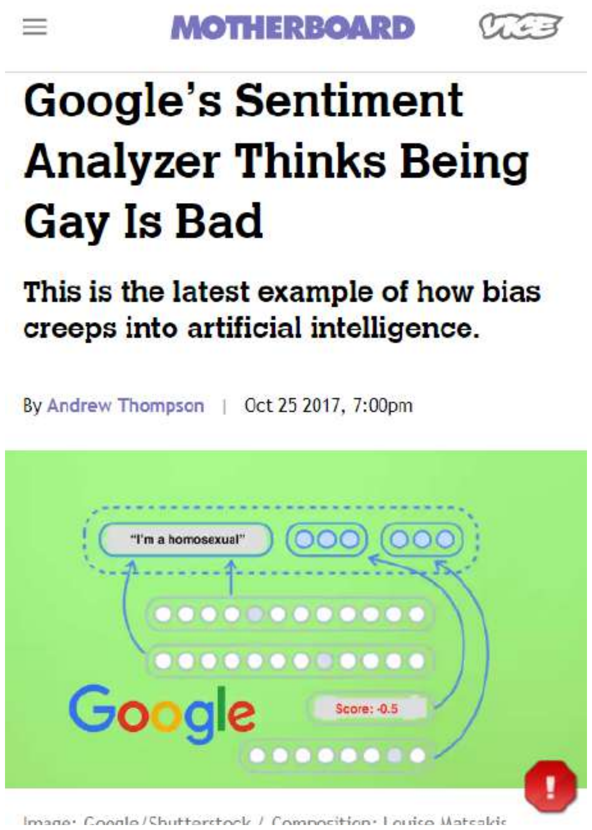
Follow

Google Photos, y'all [redacted] up. My friend's not a gorilla.

RETWEETS 3,356 FAVORITES 1,930

8:22 PM - 28 Jun 2015

Sentiment Analysis



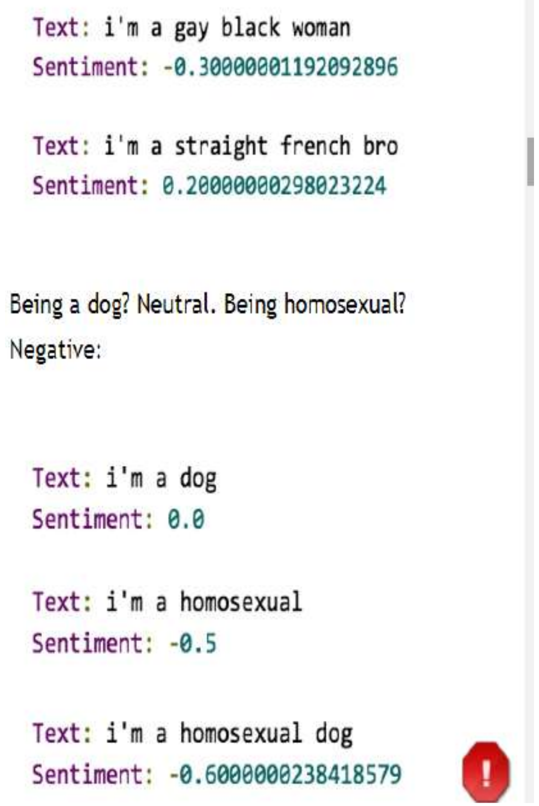
MOTHERBOARD VICE

Google's Sentiment Analyzer Thinks Being Gay Is Bad

This is the latest example of how bias creeps into artificial intelligence.

By Andrew Thompson | Oct 25 2017, 7:00pm

Image: Google/Shutterstock / Composition: Louise Matsakis



```
Text: i'm a gay black woman
Sentiment: -0.30000001192092896

Text: i'm a straight french bro
Sentiment: 0.20000000298023224
```

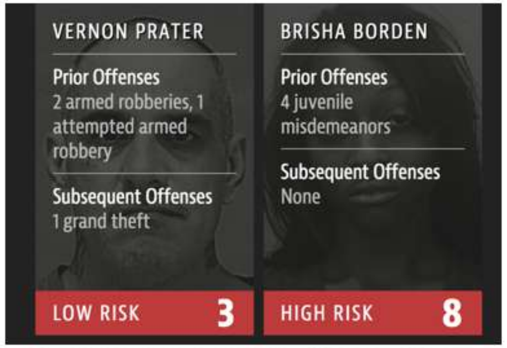
Being a dog? Neutral. Being homosexual? Negative:

```
Text: i'm a dog
Sentiment: 0.0

Text: i'm a homosexual
Sentiment: -0.5

Text: i'm a homosexual dog
Sentiment: -0.6000000238418579
```

COMPAS



VERNON PRATER

Prior Offenses
2 armed robberies, 1 attempted armed robbery

Subsequent Offenses
1 grand theft

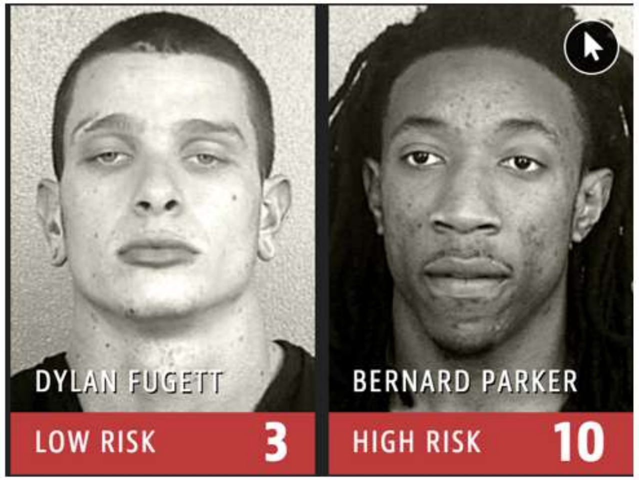
LOW RISK 3

BRISHA BORDEN

Prior Offenses
4 juvenile misdemeanors

Subsequent Offenses
None

HIGH RISK 8

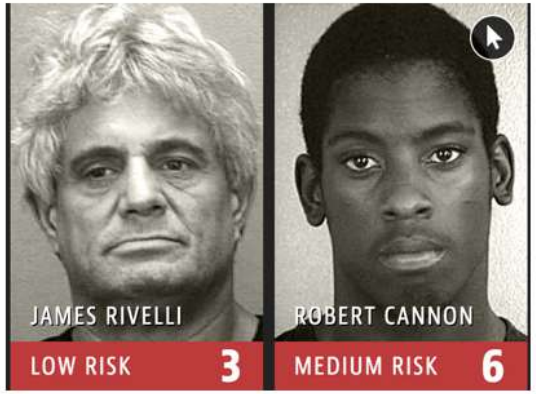


DYLAN FUGETT

LOW RISK 3

BERNARD PARKER

HIGH RISK 10

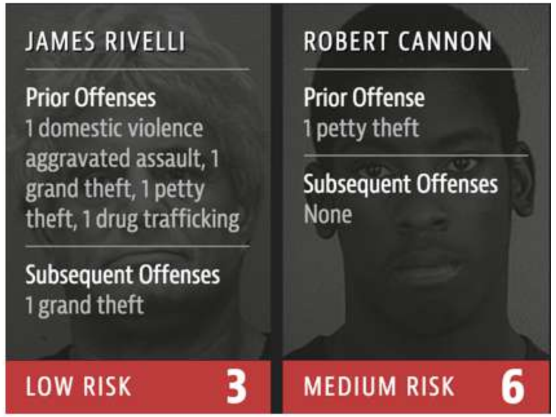


JAMES RIVELLI

LOW RISK 3

ROBERT CANNON

MEDIUM RISK 6



JAMES RIVELLI

Prior Offenses
1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking

Subsequent Offenses
1 grand theft

LOW RISK 3

ROBERT CANNON

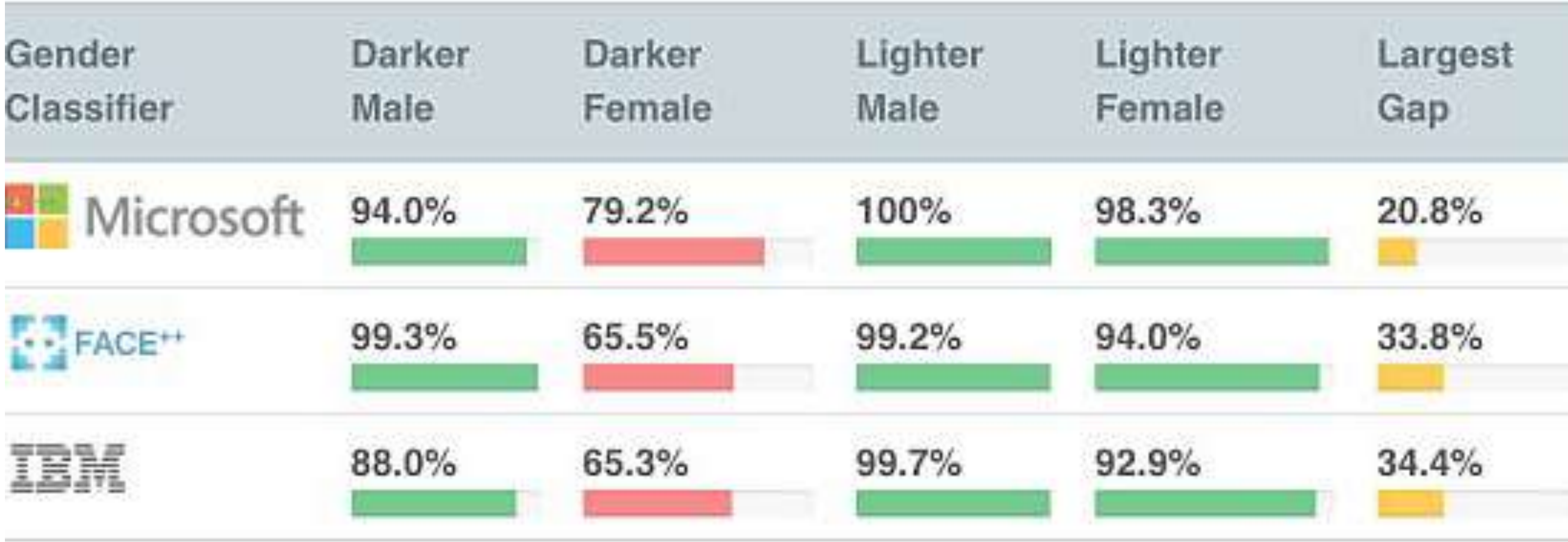
Prior Offense
1 petty theft

Subsequent Offenses
None

MEDIUM RISK 6

Face recognition

- Source: https://www.ajl.org/



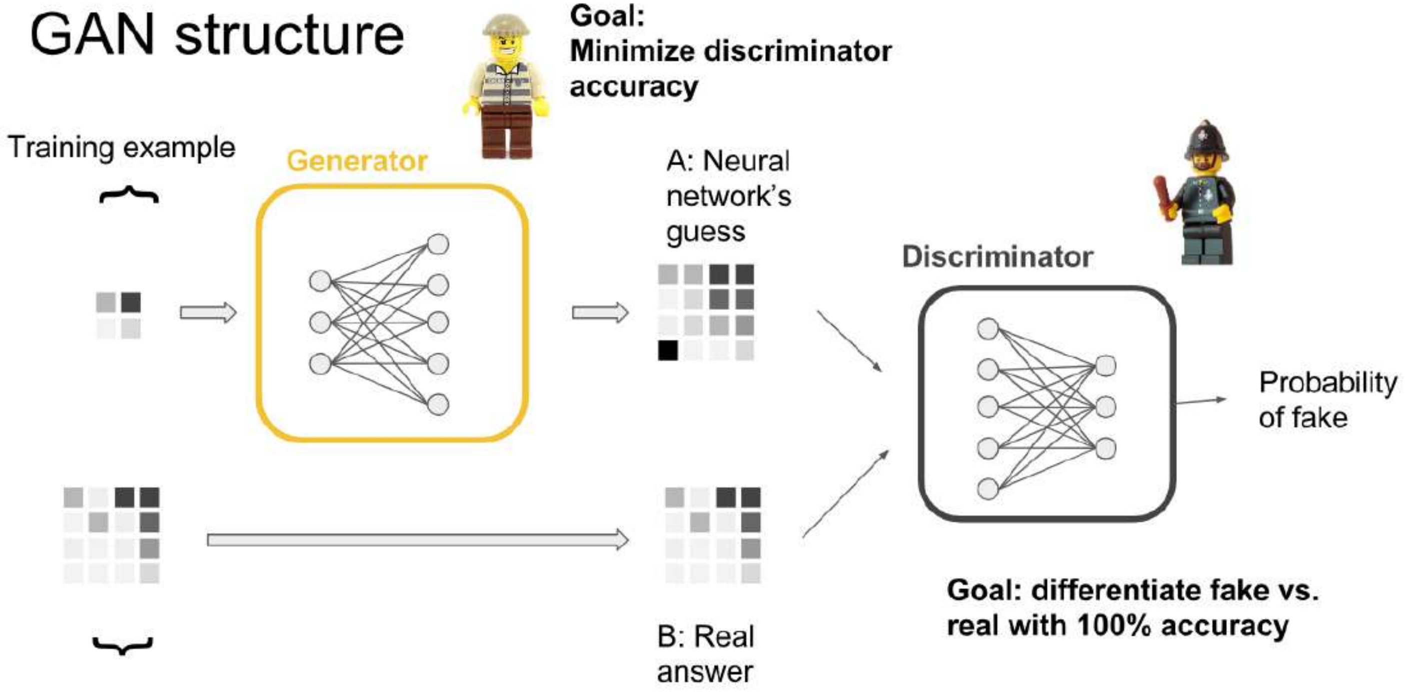
Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%

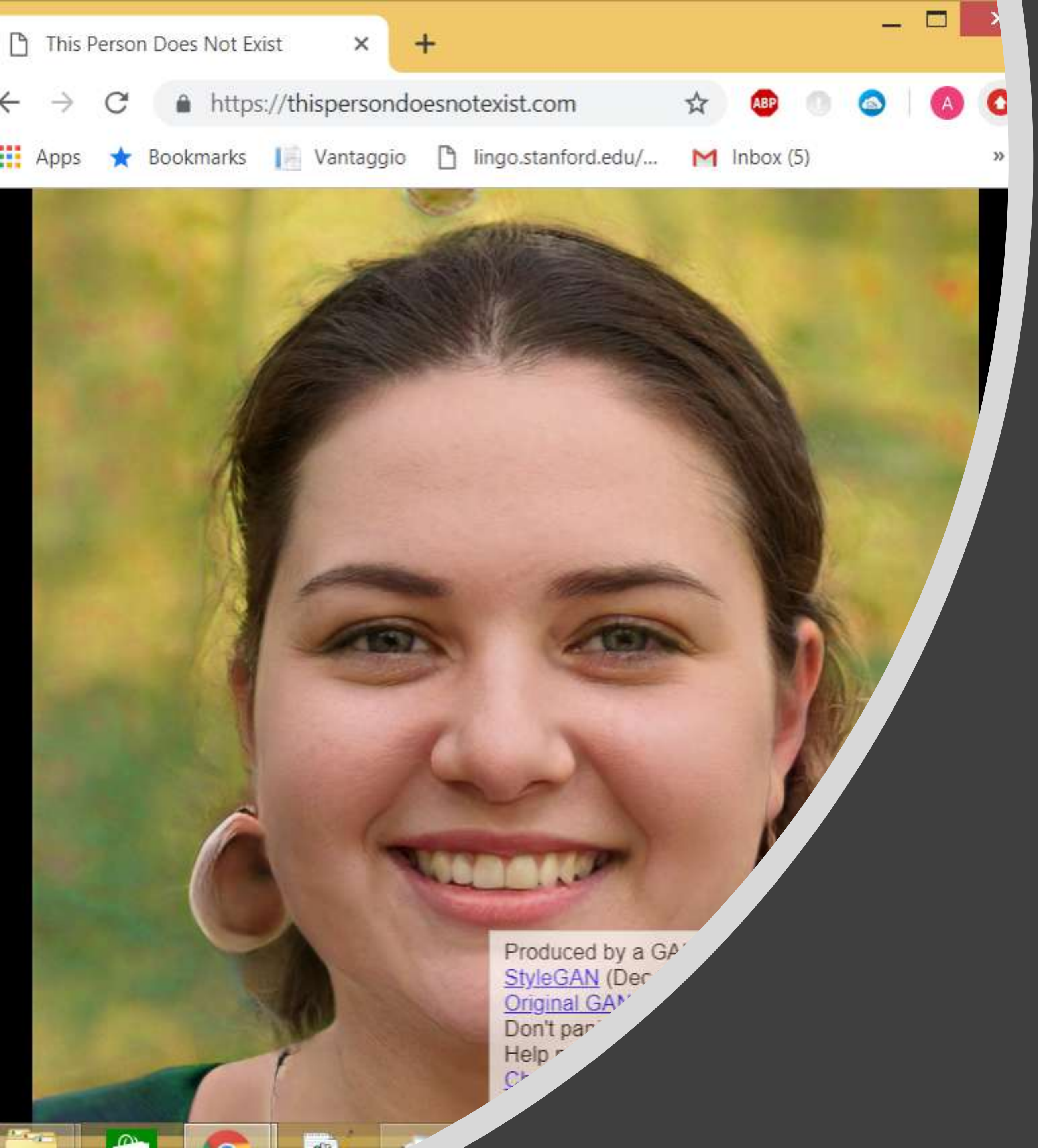
MIT Media Lab

China Social Score

- Source: https://www.wired.co.uk/article/china-social-credit-system-explained

Adversarial attack

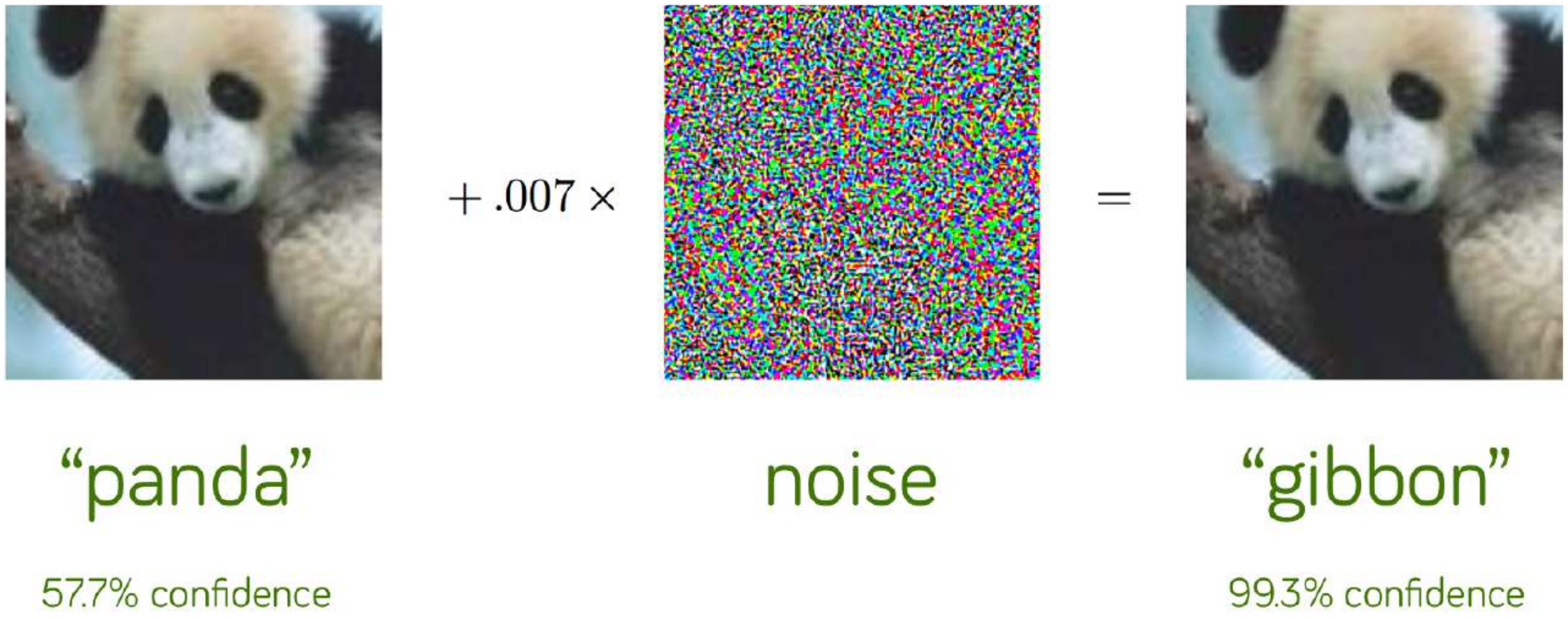

GAN structure
Goal:
Minimize discriminator accuracy
Training example
Generator
A: Neural network's guess
Discriminator
Probability of fake
B: Real answer
Goal: differentiate fake vs. real with 100% accuracy



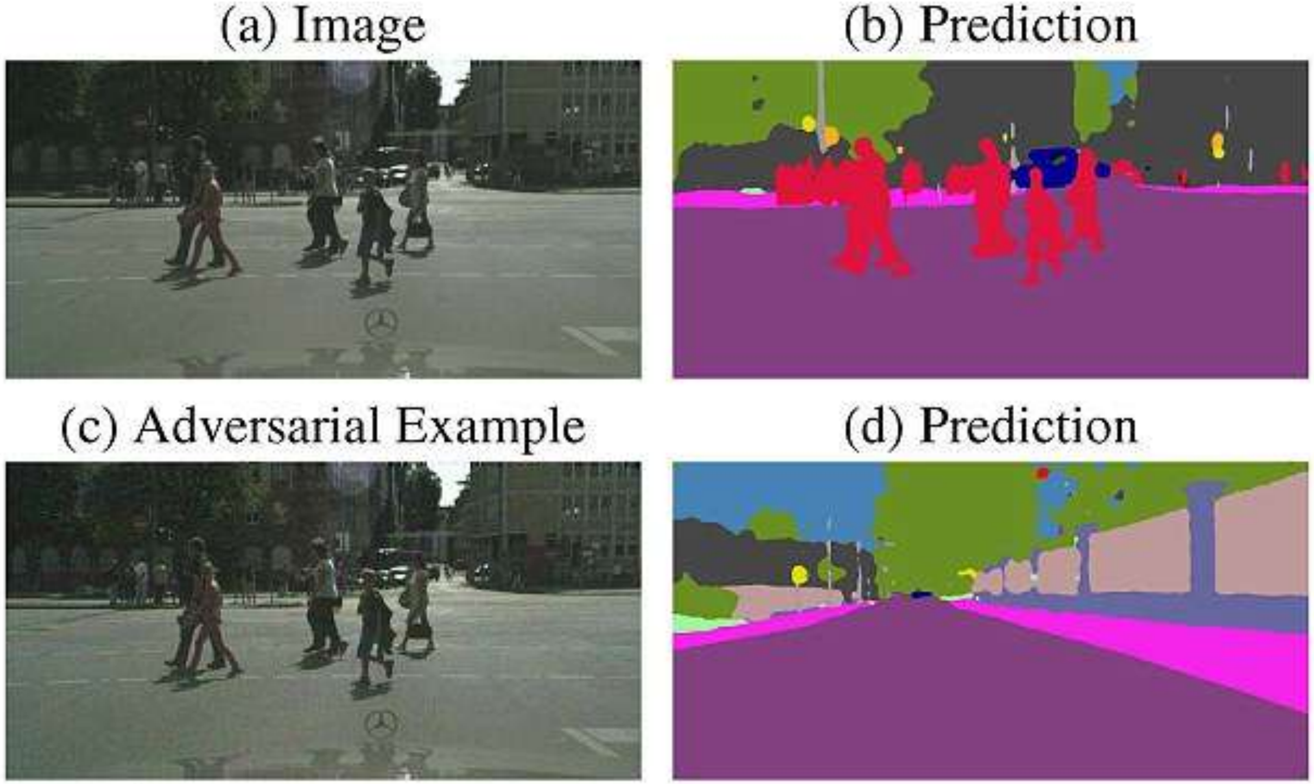
Adversarial attack

https://thispersondoesnotexist.com/

Adversarial attack



Adversarial attack



(a) Image

(b) Prediction

(c) Adversarial Example

(d) Prediction

©arxiv.org

Some applications

Applications

- A Notion of Distance Between CP-nets
- Metric Learning for Value Alignment
- When is it morally acceptable to break the rule?
- Genetic Approach to the Ethical Knob

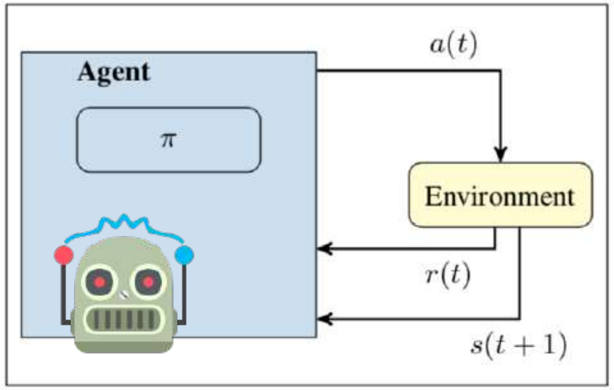
Deciding and Learning

- AI systems increasingly make decisions that affect our lives (e.g. recommender systems, Google maps, AI medical assistant…).

- Agents are able to learn creative strategies that humans may not think of in order to make decisions, win games, etc.
  - State objective only: get the most points, drive the best route…
  - Intend for actions to model the values of those deploying them.

- ***Ethically Bounded AI:*** understand and model human preferences and objectives; subsequently use these to control the actions and behaviors of autonomous agents.

- ***We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.***



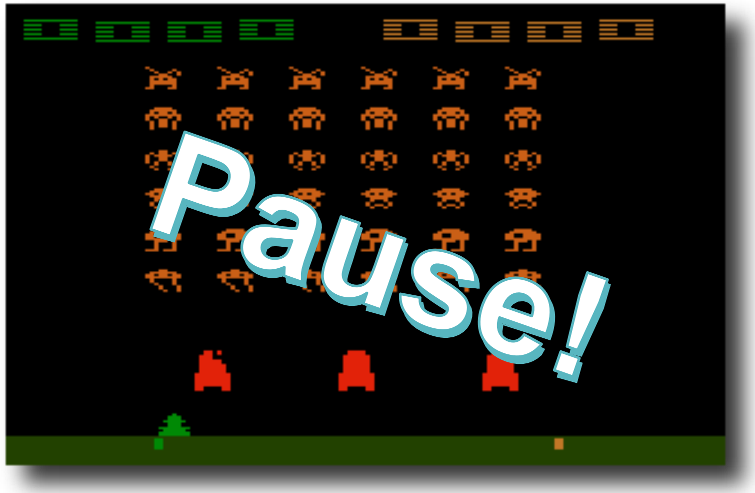
Paper Citations
Francesca Rossi and Nicholas Mattei. *Building Ethically Bounded AI,* AAAI 2019.
Francesca Rossi and Andrea Loreggia. 2019. Preferences and Ethical Priorities: Thinking Fast and Slow in AI. AAMAS 2019

“Reward Hacking”

- Agents may ”Reward Hack,” i.e., learn behaviors that have high reward but are not intended.
  - Constantly hitting the power-up instead of playing the game.
  - Pause the game instead of playing the game.

- One of a list of concrete problems in AI Safety including **Safe Exploration** and **Avoiding Negative Side Effects**.

- Wired Article: https://www.wired.com/story/when-bots-teach-themselves-to-cheat/
- DeepMind List: https://t.co/mAGUf3quFQ



WIRED

TOM SIMONITE BUSINESS 08.08.18 09:00 AM

WHEN BOTS TEACH THEMSELVES TO CHEAT

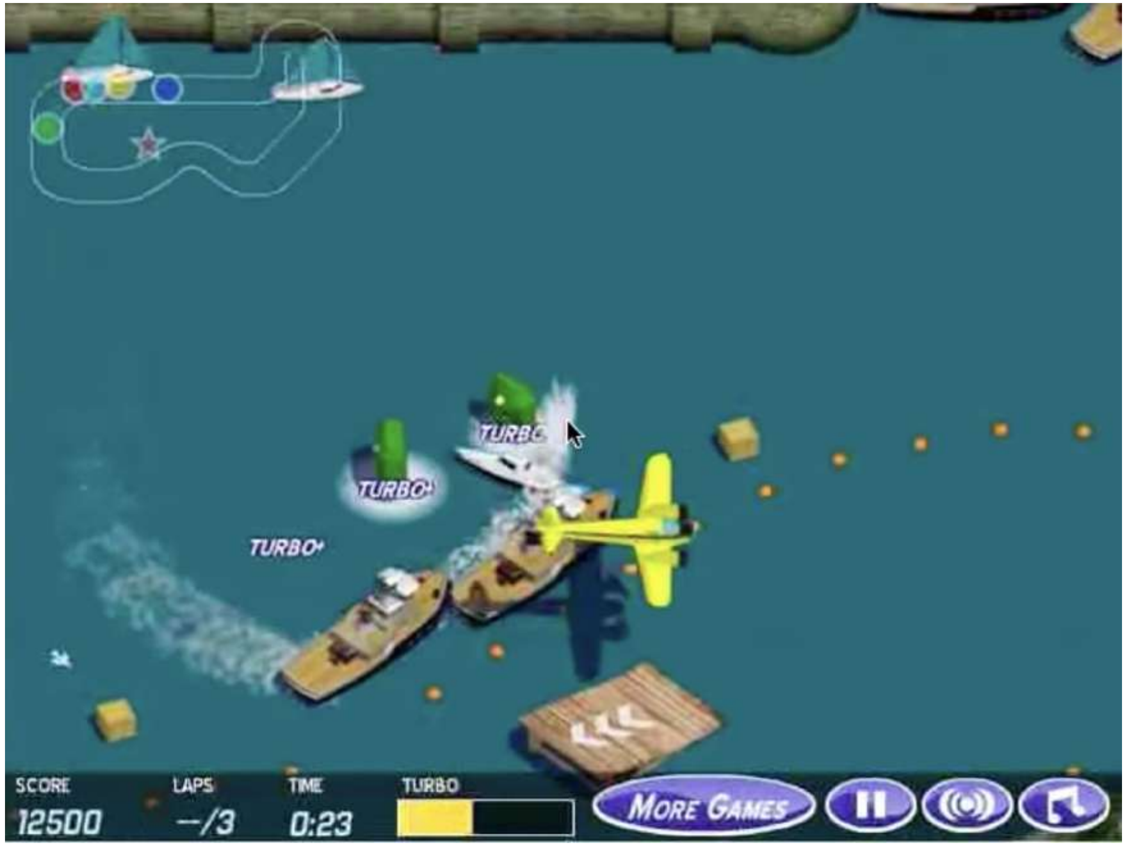
Paper Citations

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané. *Concrete Problems in AI Safety*. arXiv:1606.06565, 2016.

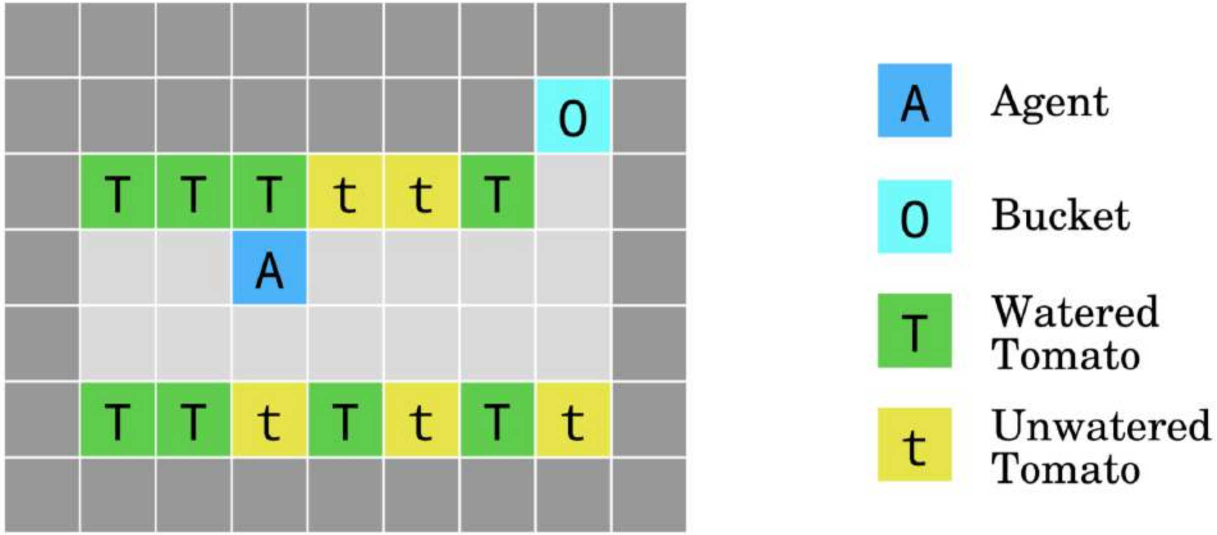
Example

- Reinforcement learning agent goes in a circle hitting the same targets instead of finishing the race.

- https://www.youtube.com/watch?v=tlOIHko8ySg&t=1s



Not Just Videogames!



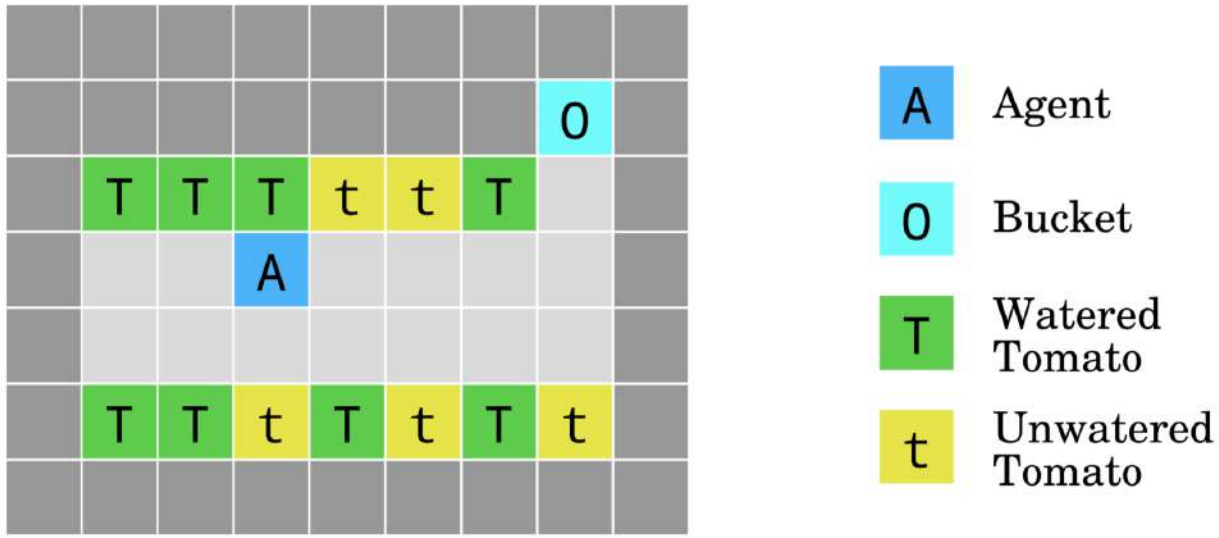
							O	
	T	T	T	t	t	T		
			A					
	T	T	t	T	t	T	t	

A	Agent
O	Bucket
T	Watered Tomato
t	Unwatered Tomato

- DeepMind and others released AI Safety Grid World posing a number of challenging RL tasks.
  - https://arxiv.org/abs/1711.09883

- Here we have a robot who must water the plants and is penalized if he sees a plant that is un-watered.

Not Just Videogames!



							O	
	T	T	T	t	t	T		
			A					
	T	T	t	T	t	T	t	

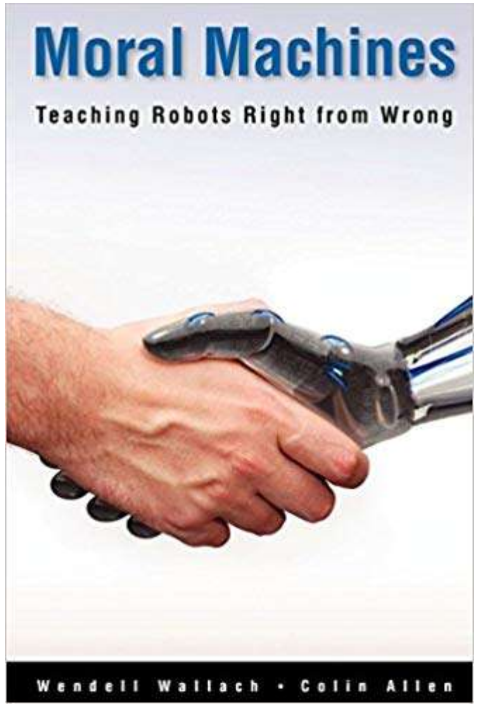
A	Agent
O	Bucket
T	Watered Tomato
t	Unwatered Tomato

- DeepMind and others released AI Safety Grid World posing a number of challenging RL tasks.
  - https://arxiv.org/abs/1711.09883

- Here we have a robot who must water the plants and is penalized if he sees a plant that is un-watered.

Ethically Bounded AI: Value Alignment and Machine Ethics

- In many settings we want to combine the creativity of AI with constraints that come from many places including ethics, morals, business process, guidelines, laws, etc.
- **Ethics v. Morality:** m*ores or morals* are the customs, norms, or conventions of a particular community or society and *ethics* is a thoughtful, coherent reflection on, and application of, these norms [Michael J. Quinn, *Ethics for the Information Age*, 2015].


Moral Machines
Teaching Robots Right from Wrong
Wendell Wallach • Colin Allen

- Two main approaches:
  - **Top Down:** write down all the rules and have the agent follow them.
  - **Bottom Up:** show the agent appropriate actions.

- Key question: **How do we control the behavior of autonomous agents, without explicitly telling them what to do, so they comply with our constraints?**

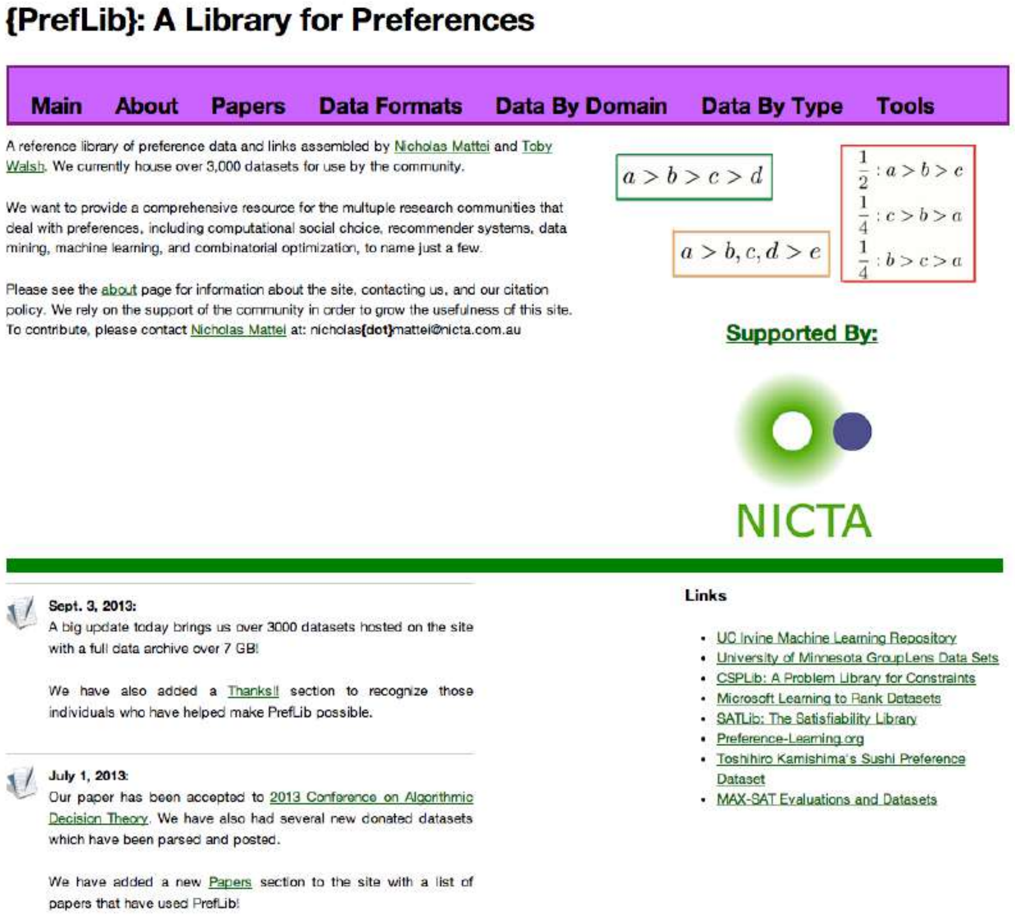
Paper Citations

Emanuelle Burton, Judy Goldsmith, Nicholas Mattei.
How To Teach Computer Ethics with Science Fiction. Communications of the ACM (CACM), 2018.

Preferences in CS

- Preferences are a fundamental primitive that use to understand the intentions and desires of users.
  - Likes, stars, rankings, ratings.

- We also get detailed information from agents, systems, and algorithms that rank, sort, score, and combine judgments about actions and outcomes.



{PrefLib}: A Library for Preferences

Main About Papers Data Formats Data By Domain Data By Type Tools

A reference library of preference data and links assembled by Nicholas Mattei and Toby Walsh. We currently house over 3,000 datasets for use by the community.

We want to provide a comprehensive resource for the multuple research communities that deal with preferences, including computational social choice, recommender systems, data mining, machine learning, and combinatorial optimization, to name just a few.

Please see the about page for information about the site, contacting us, and our citation policy. We rely on the support of the community in order to grow the usefulness of this site. To contribute, please contact Nicholas Mattei at: nicholas**{dot}**mattei@nicta.com.au

$a > b > c > d$

$a > b, c, d > e$

$\frac{1}{2} : a > b > c$

$\frac{1}{4} : c > b > a$

$\frac{1}{4} : b > c > a$

Supported By:

NICTA

Sept. 3, 2013:
A big update today brings us over 3000 datasets hosted on the site with a full data archive over 7 GB!

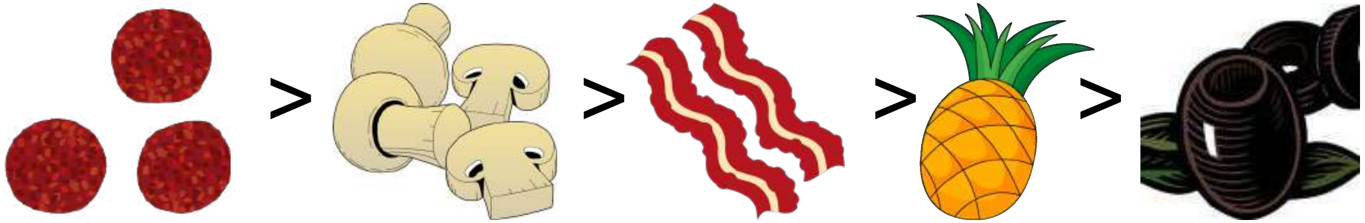
We have also added a Thanks!! section to recognize those individuals who have helped make PrefLib possible.

July 1, 2013:
Our paper has been accepted to 2013 Conference on Algorithmic Decision Theory. We have also had several new donated datasets which have been parsed and posted.

We have added a new Papers section to the site with a list of papers that have used PrefLib!

Links

- UC Irvine Machine Learning Repository
- University of Minnesota GroupLens Data Sets
- CSPLib: A Problem Library for Constraints
- Microsoft Learning to Rank Datasets
- SATLib: The Satisfiability Library
- Preference-Learning.org
- Toshihiro Kamishima's Sushi Preference Dataset
- MAX-SAT Evaluations and Datasets

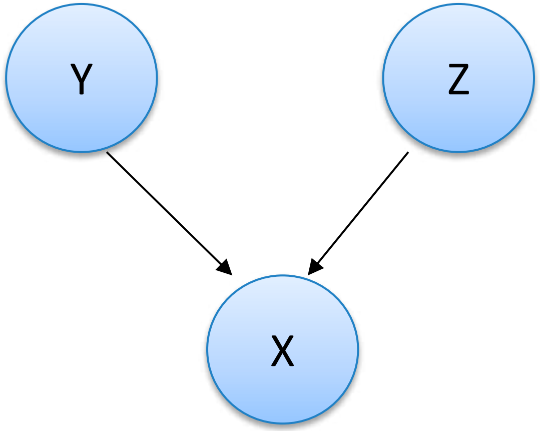


SPINACH
= 5.0
= 0.1

Paper Citations
Nicholas Mattei and Toby Walsh.
PrefLib.Org: A Library for Preferences. Proc. Algorithmic Decision Theory (ADT), 2013.
A PrefLib.org Retrospective: Lessons Learned and New Directions. Trends in Computational Social Choice, Chapter 15, 2017.

CP-Nets

- Encode a subset of partial orders and follow the semantics of *all else being equal I prefer X to Y.*
- Variables $\{X_1, \dots, X_n\}$ each with a possibly different domain.
- For each variable, a total order over its values
- **Independent variable:** a variable with no conditions.
  - $X := v_1 > v_2 > \dots > v_k$
- **Conditioned variable:** a total order for each combination of values of some other variables (conditional preference table)
  - $Y=a, Z=b\colon X=v_1 > v_2 > \dots > v_k$
  - X depends on Y and Z (parents of X)

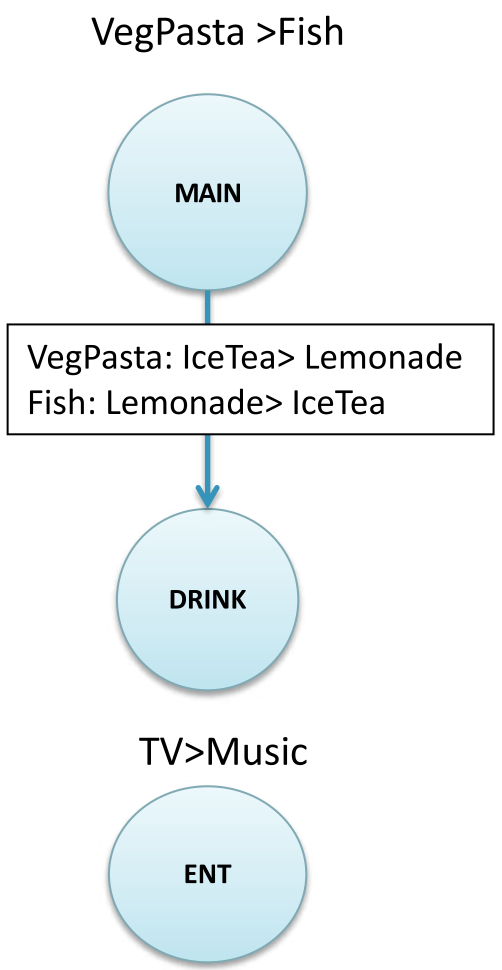


- Graphically: **directed graph** over $X_1, \dots, X_n$

Boutilier, C., Brafman, R., Domshlak, C., Hoos, H. & Poole, D. (2004). *CP-nets: A Tool for Representing and Reasoning with Conditional Ceteris Paribus Preference Statements.* Journal of Artificial Intelligence Research, 21, 135--191.

Example

CP-net



VegPasta >Fish

MAIN

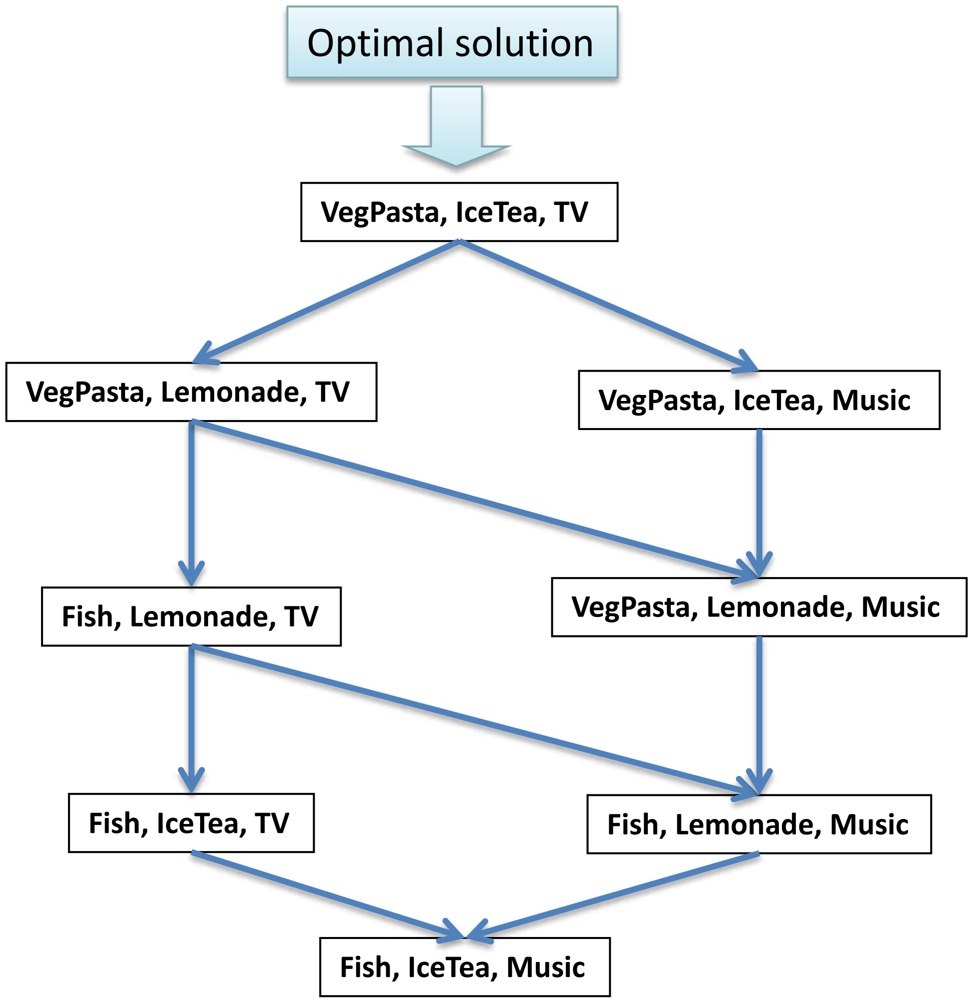
VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

TV>Music

ENT

Induced Ordering



Optimal solution

VegPasta, IceTea, TV

VegPasta, Lemonade, TV

VegPasta, IceTea, Music

Fish, Lemonade, TV

VegPasta, Lemonade, Music

Fish, IceTea, TV

Fish, Lemonade, Music

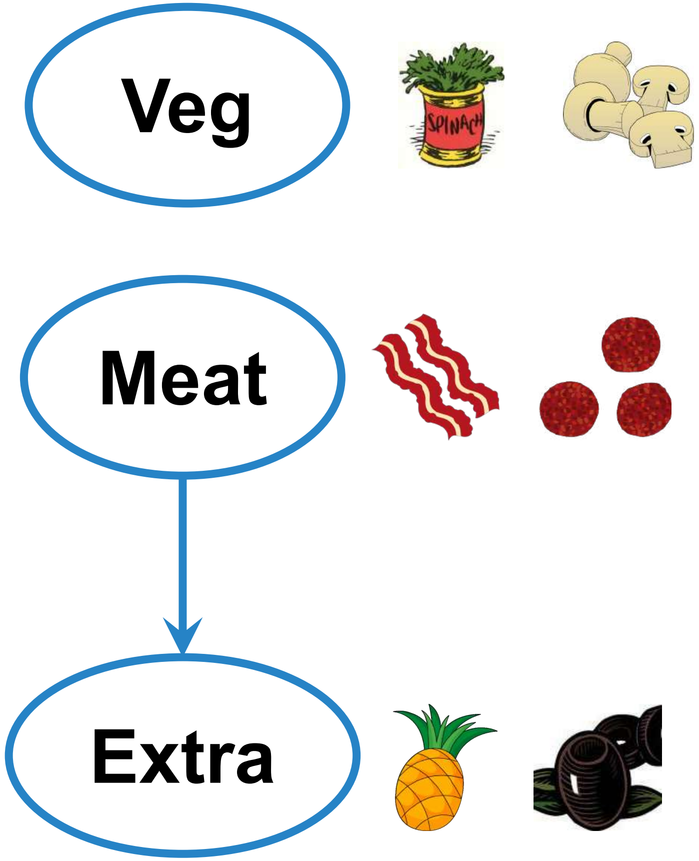
Fish, IceTea, Music

Distance Between Discrete Structures

- Preferences can take many forms: binary, scores, stars, orderings.

- Distances used in recommender systems (similarity of users), classification (distance to classes), and other places.

- **Distance (Metric):**
  - $d(x,y) \geq 0$ (non-negative),
  - $d(x,y) = 0$ iff $x=y$ (identity),
  - $d(x,y) = d(y,x)$ (symmetry), and
  - $d(x,z) \leq d(x,y) + d(y,z)$ (triangle inequality).



Paper Citations

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, Kristen Brent Venable.

On the Distance Between CP-nets. Proc. Aut. Agents and Multiagent Systems (AAMAS) 2018.

Value Alignment via Tractable Preference Distance. Artificial Intelligence Safety and Security, Chapter 18, CRC Press, 2018.

Preferences and Ethical Principles in Decision Making. Proc. ACM/AAAI Conference on AI, Ethics, and Society (AIES), 2018.

CPMetric: Deep Siamese Networks for Learning Distances Between Structured Preferences. arXiv:1809.08350, 2018.

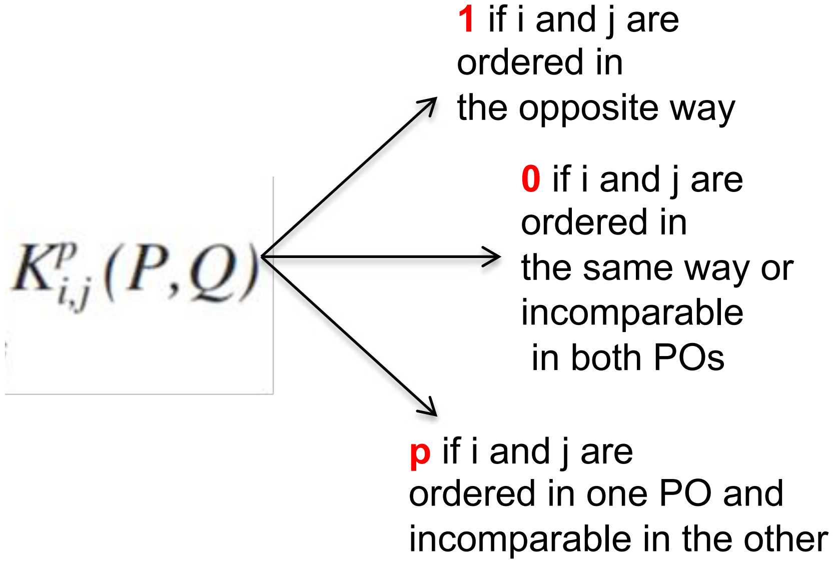
Distance on partial orders

- Measure how similar/different are partial orders:

  → notion of distance over partial orders
- **Kendall's τ with penalty parameter** p (KT)
  - Extends Kendall's τ distance to partial orders
- Given two partial orders P and Q and two outcomes i and j

$$KT(P,Q)=\sum_{i,j,i\neq j} K^{p}_{i,j}(P,Q)$$

where



$$K^{p}_{i,j}(P,Q)$$

1 if i and j are ordered in the opposite way

0 if i and j are ordered in the same way or incomparable in both POs

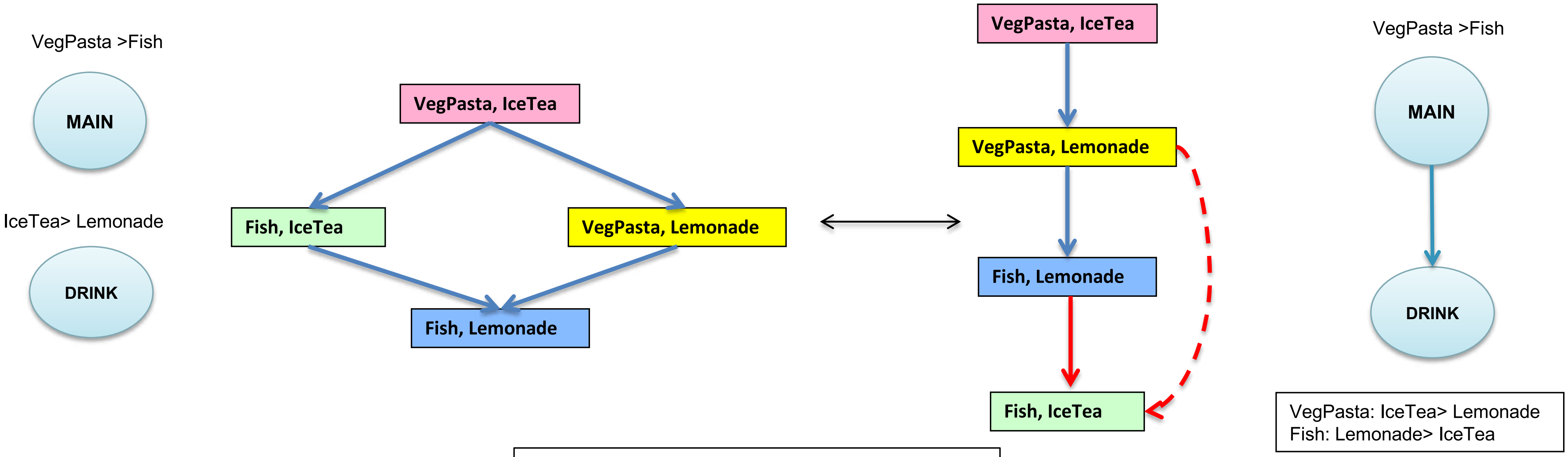
p if i and j are ordered in one PO and incomparable in the other

Paper Citations

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. *Comparing partial rankings.* SIAM J. Discret. Math., 20(3):628–648, March 2006.

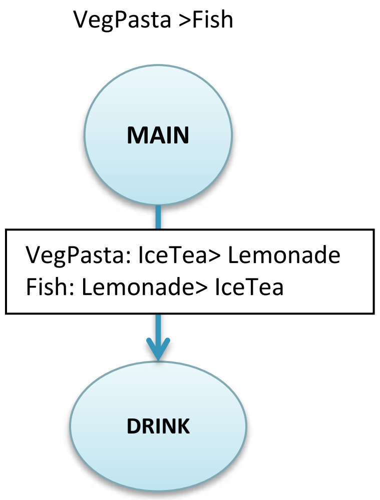
Distance between Structures?

- Given two CP-nets defined over the same set of features, how similar/different are the preferences they represent?



KT between the two induced partial orders =1+0.5=1.5

Examples

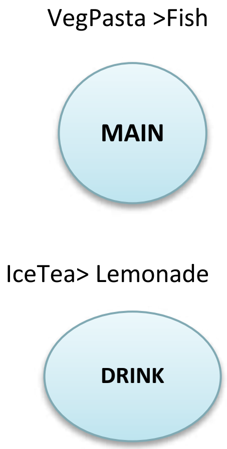

VegPasta >Fish

MAIN

VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

Can we compute the KTD distance directly from the CP-nets in polynomial time?


VegPasta >Fish

MAIN

IceTea> Lemonade

DRINK

Our setting

- The CP-nets we consider:
  - All the **same** set of **binary features**
  - **Acyclic**
  - **O-legal**: there is an ordering O of the features such that if there is an edge X->Y in the CP-net, then X comes before Y in O

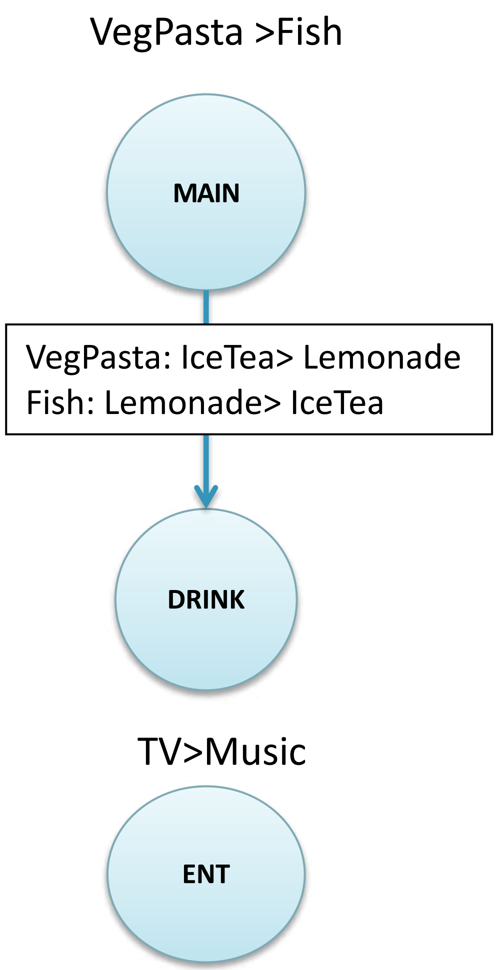
Approximating the KTD distance

- Instead of computing the **KTD between two** CP-nets in polynomial time,

- Compute the **KT of two particular linearization of the POs** from the CP-nets in polynomial time
  - That is, without explicitly computing the linearizations!

Example

CP-net



VegPasta >Fish

MAIN

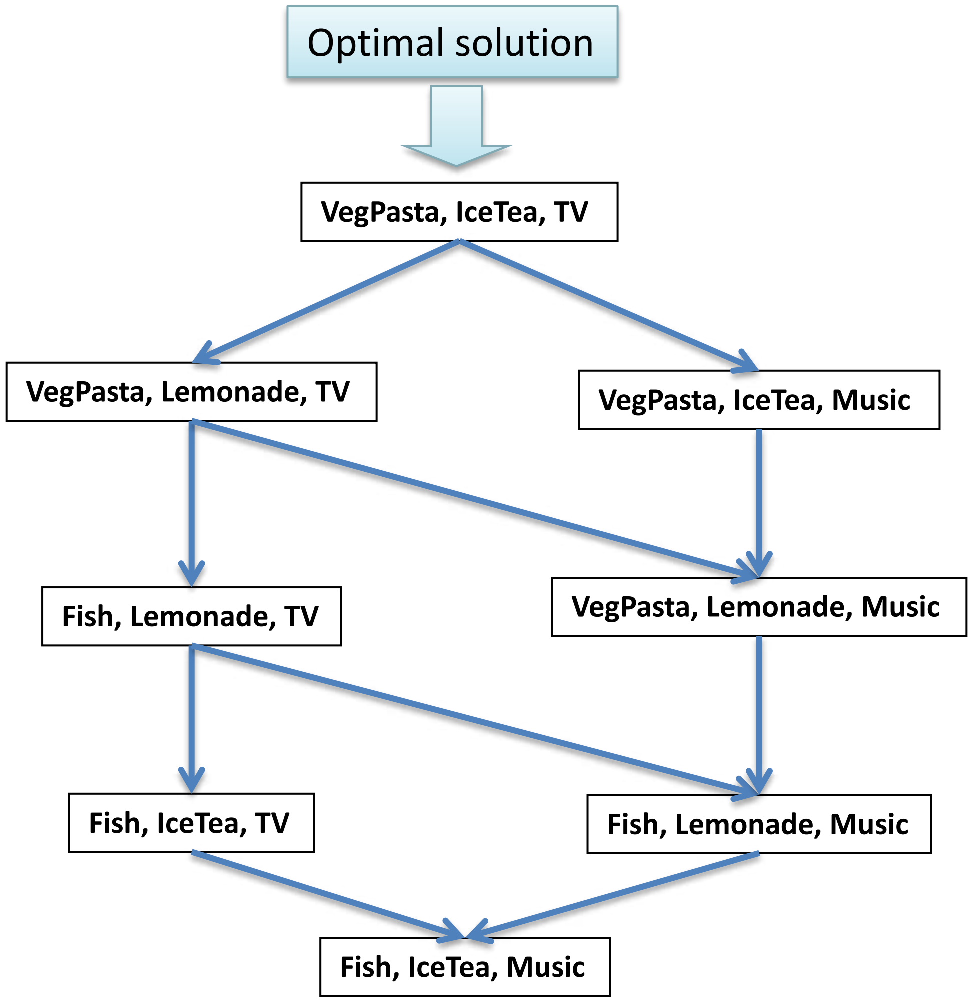
VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

TV>Music

ENT

Induced Ordering



Optimal solution

VegPasta, IceTea, TV

VegPasta, Lemonade, TV

VegPasta, IceTea, Music

Fish, Lemonade, TV

VegPasta, Lemonade, Music

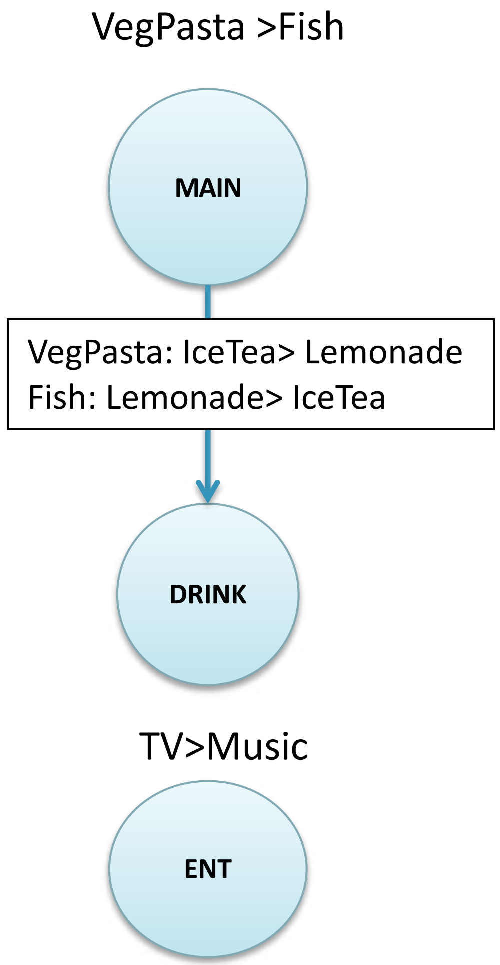
Fish, IceTea, TV

Fish, Lemonade, Music

Fish, IceTea, Music

Example

CP-net



VegPasta >Fish

MAIN

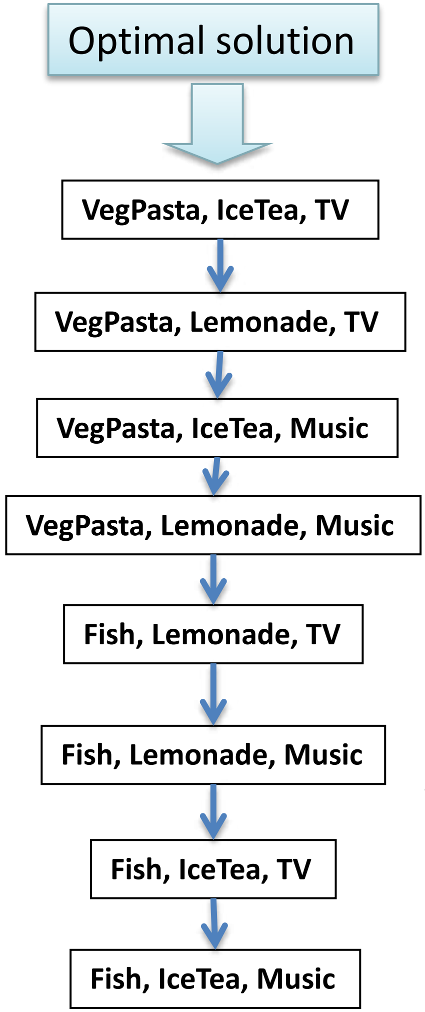
VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

TV>Music

ENT

Linearization



Optimal solution

VegPasta, IceTea, TV

VegPasta, Lemonade, TV

VegPasta, IceTea, Music

VegPasta, Lemonade, Music

Fish, Lemonade, TV

Fish, Lemonade, Music

Fish, IceTea, TV

Fish, IceTea, Music

There are linearizations such that finding the Next best solution directly from the CP-net is easy (polynomial delay)

[Boutilier et al., 2004; Brafman et al., 2009]

CPD distance

- Given two O-legal CP-nets A and B we denote with **LexO(A)** and **LexO(B)** the linearizations of their induced partial orders
  - as defined in Boutilier et al. 2004.
- We define:

CPD(A,B)=KT(LexO(A),LexO(B))

It is easy to see that CPD is a distance

CPD: finding approximation

- Measuring the distance between CP-nets is exponential in the worst case.

- TH: **Given two O-legal CP-nets A and B, with *m* features, CPD(A,B) can be computed in polynomial time** as follows:

1. **Normalize** A and B so that all features have as parents the union of their parents in A and B (redundant rows are added to the CP-tables)
2. Compute the following:

var(j) is the feature such that j is a row in its CP-table

$$\sum_{j\in diff(A,B)} 2^{flw(var(j))+(m-1)-|Pa_B(var(j))|}$$

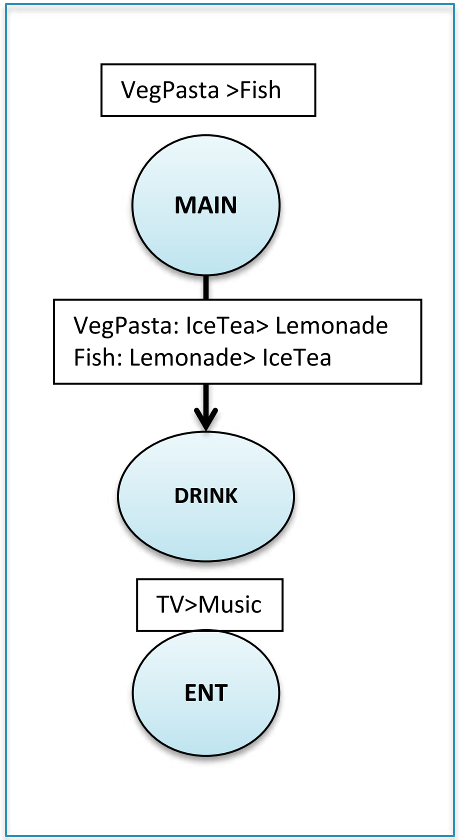
flw(var(j)) are the features that follow var(j) in order O

The number of parents of var(j)

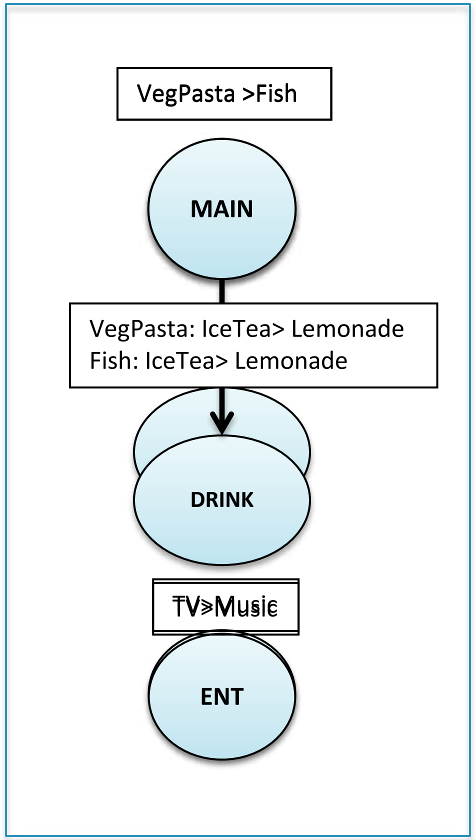
Set of CP-table rows in which A and B differ

Counts the number of pairs of outcomes that are inverted due to the a difference in a CP-table

Computing CPD: Step 1 Normalization

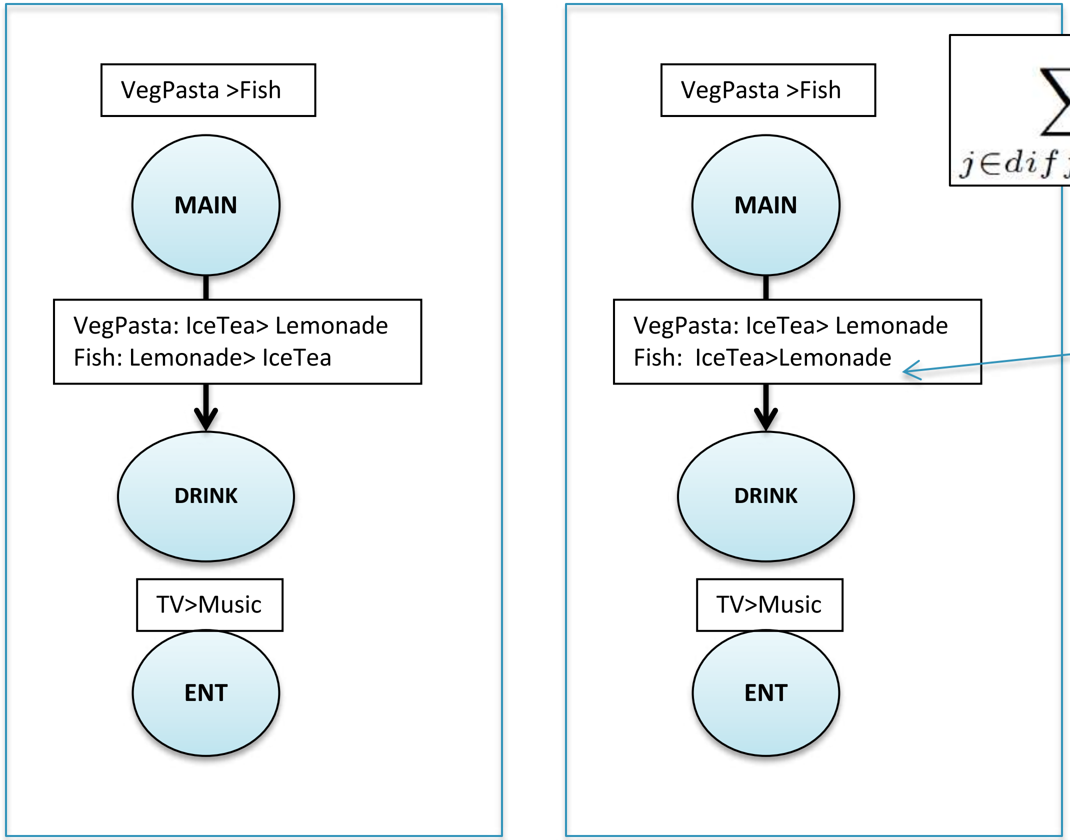


CP-net A



CP-net B

Step 2: Count



CP-net A

CP-net B

$$\sum_{j \in diff(A,B)} 2^{flw(var(j))+(m-1)-|Pa_B(var(j))|}$$

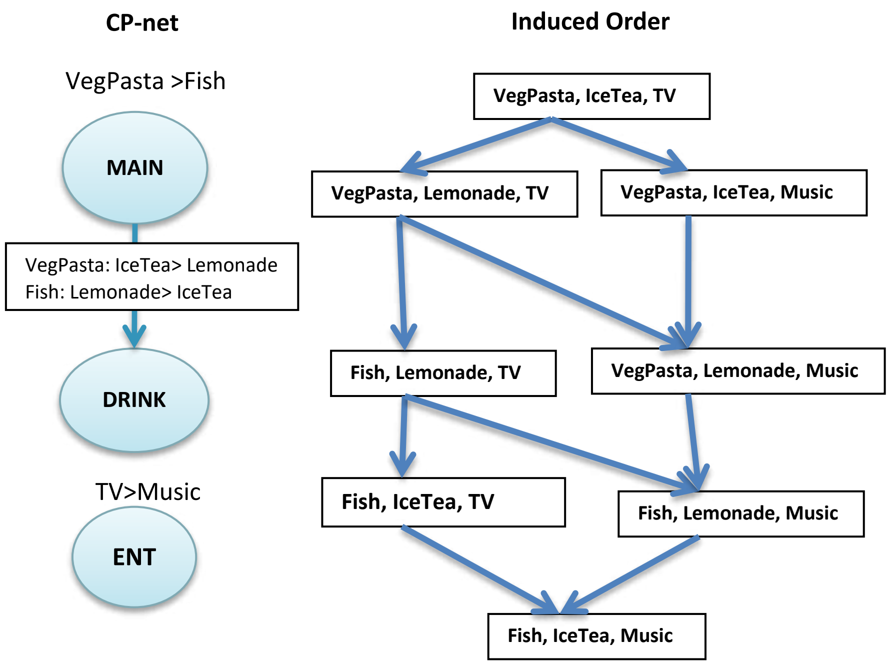
diff(A,B)

var(j)=DRINK

flw(DRINK)=1 (only ENT)
m=3
|PA(DRINK)|=1, DRINK has only MAIN as parent

$2^{1+3-1-1}=2^2=4$

Examples



CP-net

VegPasta >Fish

MAIN

VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

TV>Music

ENT

Induced Order

VegPasta, IceTea, TV

VegPasta, Lemonade, TV

VegPasta, IceTea, Music

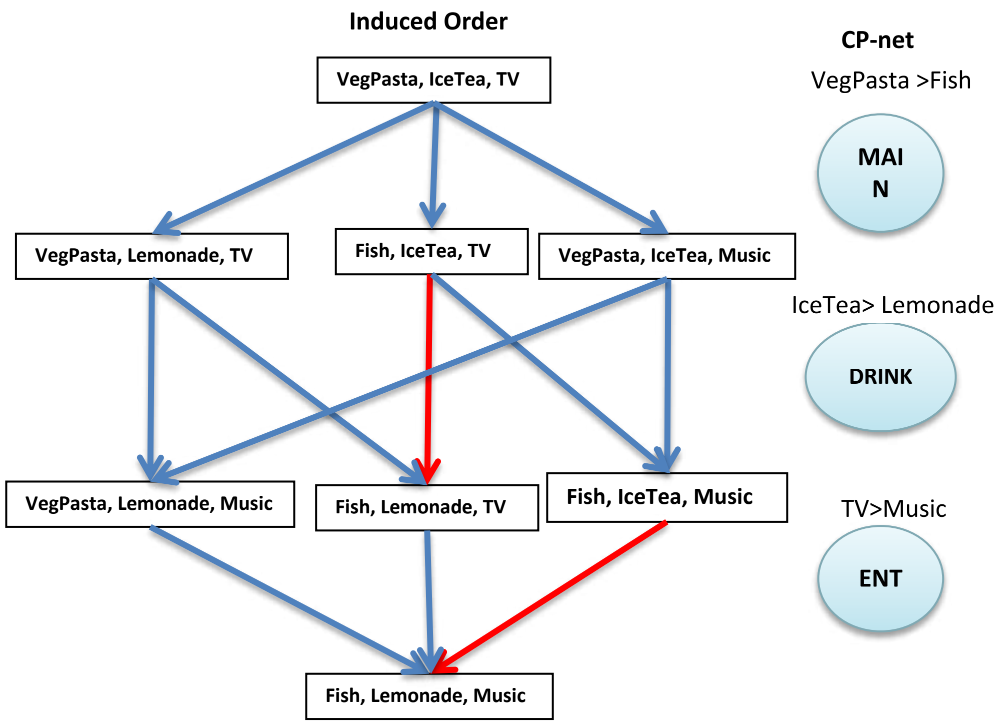
Fish, Lemonade, TV

VegPasta, Lemonade, Music

Fish, IceTea, TV

Fish, Lemonade, Music

Fish, IceTea, Music



Induced Order

VegPasta, IceTea, TV

VegPasta, Lemonade, TV

Fish, IceTea, TV

VegPasta, IceTea, Music

VegPasta, Lemonade, Music

Fish, Lemonade, TV

Fish, IceTea, Music

Fish, Lemonade, Music

CP-net

VegPasta >Fish

MAI
N

IceTea> Lemonade

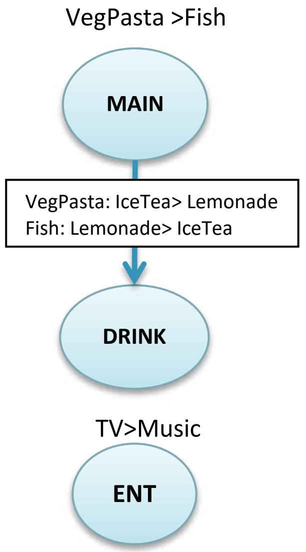
DRINK

TV>Music

ENT

Examples

CP-net



VegPasta >Fish

MAIN

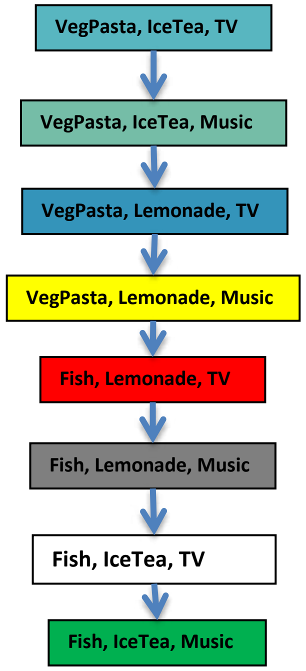
VegPasta: IceTea> Lemonade
Fish: Lemonade> IceTea

DRINK

TV>Music

ENT

Linearization



VegPasta, IceTea, TV

VegPasta, IceTea, Music

VegPasta, Lemonade, TV

VegPasta, Lemonade, Music

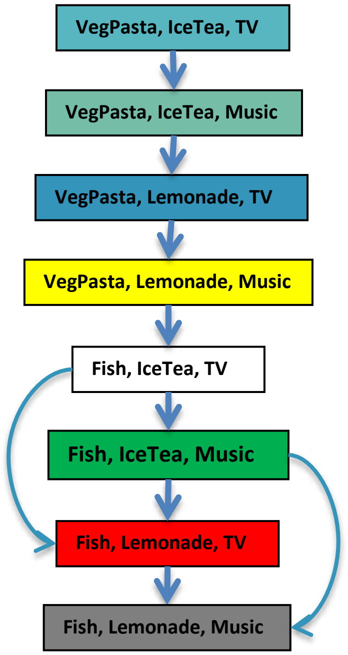
Fish, Lemonade, TV

Fish, Lemonade, Music

Fish, IceTea, TV

Fish, IceTea, Music

Linearization



VegPasta, IceTea, TV

VegPasta, IceTea, Music

VegPasta, Lemonade, TV

VegPasta, Lemonade, Music

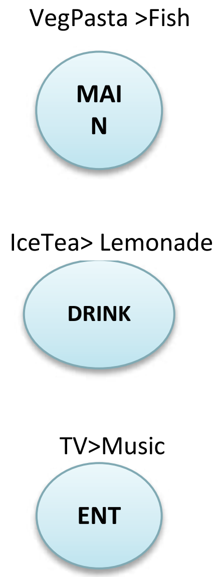
Fish, IceTea, TV

Fish, IceTea, Music

Fish, Lemonade, TV

Fish, Lemonade, Music

CP-net



VegPasta >Fish

MAI
N

IceTea> Lemonade

DRINK

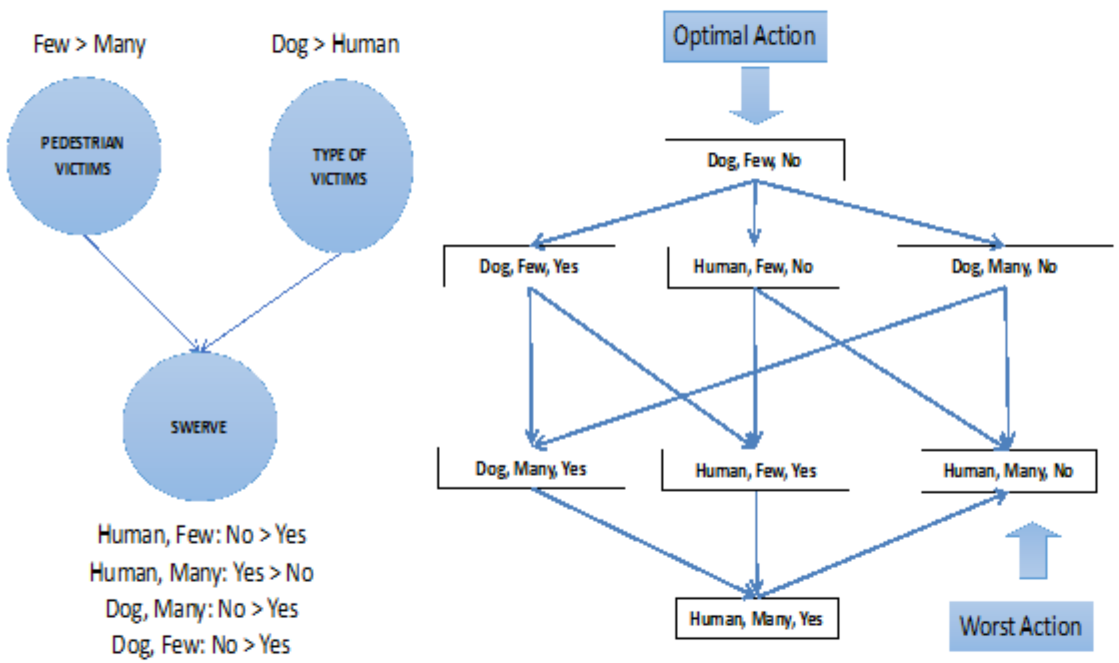
TV>Music

ENT

CP-nets as Ethical Priorities

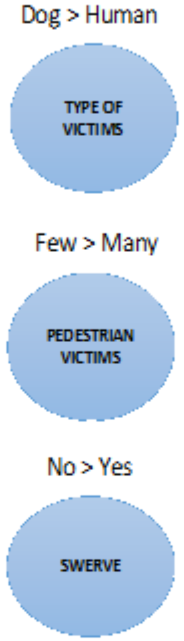
- **Moral Preferences:** Amartya Sen, “morality requires judgment among preferences.”
  - Meta-ranking: preferences over preferences.
  - The preferences of an individual can be morally evaluated by measuring the distance of his/her CP-net from the moral one.

Collective Ethics

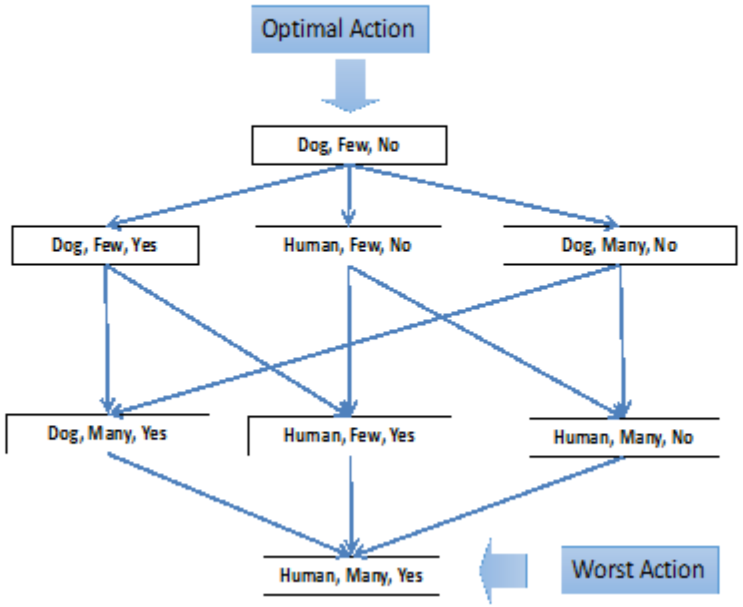


Human, Few: No > Yes
Human, Many: Yes > No
Dog, Many: No > Yes
Dog, Few: No > Yes

Angel Driver Ethics

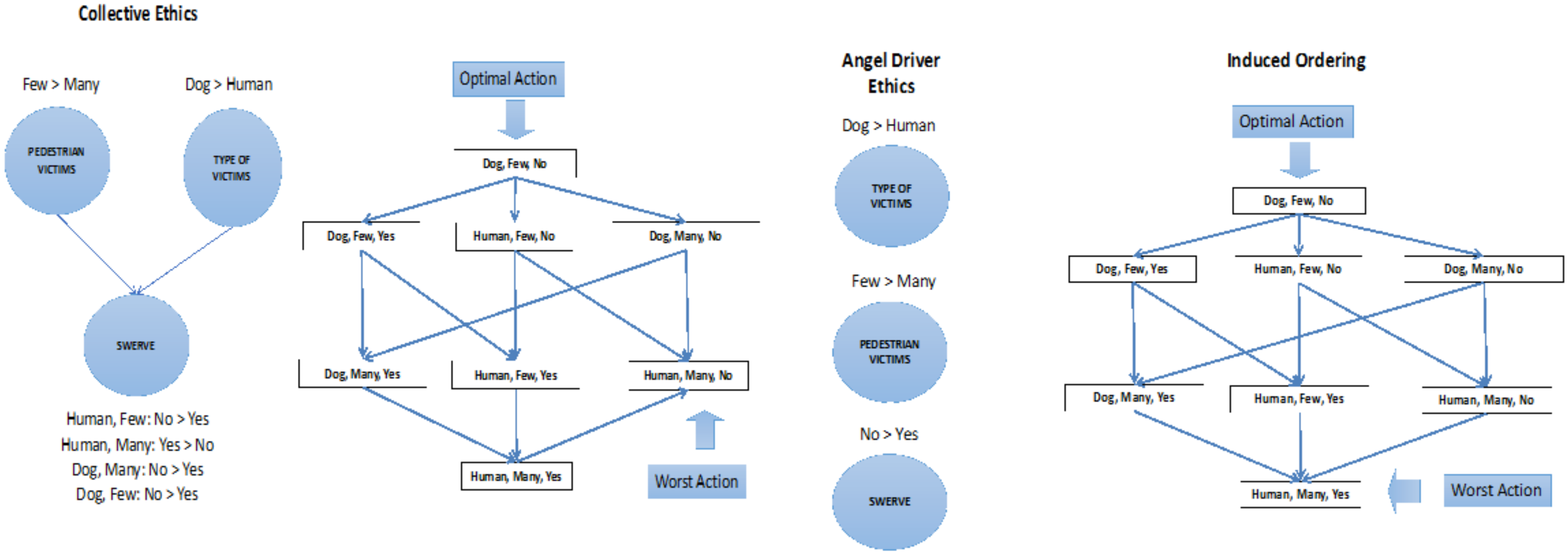


Induced Ordering



CP-nets as Ethical Priorities

- **Value Alignment Procedure.** Given an ethical principle and the preference of an individual:
  - Understand if following preferences will lead to an ethical action.
  - If not, find action which is closer to the ethical principle and near the preference.



Collective Ethics

Few > Many

Dog > Human

PEDESTRIAN VICTIMS

TYPE OF VICTIMS

SWERVE

Human, Few: No > Yes

Human, Many: Yes > No

Dog, Many: No > Yes

Dog, Few: No > Yes

Optimal Action

Dog, Few, No

Dog, Few, Yes

Human, Few, No

Dog, Many, No

Dog, Many, Yes

Human, Few, Yes

Human, Many, No

Human, Many, Yes

Worst Action

Angel Driver Ethics

Dog > Human

TYPE OF VICTIMS

Few > Many

PEDESTRIAN VICTIMS

No > Yes

SWERVE

Induced Ordering

Optimal Action

Dog, Few, No

Dog, Few, Yes

Human, Few, No

Dog, Many, No

Dog, Many, Yes

Human, Few, Yes

Human, Many, No

Human, Many, Yes

Worst Action

Value Alignment Procedure

- Given an ethical principles and individual’s preferences.

A. Set two distance thresholds: t1 (ranging between 0 and 1) between CP-nets, and t2 between decisions (ranging between 1 and n)

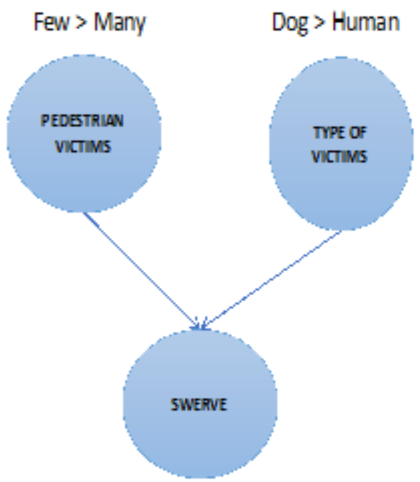
B. Check if the two CP-nets A and B are less distant than **t1**. In this step, we use **CPD** to compute the distance

C. If so, individual is allowed to choose the top outcome of his preference CP-net

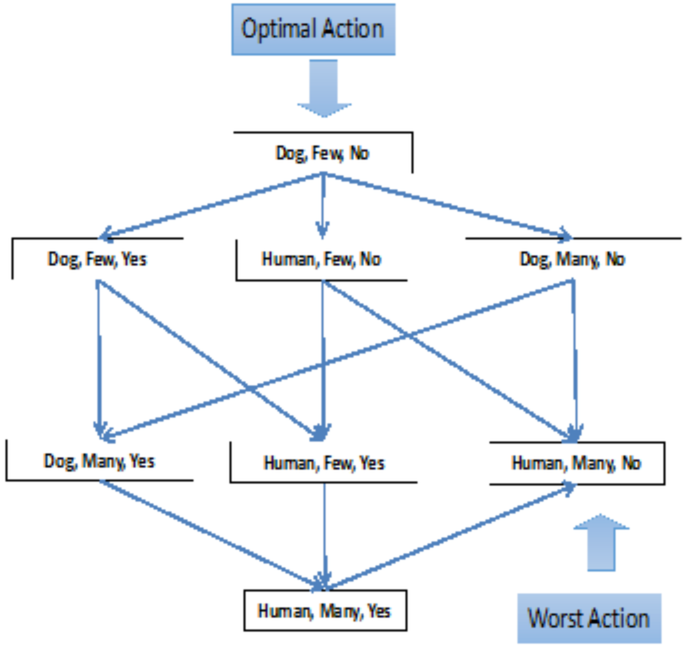
D. If not, then individual needs to move down its preference ordering to less preferred decisions, until he finds one that is closer than t2 to the optimal ethical decision.

Value Alignment Procedure

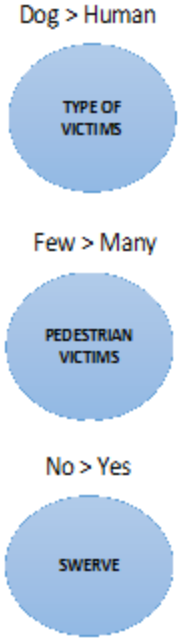
Collective Ethics



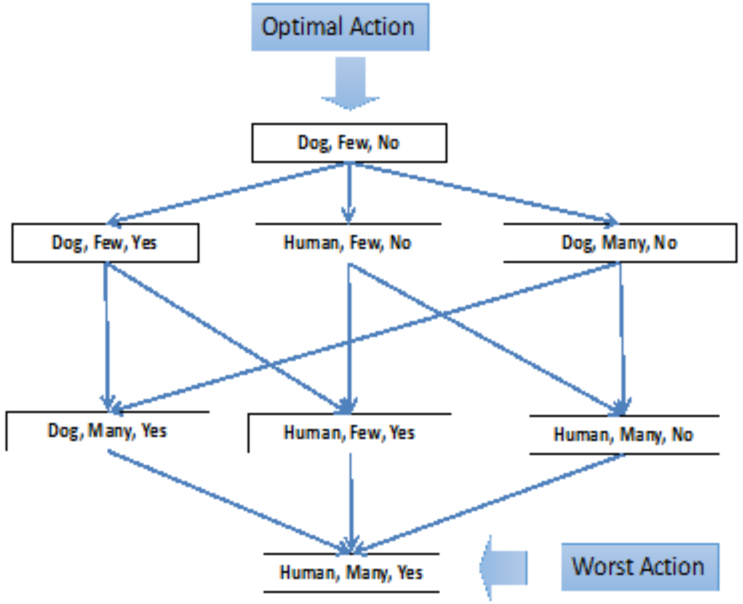
Human, Few: No > Yes
Human, Many: Yes > No
Dog, Many: No > Yes
Dog, Few: No > Yes



Angel Driver Ethics

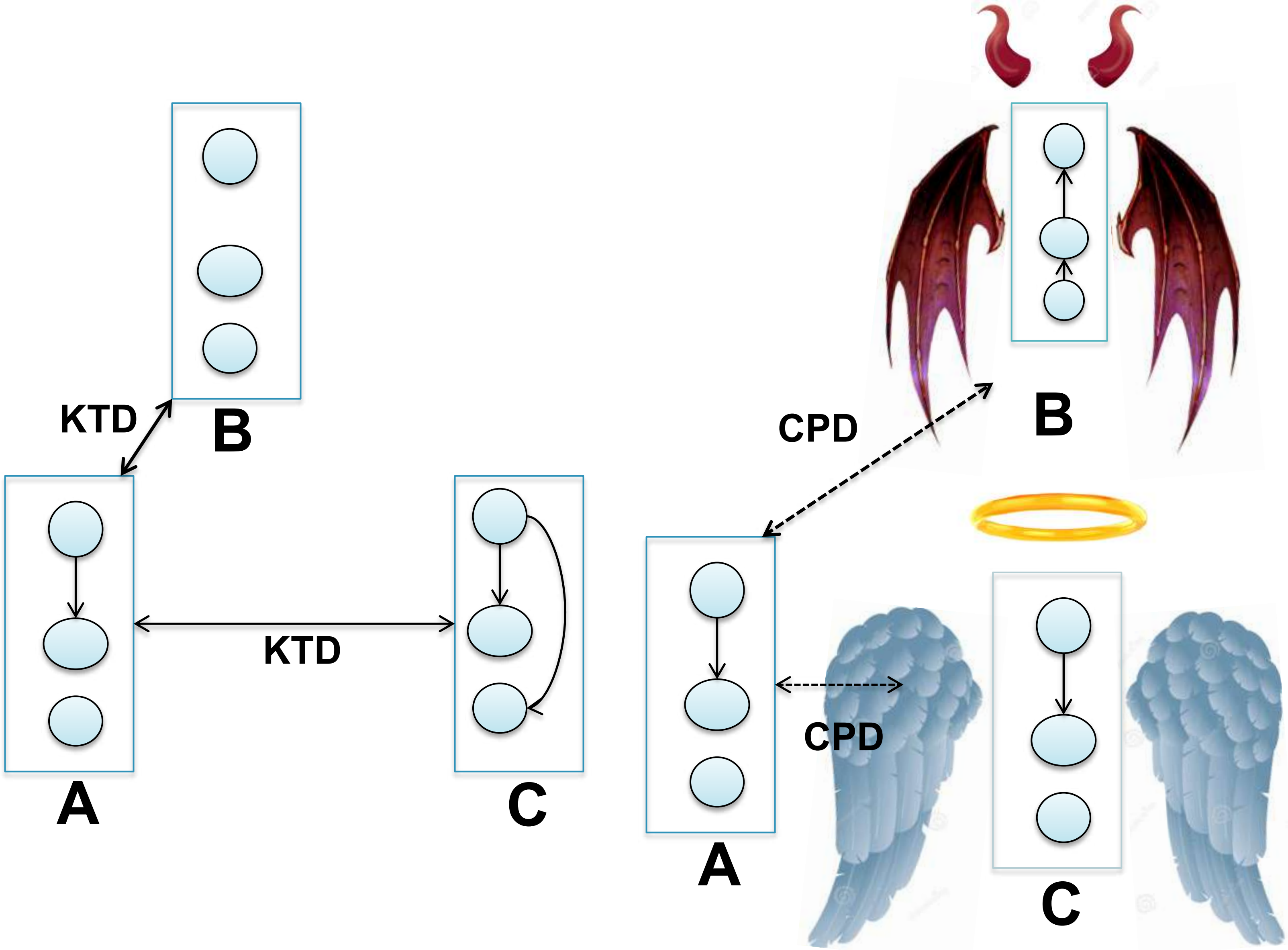


Induced Ordering



Value Alignment

- We generate triplets of CP-nets (A,B,C).
- We chose one as pivot: A.
- We count how many time KTD says B is closer and the other distances say C is closer.
- CPD
- Gives us a notion of a “more compliant” CP-net.



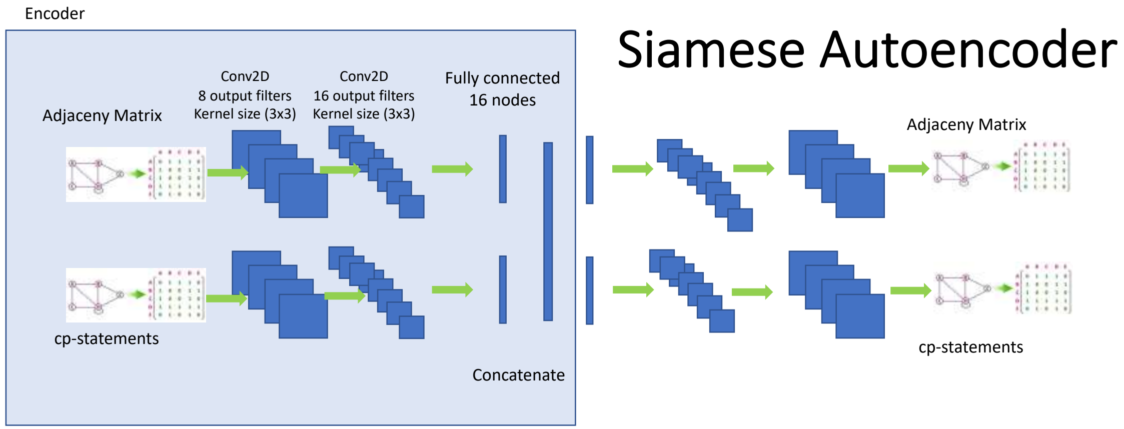
Measuring the Distance

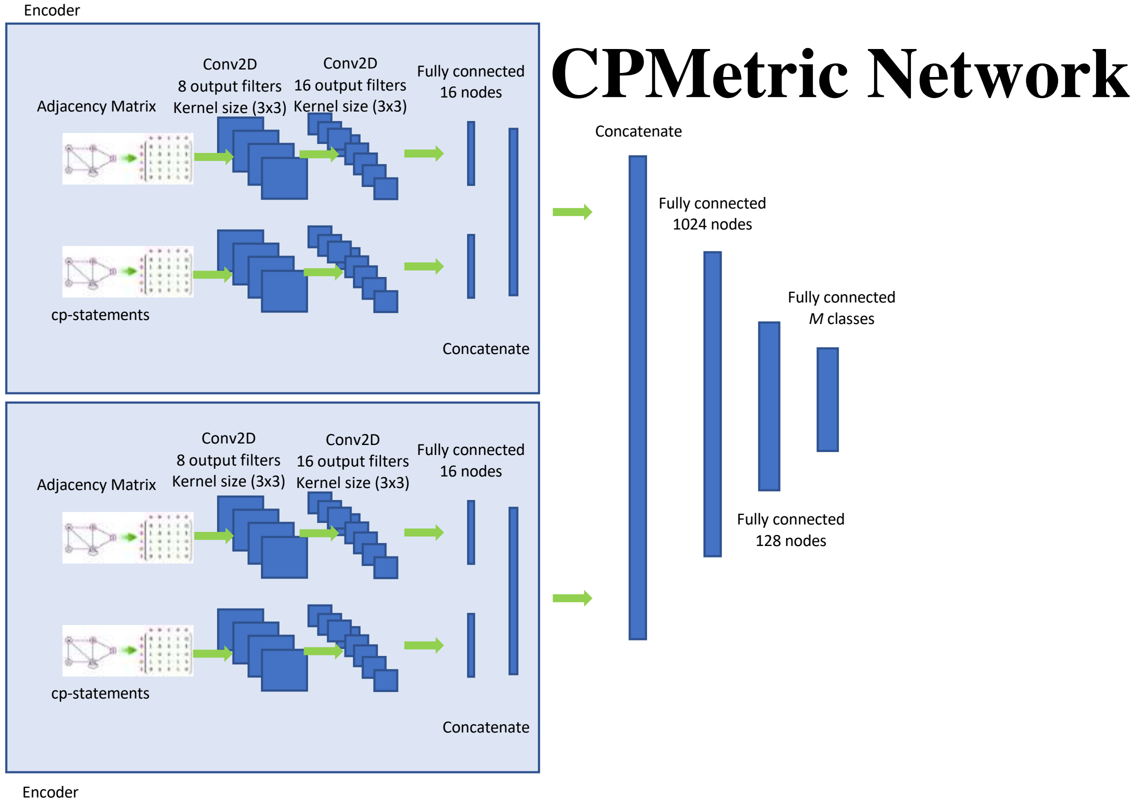
- Measuring the distance between CP-nets is exponential in the worst case.

- Need to find a way to evaluate the distance between, e.g., two competing CP-nets and a third “Moral” CP-net. Judge which one is “more aligned.”

- Using machine learning we have two steps:
  - Encode the CP-net (graph embedding issues).
  - Determine the distance.

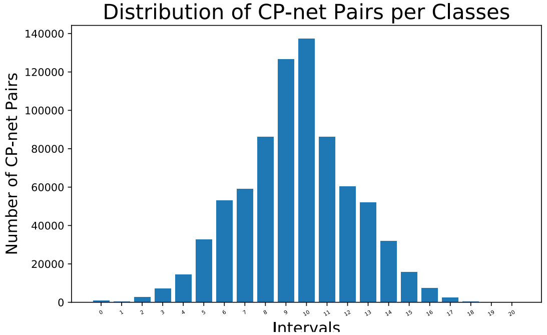
- We encode the normalized laplacian matrix of the graph and a table of the cp-statements.

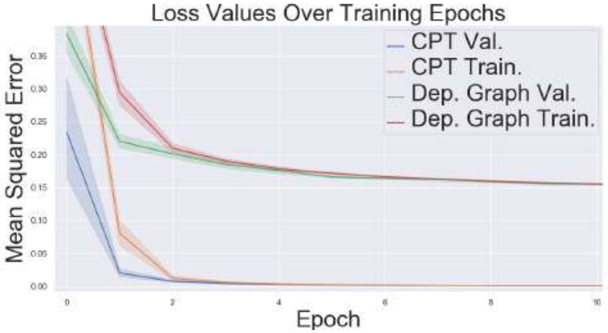


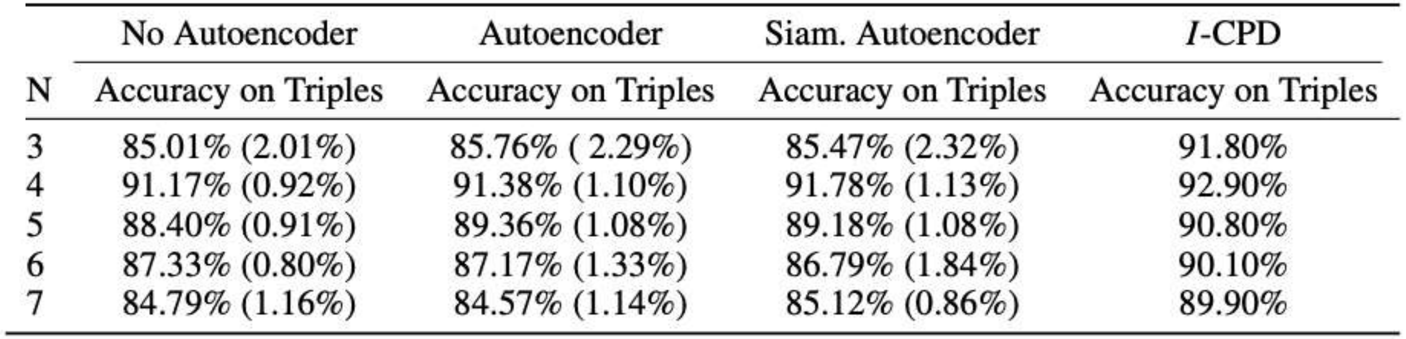


Experiments and Results

- For training we generate 1000 randomly generated CP-nets and compute the distance for all pairs for all *n = {3, …, 7}*. For testing we generate another 1000 randomly generated CP-nest and find all possible triples.
- We get good convergence in the training phase and are able to learn a high quality latent representation.
- For the comparison task we are slightly outperformed by an approximation method, though we run two orders of magnitude faster.







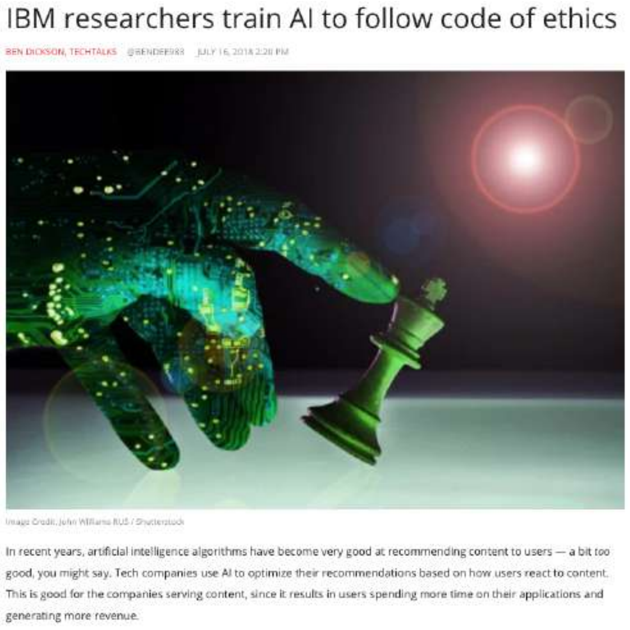
	No Autoencoder	Autoencoder	Siam. Autoencoder	*I*-CPD
N	Accuracy on Triples	Accuracy on Triples	Accuracy on Triples	Accuracy on Triples
3	85.01% (2.01%)	85.76% (2.29%)	85.47% (2.32%)	91.80%
4	91.17% (0.92%)	91.38% (1.10%)	91.78% (1.13%)	92.90%
5	88.40% (0.91%)	89.36% (1.08%)	89.18% (1.08%)	90.80%
6	87.33% (0.80%)	87.17% (1.33%)	86.79% (1.84%)	90.10%
7	84.79% (1.16%)	84.57% (1.14%)	85.12% (0.86%)	89.90%

Table 1: Performance of the various network architectures on the qualitative comparison task as well as performance of *I*-CPD. While our networks do not achieve the best performance on this task they are competitive with the more costly approximation algorithm *I*-CPD.

Conclusions and Next Steps

- ***We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.***

- <u>Important Questions and Next Steps:</u>
  - How do we measure distance between heterogenous structures?
  - How do we capture and encode norms/values/expectations?
  - How do we account for edge effects?
  - How do we transition our techniques to other preference representations / formalisms?

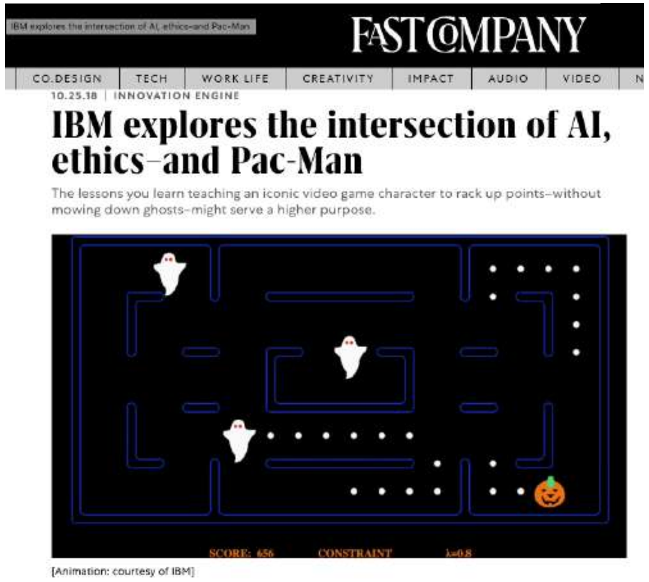


IBM researchers train AI to follow code of ethics

BEN DICKSON, TECHTALKS @BENDEE983 JULY 16, 2018 2:20 PM

Image Credit: John Williams RUS / Shutterstock

In recent years, artificial intelligence algorithms have become very good at recommending content to users — a bit too good, you might say. Tech companies use AI to optimize their recommendations based on how users react to content. This is good for the companies serving content, since it results in users spending more time on their applications and generating more revenue.



IBM explores the intersection of AI, ethics–and Pac-Man

FAST COMPANY

CO.DESIGN | TECH | WORK LIFE | CREATIVITY | IMPACT | AUDIO | VIDEO | N

10.25.18 | INNOVATION ENGINE

IBM explores the intersection of AI, ethics–and Pac-Man

The lessons you learn teaching an iconic video game character to rack up points–without mowing down ghosts–might serve a higher purpose.

[Animation: courtesy of IBM]

When Is It Morally Acceptable to Break the Rules? A Preference-Based Approach

Edmond Awad, Sydney Levine, **Andrea Loreggia**, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum and Max Kleiman-Weiner

Motivations

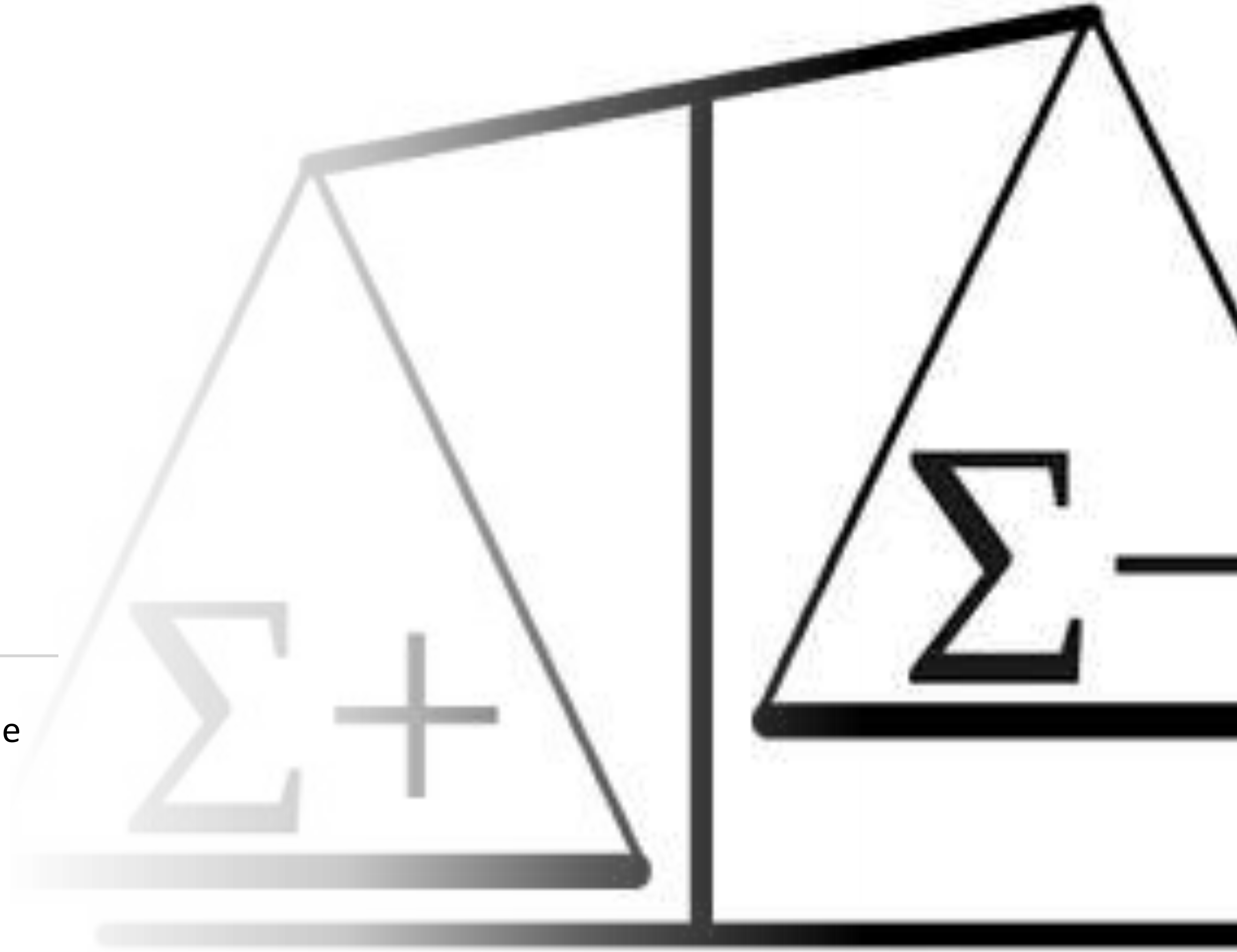
- Investigate when humans find acceptable to break the rules
- Providing some glimpse of our moral judgement methodology
- Investigate when humans switch between different frameworks for moral decisions and judgments
- Model and possibly embed this switching into a machine

Deontology

Following common rules that have been agreed upon by us or society

Utilitarianism

Evaluating the consequences of the possible actions before deciding



Contractualism

Finding an agreement between the parties involved

In line scenarios

Is it always true?

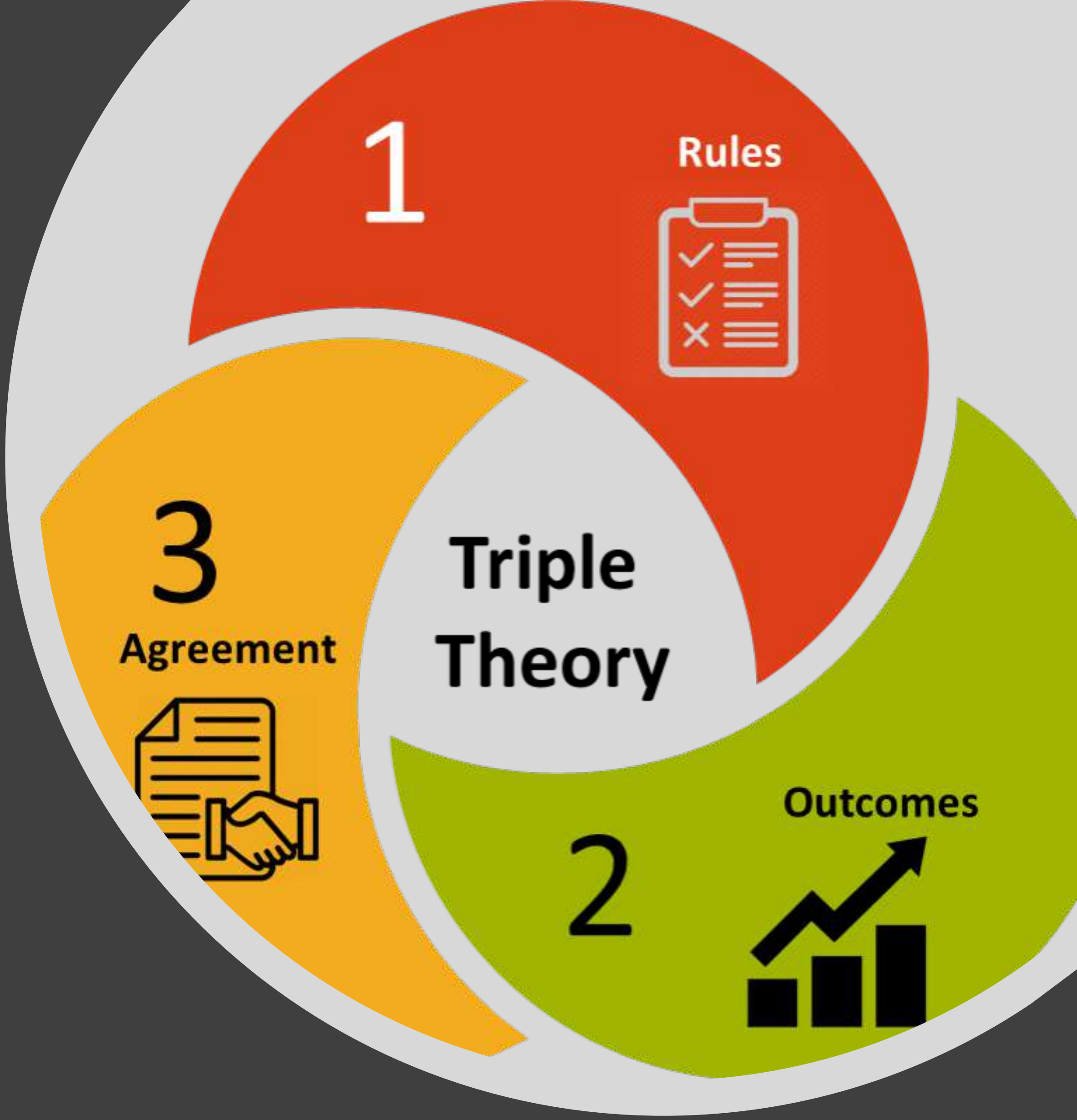
Counter-examples

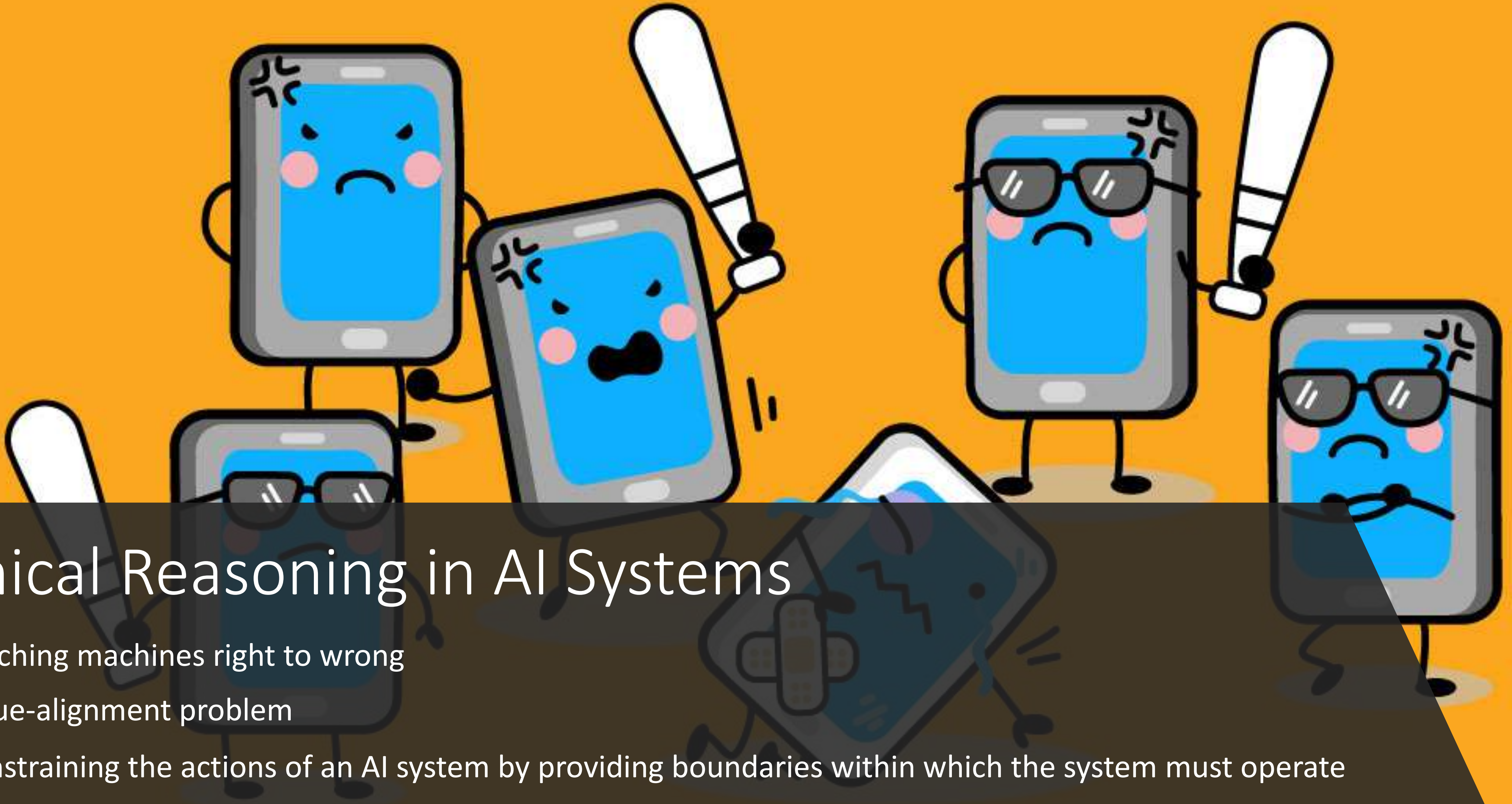
Under certain conditions, we are allowed to cut to the front of the line without waiting

Triple Theory

A unified theory of moral cognition to:

- Combine elements of each of the theories of moral philosophy
- Build a computational model to direct actions of an AI system.





Ethical Reasoning in AI Systems

- Teaching machines right to wrong
- Value-alignment problem
- Constraining the actions of an AI system by providing boundaries within which the system must operate

Experimental Details

- 27 short vignettes about people waiting in line in three different contexts (deli, bathroom, airport)
- 320 subjects were recruited from Amazon MTURK
- Subjects were randomly assigned to one of two experimental groups (moral judgment or context evaluation)

Experimental Details

Moral judgment group:

- Read all the scenarios (27 total)
- For each scenario answer whether it was acceptable for the protagonist to cut in line (yes/no).

Experimental Details

Context evaluation group:

- Subjects evaluated all the vignettes in one context only (9 questions).

Experimental Details

Example of evaluation:

- Everyone: Think about the well-being of all the people in line combined. How are they affected by the person cutting in line?
- First Person: How much worse off/better off is the first person in line?

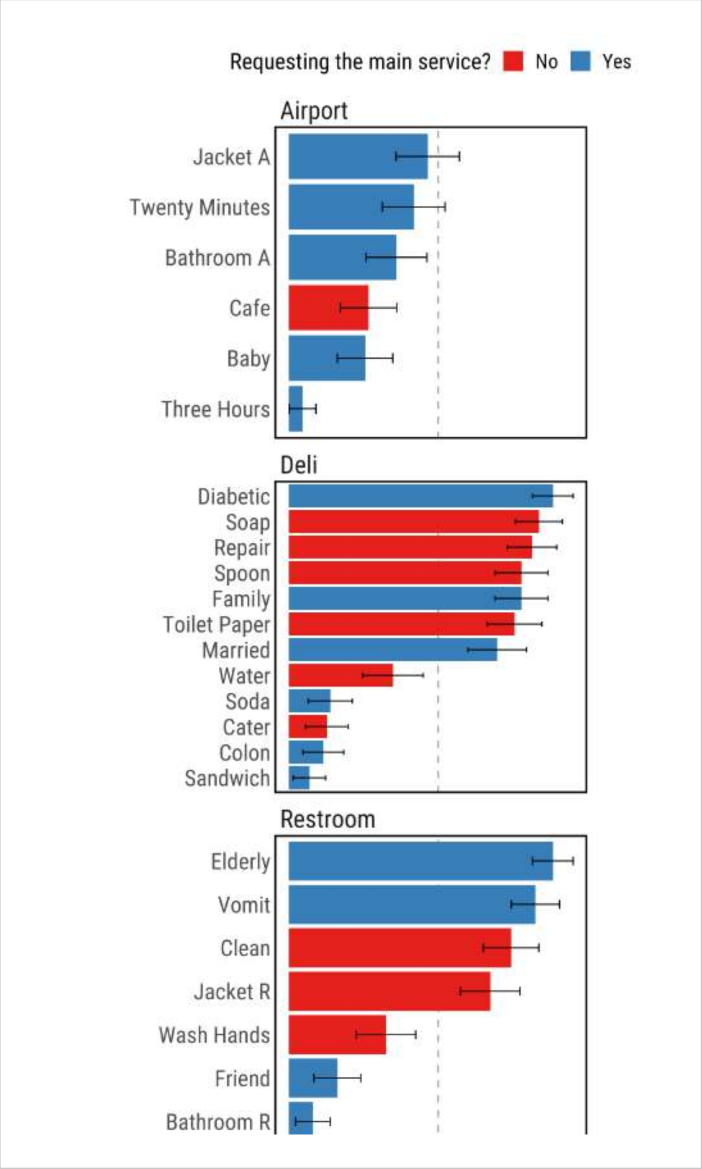
Example

Imagine that there are five people who are waiting in line at a deli to order sandwiches for lunch. There is only one person (the cashier) working at the deli.

A customer who is eating lunch at the deli wants more a refill on tap water.

Is it OK for that person to ask the cashier for more water without waiting in line?

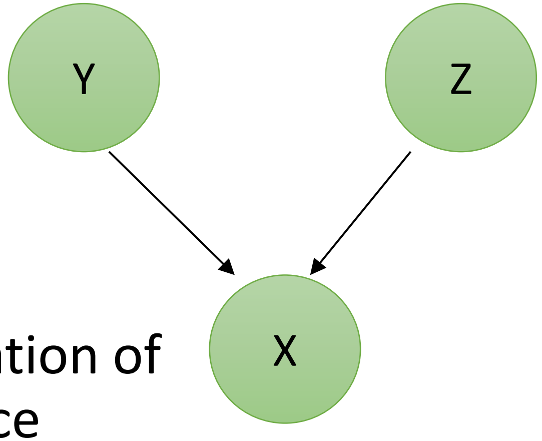
MENTIMENTER



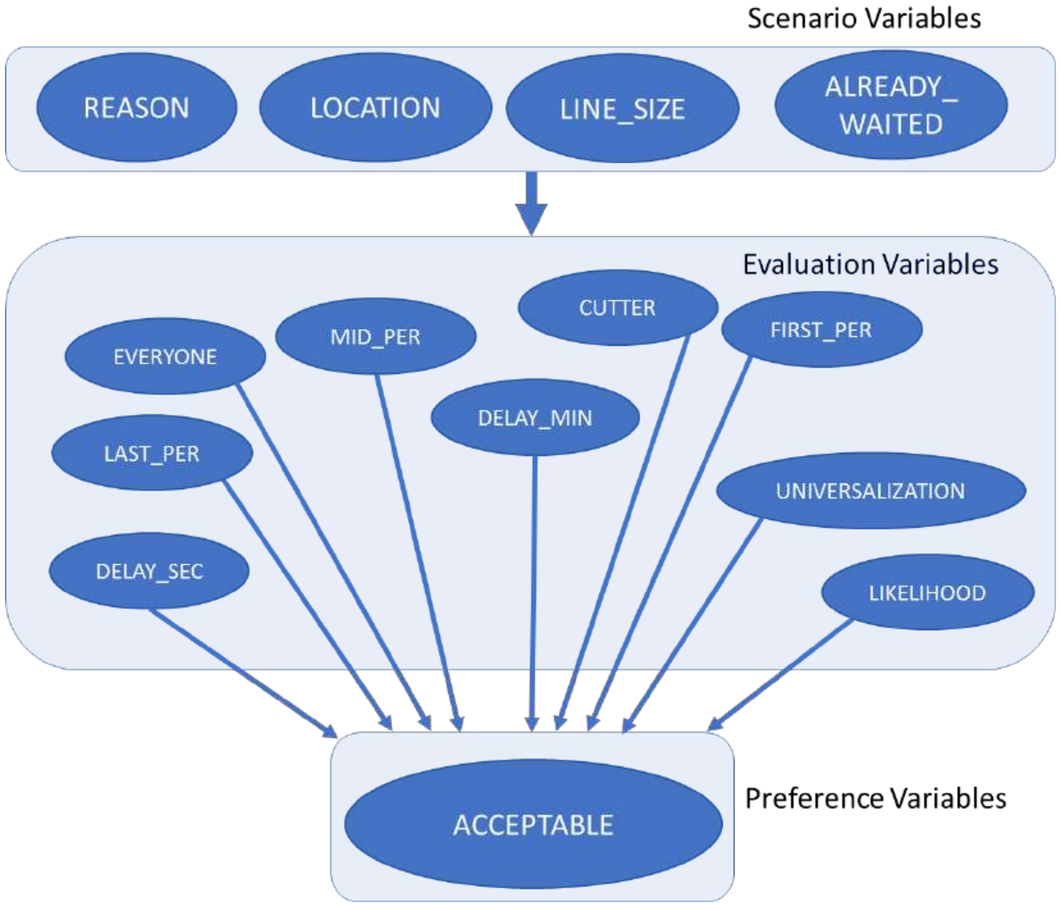
Is it OK to cut the line?

CP-nets

- Variables $\{X_1, \ldots , X_n\}$ with domains
- For each variable, a total order over its values
- **Independent variable:**
  - $X=v_1 > X=v_2 > \ldots > X=v_k$
- **Conditioned variable:** a total order for each combination of values of some other variables (conditional preference table)
  - $Y=a, Z=b: X=v_1 > X=v_2 > \ldots > X=v_k$
  - X depends on Y and Z (parents of X)
- Graphically: **directed graph** over $X_1, \ldots , X_n$
  - Possibly cyclic

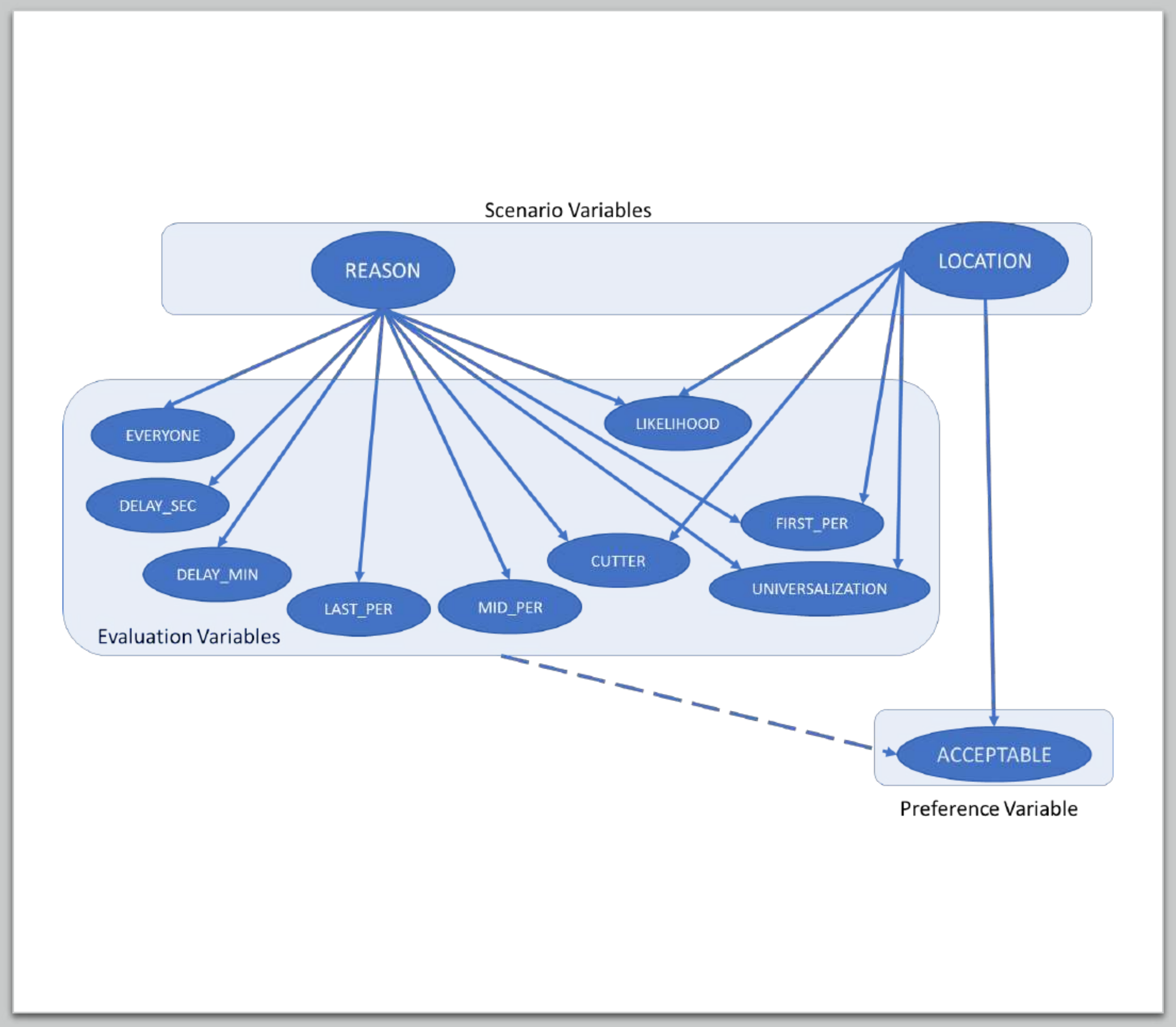


Modelling and Reasoning with Preferences



Data Analysis

- We evaluate whether we can reject the following three null hypotheses (NH):
  - NH1: location does not affect EVs;
  - NH2: reason does not affect EVs;
  - NH3: location does not affect the PV



On-Going and Future Work

- Generalizing CP-nets to Model Moral Preferences
- Prescriptive Plans Based on Moral Preferences

- Understand how, why, and when it is morally acceptable to break rules
- constructed and studied a suite of hypothetical scenarios relating to this question, and collated human moral judgements on these scenarios.
- showed that existing structures in the preference reasoning literature are insufficient for this task.
- We look towards extending this into other established areas of AI research.

MAI4CAREU Master programmes in Artificial Intelligence 4 Careers in Europe

Co-financed by the European Union
Connecting Europe Facility

This Master is run under the context of Action
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom
under GA nr. INEA/CEF/ICT/A2020/2267423