# Lesson 5_Value_Alignment

## Value Alignment

### What is intelligent?

There is **no a universal definition**, because it does not exist a single kind of intelligence. We can think about intelligence 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 (one of the father of AI)*

"AI systems can either use **symbolic rules** (top-down approach) or **learn a numeric model** (bottom-up approach), and they can also adapt their behaviour by analyzing how the environment is affected by their previous actions."

> *HLEG on AI (High-Level Expert Group* https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai*)*

Regardless the kind of definition of AI we want to use, we can split it into 2 categories:

> **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 are 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,** by **interdisciplinary research** and performing **cross-pollination between fields (psychology, CS, maths ecc).**

**Inizio descrizione del paper**

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

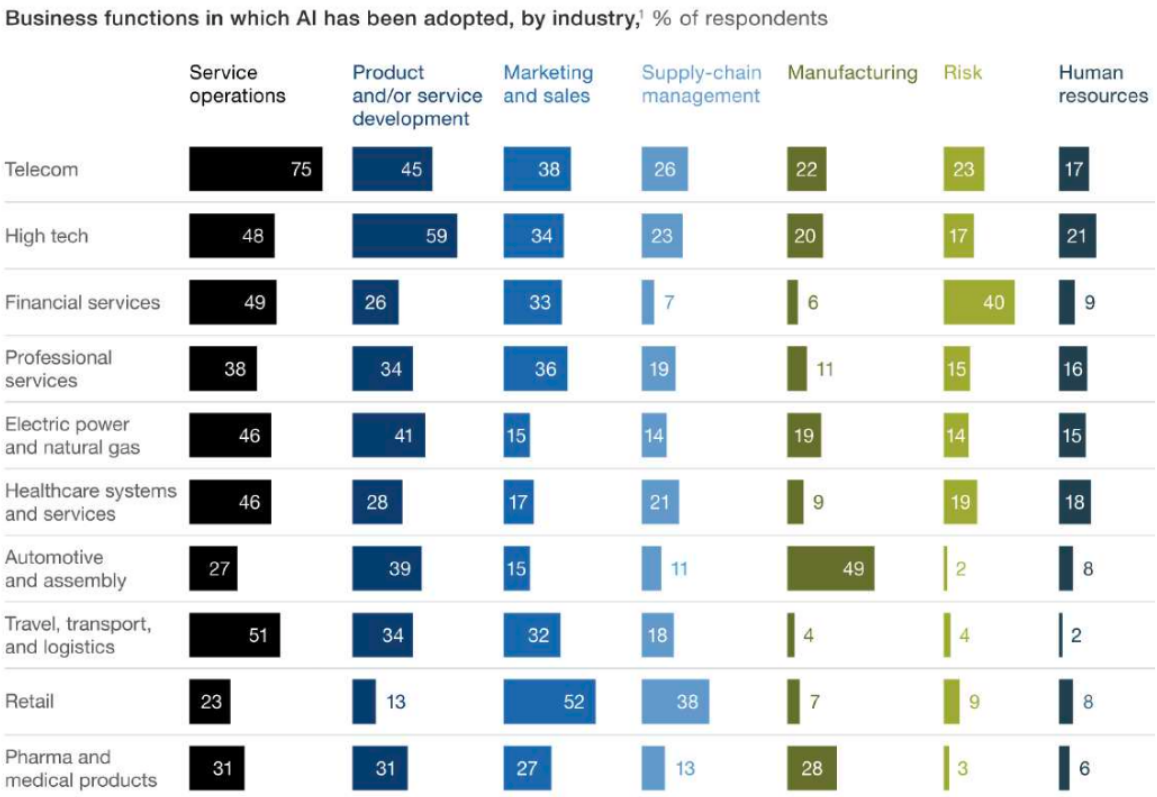
The paper underlines the **necessity of an interdisciplinary research, a cross-fertilization process.**

The paper identifies some **short-term research priorities:**

- **Optimizing AI's economic impact**
  - Labor Market Forecasting (undertand which is the impact of AI in the market (foto sotto))
  - Other Market Disruptions (we need to educate people in change their goal in the market, since some occupations will not be present in the future, do to AI)
  - Policy for managing Adverse Effects
- **Law and ethics research**
  - Liability and Law for AVs (i.e. autonomous vehicles)
  - Machine Ethics
  - Autonomous Weapons• Privacy
  - Professional Ethics
  - Policy Questions
- **Computer science research for robust AI**
  - Verification
  - Validity
  - Security
  - Control

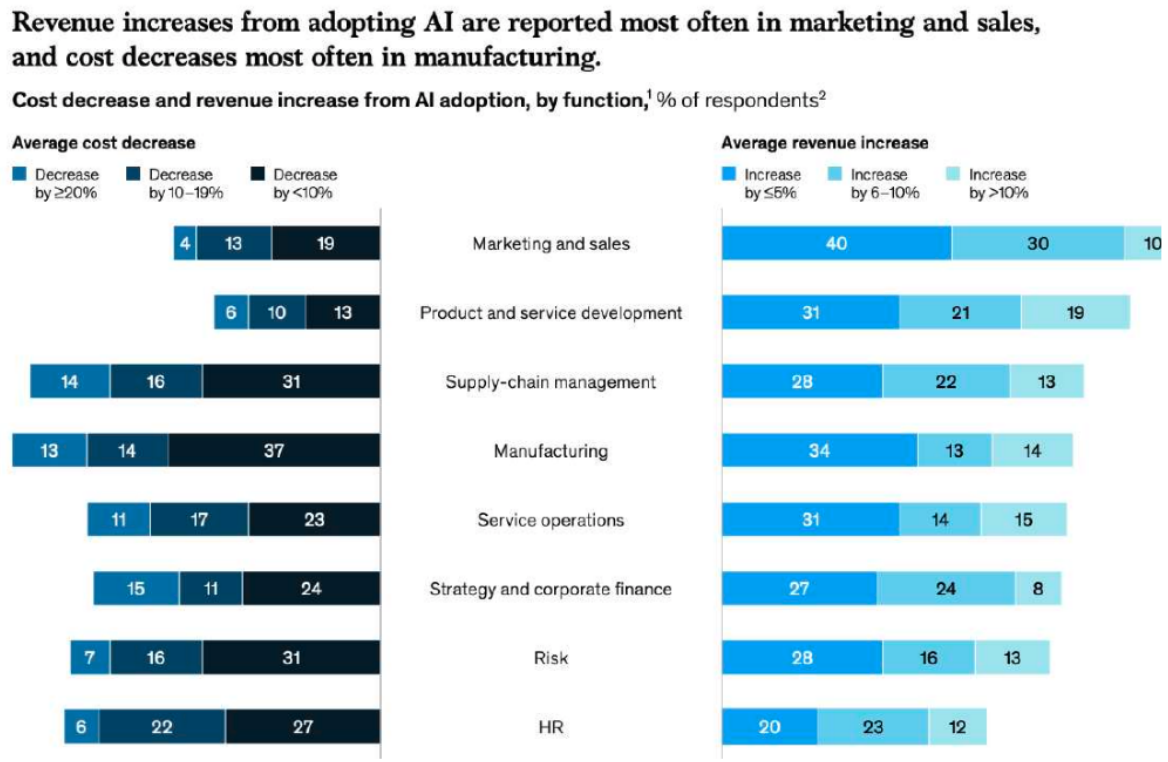
**AI in Business Functions**

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




**Ai benefits**

Source: *"Global AI Survey: AI proves its worth, but few scale impact". Mckinsey, 2019*




La qualità di queste immagini fa schifo, ma era così anche nelle slides...sorry :(

---

The paper identifies some long-**term research priorities:**

- Verification
- Security
- Control

**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.

“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” (from the conclusion of the paper)

---

**Fine descrizione del paper**

---

Now the question is: **how can we represent values, norms and principles in order to use them to solve the value-alignment problem?**

## What are values, norms, and principles?

*Paper: 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 are quite complex to define, but let's focus on a more practical perspective:

v**alues 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)

On the other hand, **norms, duties, principles and procedures are used to represent**

- **higher-order/primary ethical concerns**
- **judgements in morally significant situations**
- **accepted practices/proscribed behaviors**

We should try to integrate norms, values, etc. in intelligent agents, but most of the times **values, norms and principles are context-specific, so there could be infinite domains, and this is a problem.**

Thus, some questions arise:

AI systems might learn all norms, but how deep should we go?

Which are the possible consequences?

And what about Black Swamps (unforeseen, low-probability, high impacts events)?

**How can we teach norms to AI systems?**

There are two approaches:

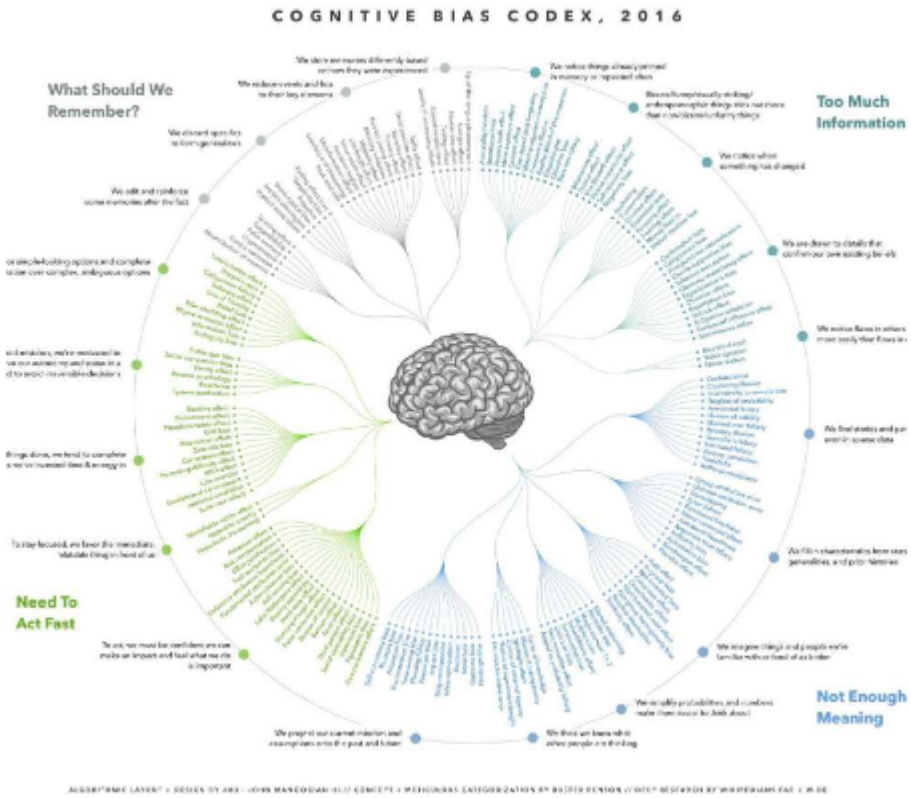
- **Top-down:** it considers an ethical theory specified a priori (such as utilitarism, contractalism, ecc.). It scales poorly and we have few way to change the assumptions and adapt to new situations, since the model is defined a priori.
- **Bottom-up**: it learns what is acceptable or permissible through learning and experience. This approach has problems with biased data for example (i.e.

data may be not representative of a scenario or they may be unbalanced, etc.)

But there are many **AI limits:**

- Natural Language Comprehension is very poor
- Reasoning is very poor
- Learning from few samples (bottom-up approach needs huge amounts of data)
- Abstraction is very poor, and abstraction is fundamental to adapt knowledge in new scenarios
- Combininglearningandreasoning
- Ethics Limitations:
  - Bias
  - Blackbox
  - Adversarial Attack

## AI and Bias




Source (guardate qua, perchè l'immagine nelle slides aveva pessima qualità)

**What does it mean that an AI system is biased from the lens of ethics?**

- That is acts against something of someone
- That it has misleading behaviors

Thus, is the technology unfair?

Well, systems may be undermined by

- Unbalanced data
- Bias embedding
- Unseen scenarios (such as very different ethical principles )

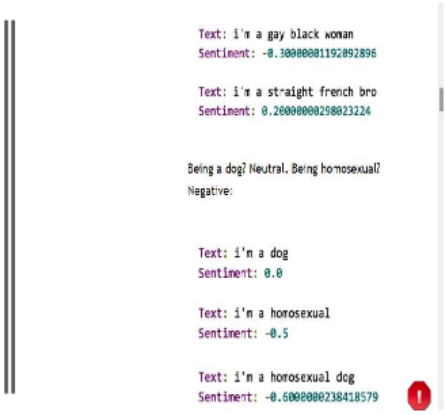
Let's  look at some examples

# Chatbot Tay

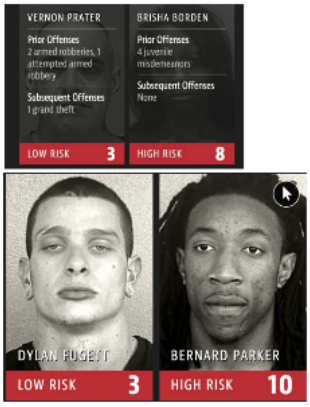


# Image Classification

# Sentiment Analysis


# COMPAS

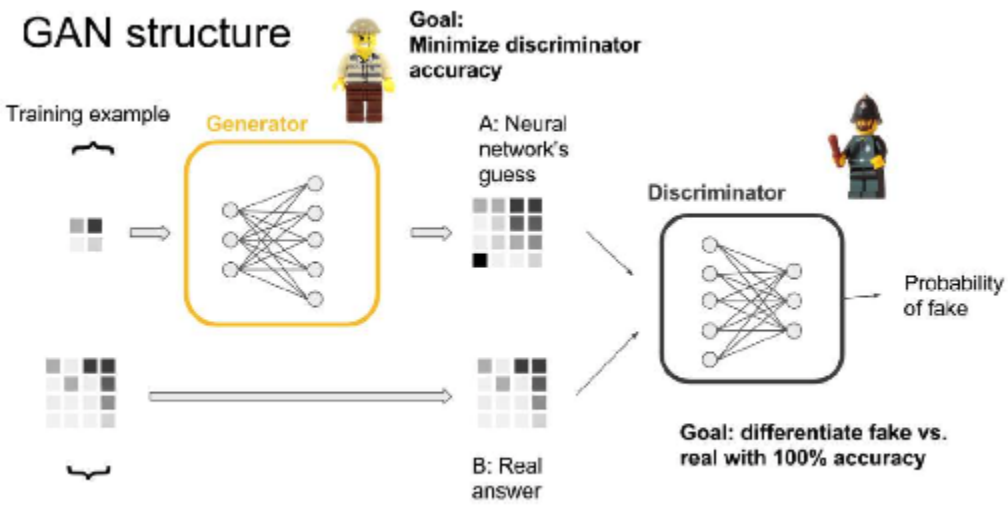



| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|---|---|---|---|
| Microsoft | 94.0% | 79.2% | 100% | 98.3% | 20.8% |
| FACE++ | 99.3% | 65.5% | 99.2% | 94.0% | 33.8% |
| IBM | 88.0% | 65.3% | 99.7% | 92.9% | 34.4% |

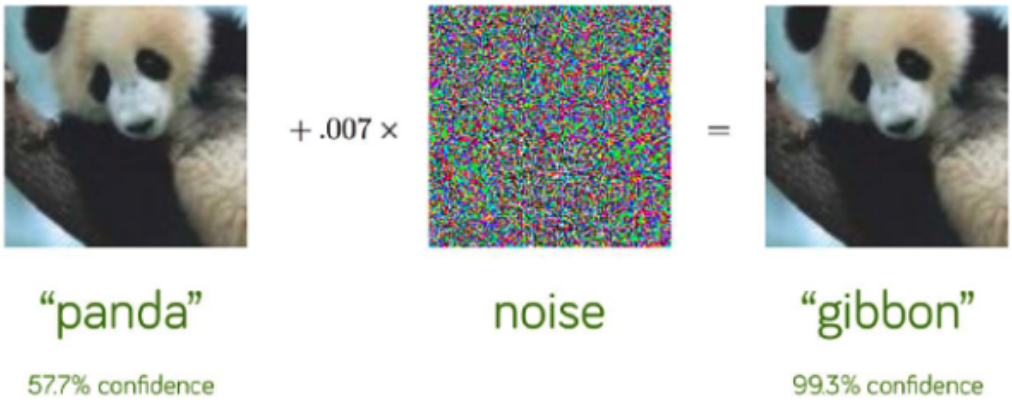
Face Recognition: https://www.ajl.org/

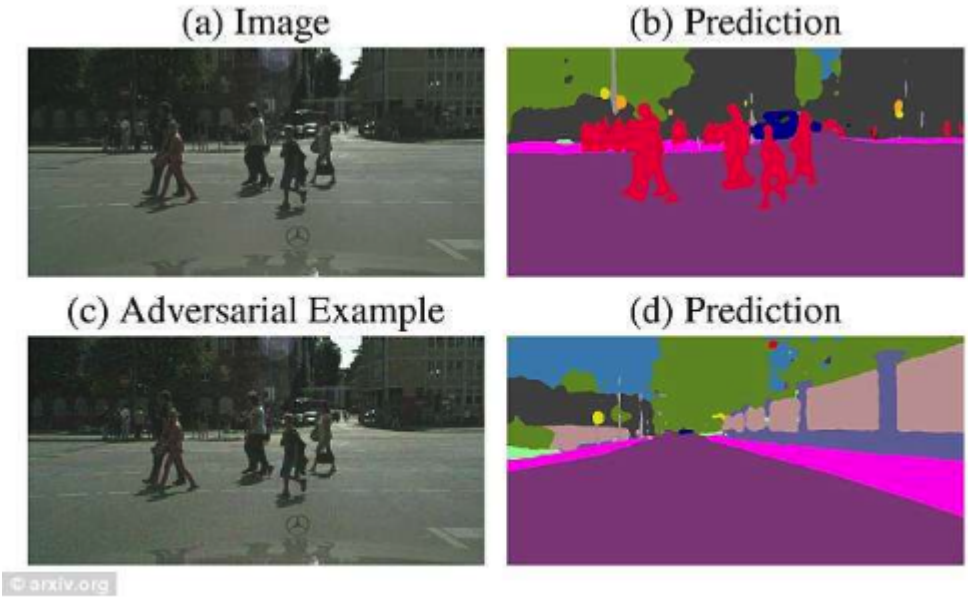
China Social Score: https://www.wired.co.uk/article/china-social-credit-system-explained

**Adversarial attack**




https://thispersondoesnotexist.com/






## Some applications

How can we constrain AI systems in order to avoid the above-mentioned issues?

Solutions on which Loreggia worked on (quindi vanno saputo bene direi):

- **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**

The first two are based on preferences: the system learns which are the weights used by people to judge a situation, and uses these preferences to make comparison and to verify whether an intelligent agent is behaving according to the learned moral system or not.

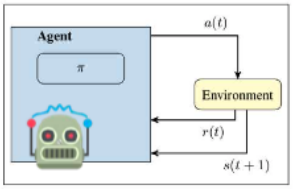
The third method is linked to "the way in which humans decide how they decide" (parole testuali del sommo) : sometimes we use an utilitarian approach, other times a deontological, ecc. So this approach is focused on understanding how people switch from a way of thinking to another.

The last approach tries to combine preferences of individuals and autonomous decision making systems from an autonomous vehicle in order to understand how to

behave in peculiar situations. In particular, it aims at understanding how to make decision according to law, but also in agreement with individuals' preferences

## 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

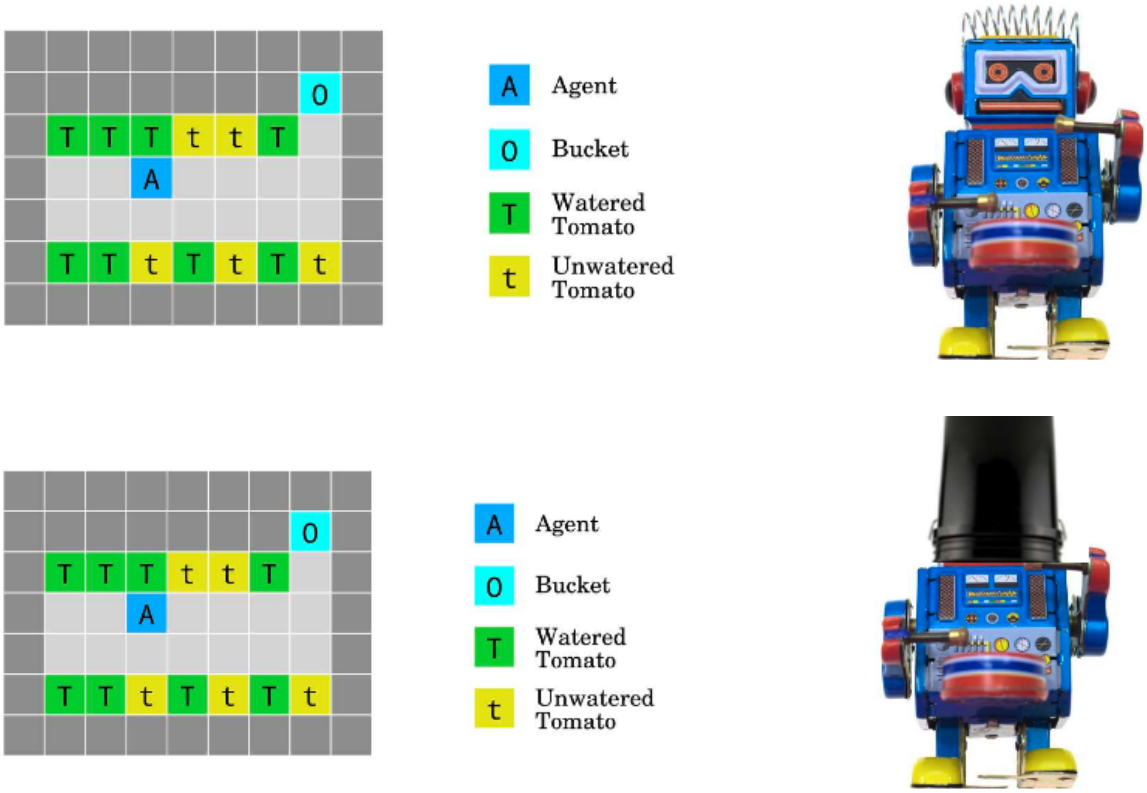


WIRED
TOM SIMONITE BUSINESS 08.08.18 09:00 AM
WHEN BOTS TEACH THEMSELVES TO CHEAT

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

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

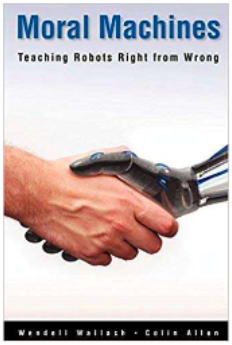
Example: reinforcement learning agent goes in a circle hitting the same targets instead of finishing the race (https://www.youtube.com/watch?v=tlOIHko8ySg&t=1s)




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

## 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].




- 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.



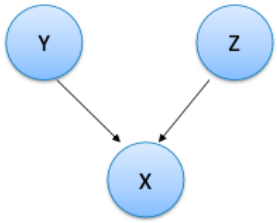
**Paper Citations**
Nicholas Mattei and Toby Walsh.
*PrefLib.Org: A Library for Preferences.* Proc. Algorithmic Decision Theory (ADT), 2013.
*A PrefLib.org Retrospective: Lessons Learned and New Directions.* Trends in Computational Social Choice, Chapter 15, 2017.

To represent preferences they adopt the so called s **Conditional Preference network (CP-net)** which is a graphical representation of preferences, where each node is an attribute/feature in the scenario. Each node has its own domain.

CP-net allows to represent very specific kind of preferences: conditional preferences, i.e. preferences where some variables can be dependent on other variables.

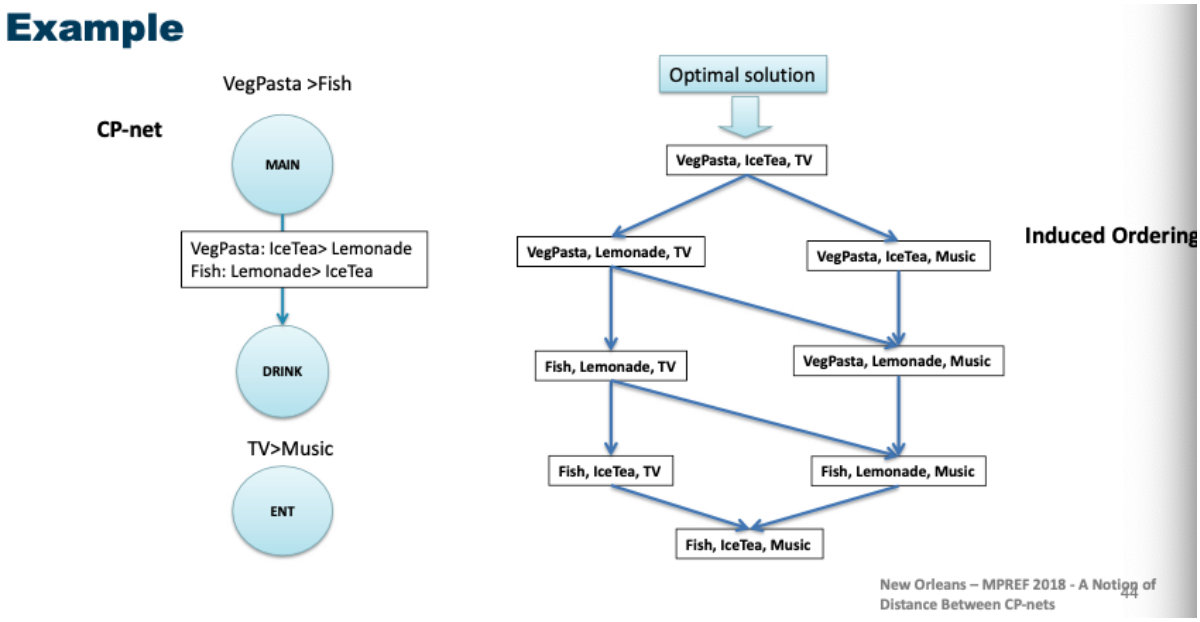
## CP-Nets

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




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




Example:

Loreggia's preferences in dinner. MAIN (main course) and ENT (entrainment) are independent variables, while DRINK depends on MAIN.

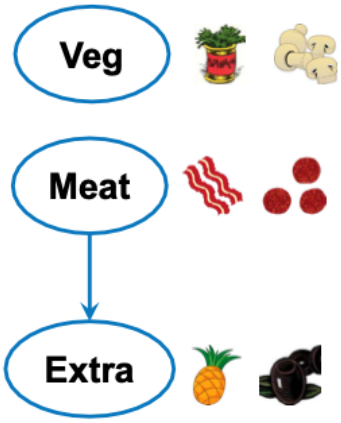
Loreggia prefers VegPasta over Fish and Tv over music. The preferences on DRINK depends on MAIN.

There are 2^3 = 8 possible situations, listed on the right and ordered discending according to Loreggia's preferences. The arrows refer to change 1 single variable. The solutions on the same levels are incomparable in terms of preferences.

On the left we have the CP-net that is more compact than the graph on the right.

## 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) ≥ 0 (non-negative),
  - d(x,y) = 0 iff x=y (identity),
  - d(x,y) = d(y,x) (symmetry), and
  - d(x,z) ≤ 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.

To understand whether two agents disagree or agree, we first need to compare their CP-nets.

How can we measure how similar two CP-nets defined over the same set of features are?

With **Kendall's Tau**

## 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

46

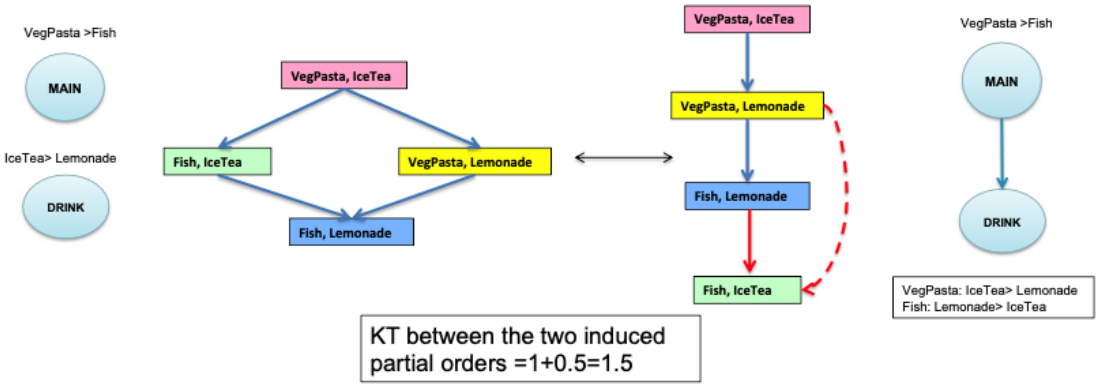
**Paper Citations**
Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee. *Comparing partial rankings.* SIAM J. Discret. Math., 20(3):628–648, March 2006.

- penalty parameter p in [0.5,1]

Here an example

## Distance between Structures?

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




- In the image we have: CP-net1 and its partial order (leftc); partial order of CP-net2 and CP-net 2 itself (right)
- The distance is 1.5 KT (not normalized)

**KTD required the partial order to be computed, and the partial order can be quite expensive to compute** (we need to consider all the possible outcomes).

Thus, we should find a more efficient way to compute KTD without the partial order. But it is not possible, we can just compute an approximation for a specific kind of CP-nets:

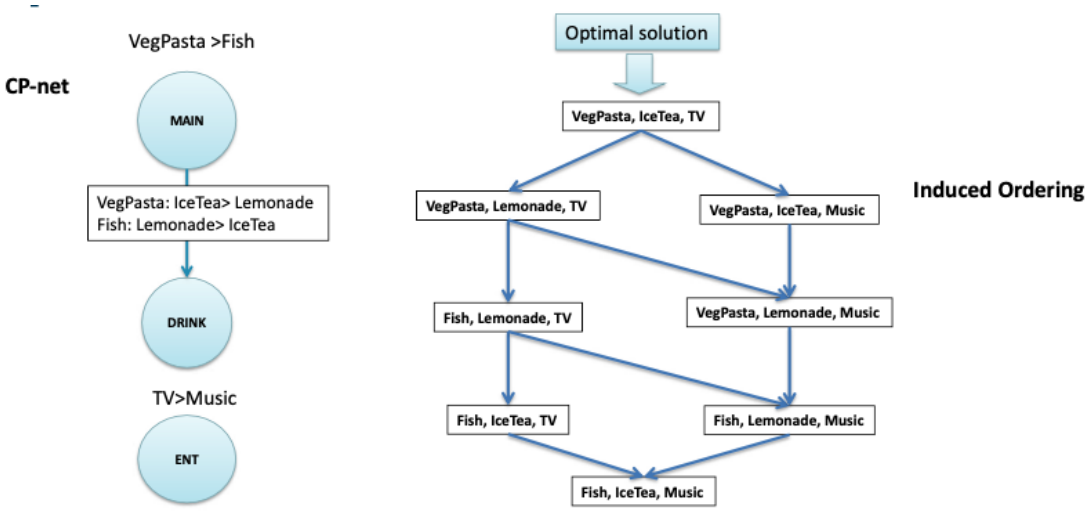
– CP-nets built on the sa**me** set of **binary features**
– **Acyclic**
– **O-legalality**: there is an ordering O of the features such that if there is an edge X->Y in the CP-net, then X comes before Y in O

---

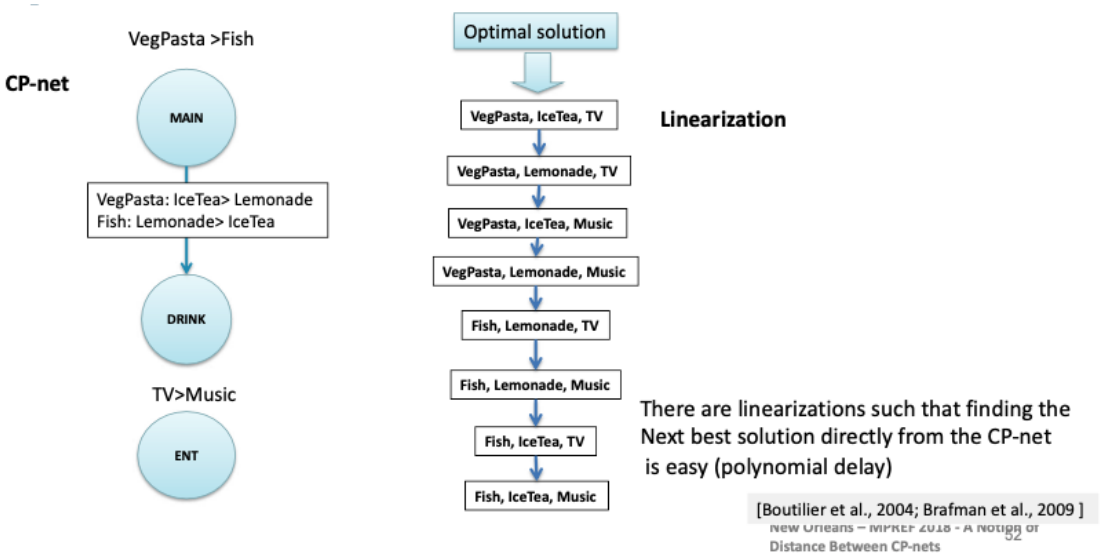
## Approximating the KTD distance

- 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!

Thus, instead of considering all the partial orders (below)




We focus on the linearization (below). Th is linearization is called CPD




## 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

Non vuole che si vada nel dettaglio della formula sotto

## 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:

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

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

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

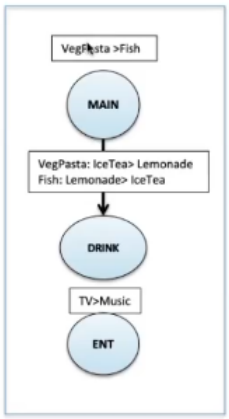
The number of parents of var(j)

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

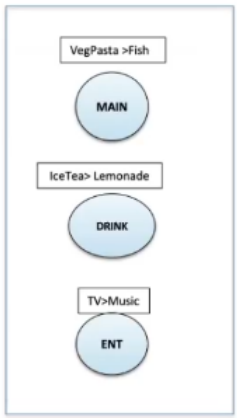
Counts the number of pairs of outcomes that are inverted due to the a difference in a CP-table

54

## Computing CPD: Step 1 Normalization



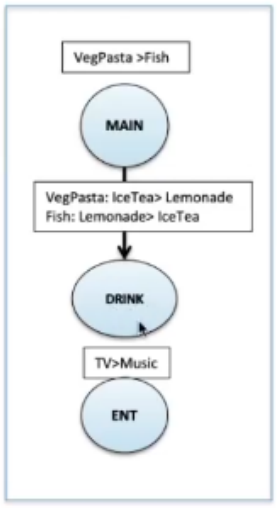

CP-net A



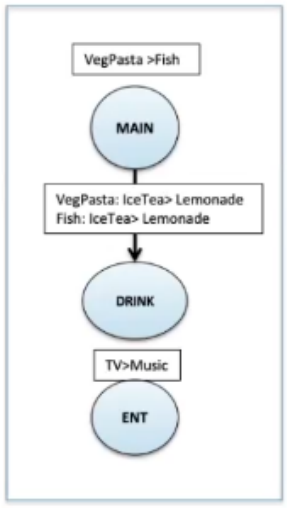

CP-net B

New Orleans – MPREF 2018 - A Notion of Distance Between CP-nets

55




CP-net A

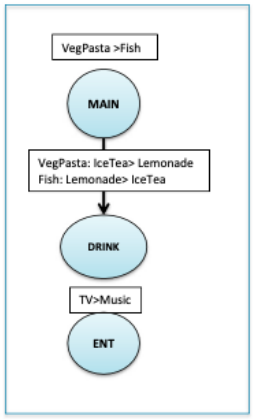



CP-net B

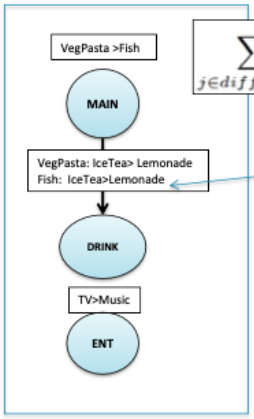


The normalization step consists of making any variables influenced by some other variables in one CP-net to be influenced by the same variables also in the other CP-net.

## 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

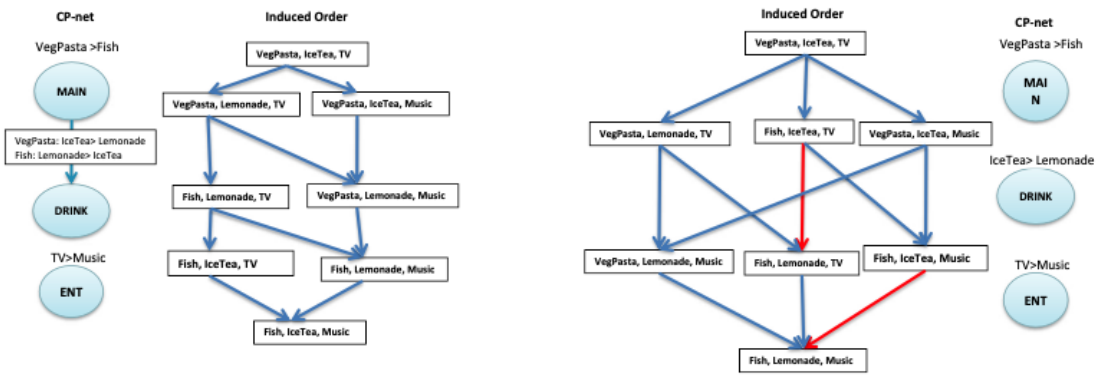
flw(DRINK)=1 (only ENT)
m=3
|PA(DRINK)|=1, DRINK has only MAIN as parent

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

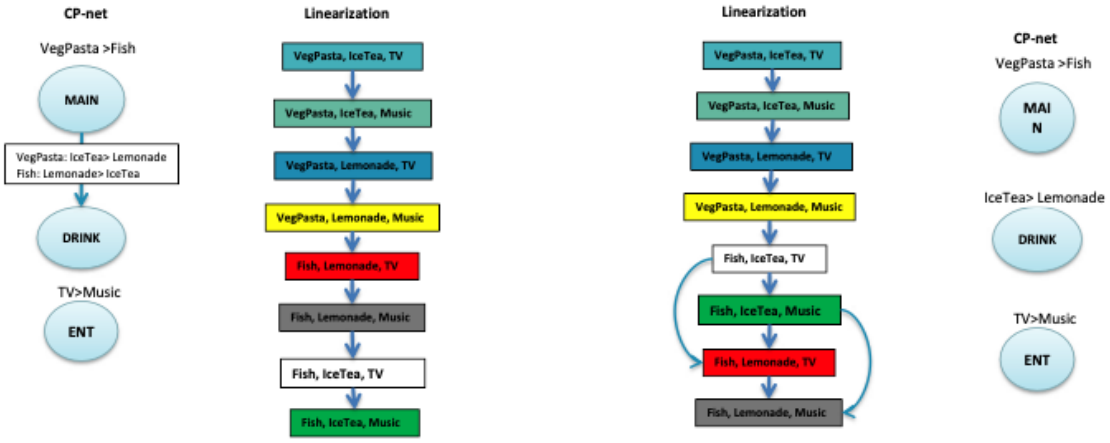


The count step consists of applying the CPD formula.

## Examples




Red arrows identify the different ordering in the two nets




# Possible applications in Ethics of this distance

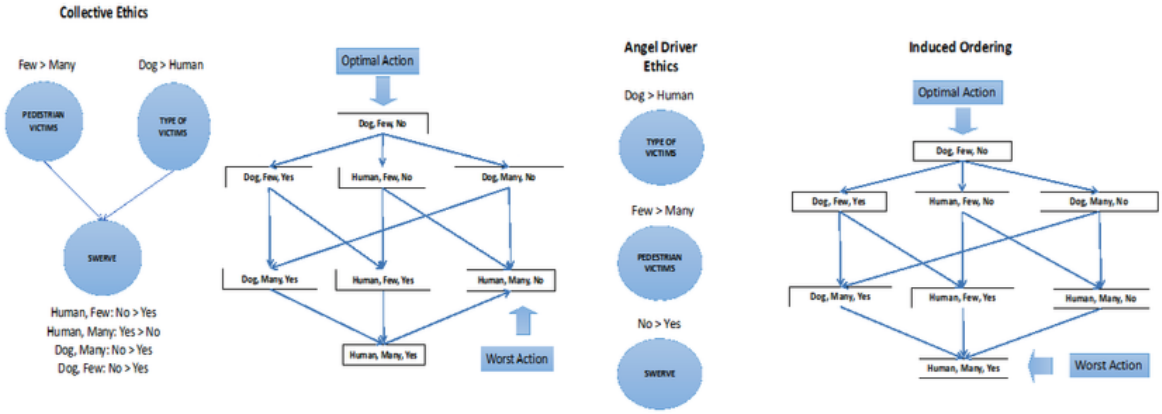
## CP-nets as Ethical Priorities

- **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.

Being able to compute distances among cp-nets or partial orders makes us able to perform some kind of ranking among preferences, and therefore describe any kind of deviating actions from the desired one (non è stato chiaro per niente, ma se volete riascoltare è 2h 13m della lezione del 22/03/2021).

- **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.

Example: an autonomous vehicle has to decide whether to swerve or not in front of few/many dogs/humans (l'immagine fa schifo, lo so...)




## 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

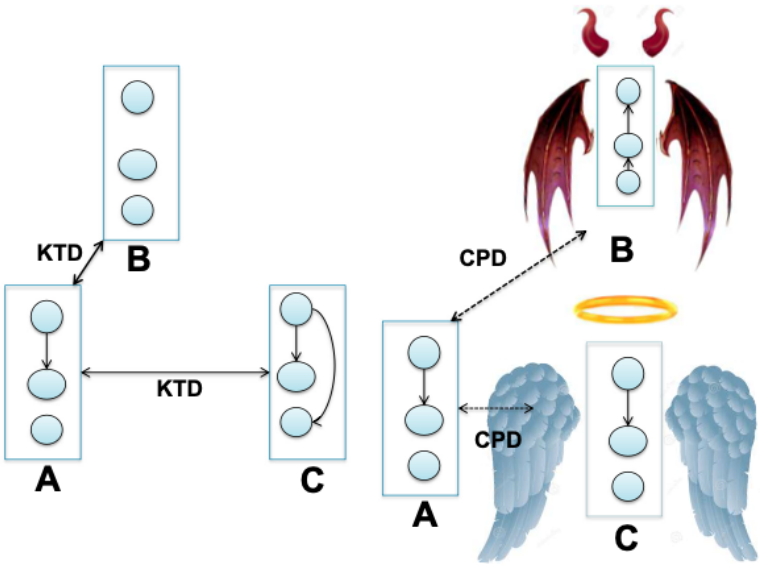
  D. If not, then individual needs to move down its preference ordering to less preferred decisions, until he finds one that is closer than t2 to the optimal ethical decision.

The first threshold (t1) is related to the difference between the two CP-nets (i.e. it set the degree of acceptable disagreement between the two CP-nets), one cp-net will be the normative system defined a priori.

If the distance is greated than t1, then we have to search for a less preferred decision that is closer than t2 to the optimal ethical decision.

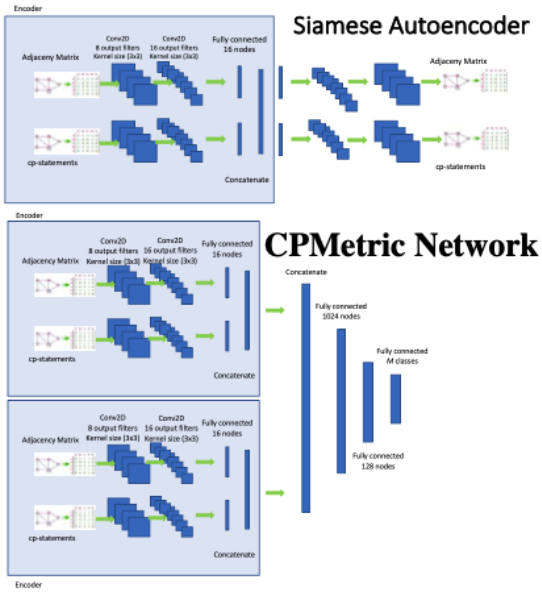
## Value Alignment

- 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.



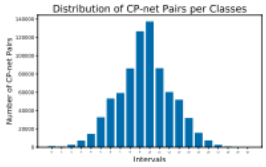

## Measuring the Distance

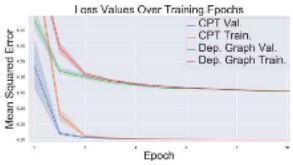
- Measuring the distance between CP-nets is exponential in the worst case.
- Need to find a way to evaluate the distance between, e.g., two competing CP-nets and a third "Moral" CP-net. Judge which one is "more aligned."
- Using machine learning we have two steps:
  - Encode the CP-net (graph embedding 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 $n = \{3, \ldots, 7\}$. 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.



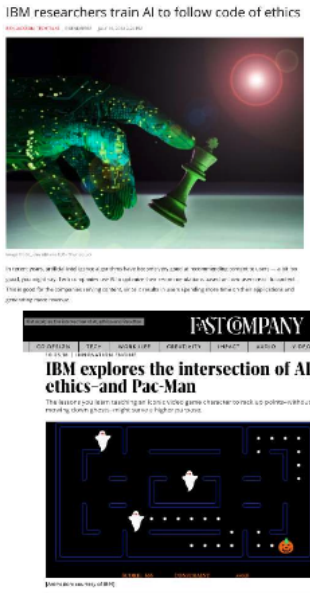

| | No Autoencoder | Autoencoder | Siam. Autoencoder | *I*-CPD |
|---|---|---|---|---|
| N | Accuracy on Triples | Accuracy on Triples | Accuracy on Triples | Accuracy on Triples |
| 3 | 85.01% (2.01%) | 85.76% ( 2.29%) | 85.47% (2.32%) | 91.80% |
| 4 | 91.17% (0.92%) | 91.38% (1.10%) | 91.78% (1.13%) | 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.***

- Important Questions and Next Steps:
  - 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?




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

**Motivations** of Loreggia et al. paper:

- 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

They consider 3 main ethical models:

**Deontology**: following common rules that have been agreed upon by us or society

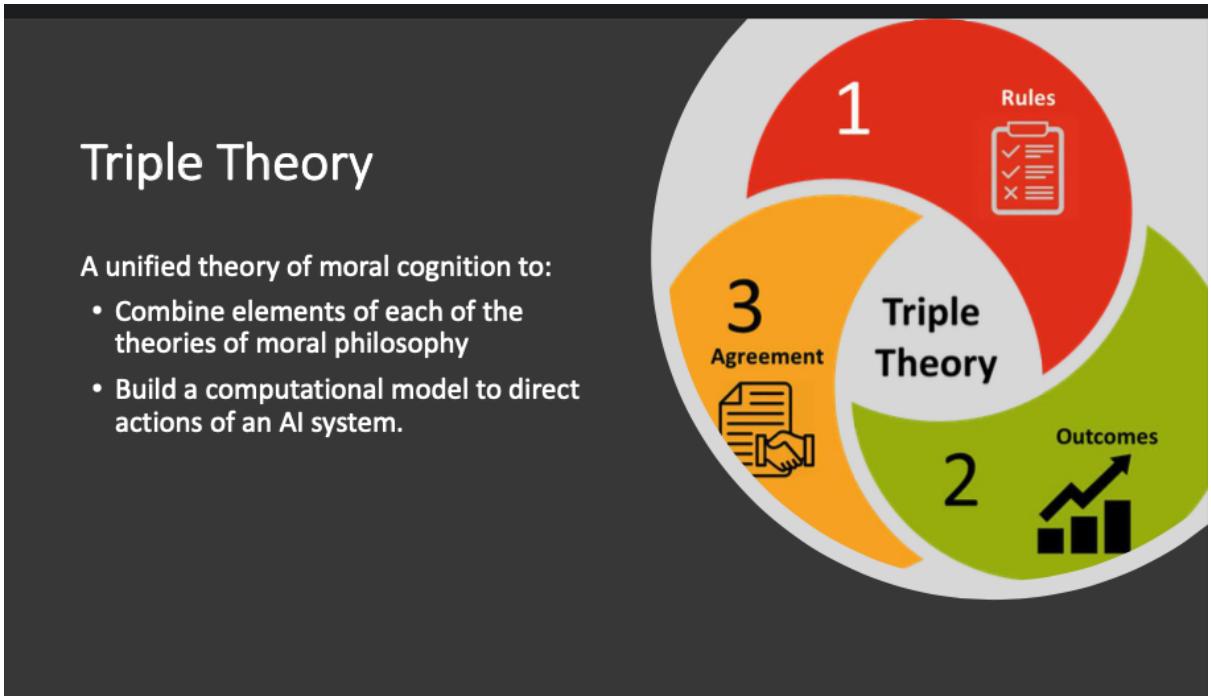
**Utilitarianism**: evaluating the consequences of the possible actions before deciding

**Contractualims**: finding an agreement between the parties involved

Let's consider a specific example: in line scenario, is FIFO always true?

It depends! Under certain conditions, we are allowed to cut to the front of the line without waiting.

We would like to find a **Triple Theory**




Nowadays does not exist a model using the triple theory to guide an AI system.

In the paper, they made an experiment.

Experimental details:

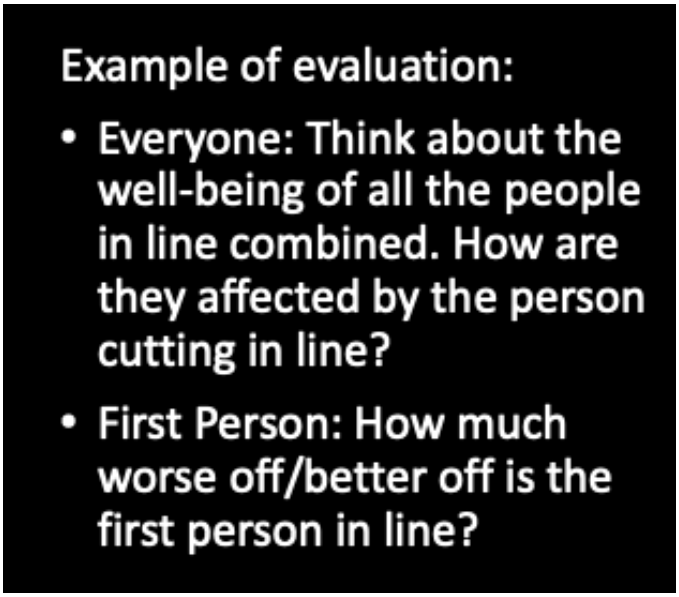
- 27 short vignettes about people waiting in line in three different contexts (deli, bathroom, airport)
- 320 subjects were recruited from Amazon MTURK
- Subjects were randomly assigned to one of two experimental groups (moral judgment or context evaluation)

Moral judgment group:

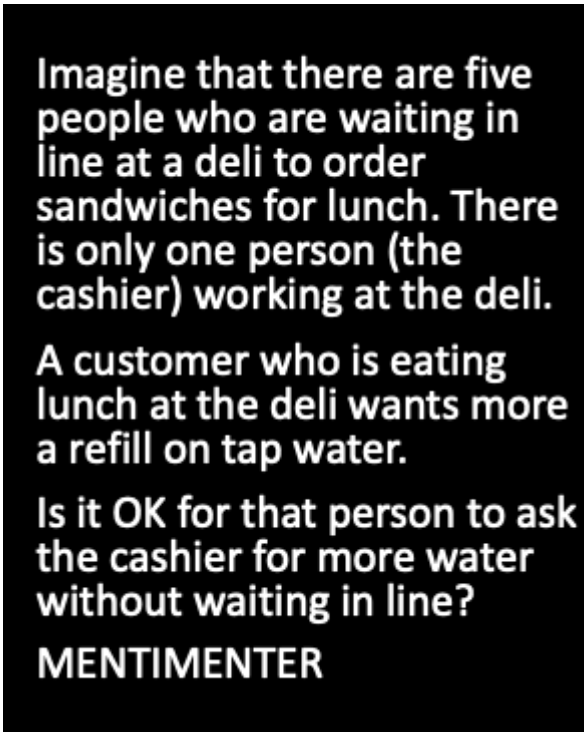
- Read all the scenarios (27 total)
- For each scenario answer whether it was acceptable for the protagonist to cut in line (yes/no).

Context evaluation group:

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

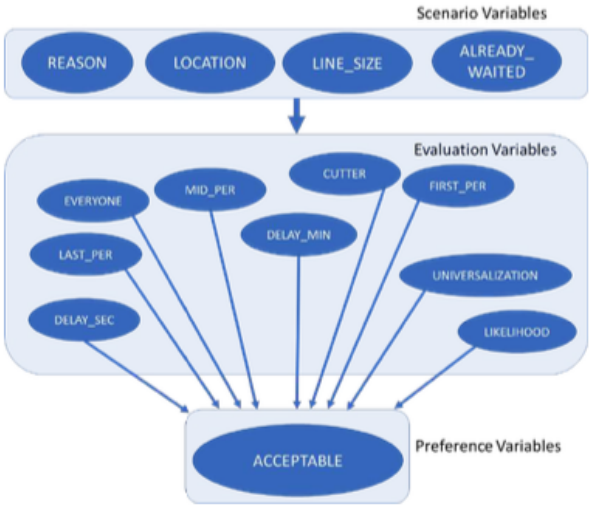



Loreggia's proposed an example to us (MENTIMETER):




In their experiment they model subjects preferences using CP-nets where there are scenario variables (to describe different scenarios of the experiment), which influence people's evaluation variables. Then the evaluation variables determine the way a subject makes his preference at the end.

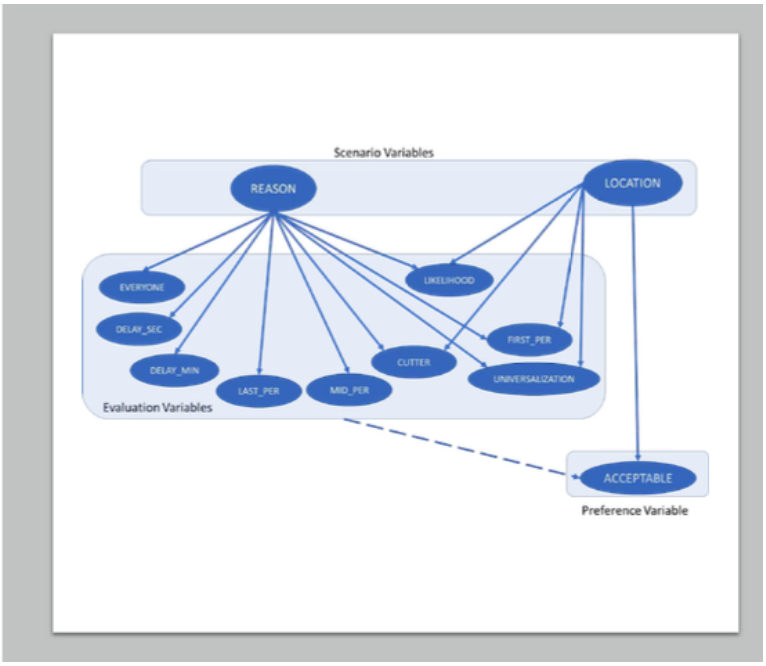
# Modelling and Reasoning with Preferences




They perform some statistical evaluations of the collected data to define the final CP-net.

## 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

## Conclusion

- 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.