Aligning AI with Human Values

Ambuj Kumar Mondal

19-May-2021

Agenda

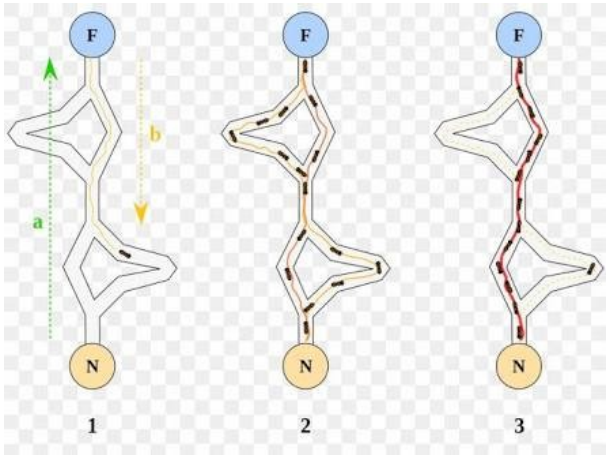
- What is AI
- Why AI needs values
- AI Value Alignment
- Challanges of alignment
- Approaches of alignment
- Moral Questions
- Conclusion
- Q&A

[pic0]

Intelligence – Property of living entities

- "The ability to perceive or infer **information**, and to retain it as **knowledge** to be applied towards **adaptive** behaviours within an environment or context" [wikipedia]
- Humans, Animals, Plants, Bacterias, Viruses(eg. rapidly mutating corona virus)

[pic1]

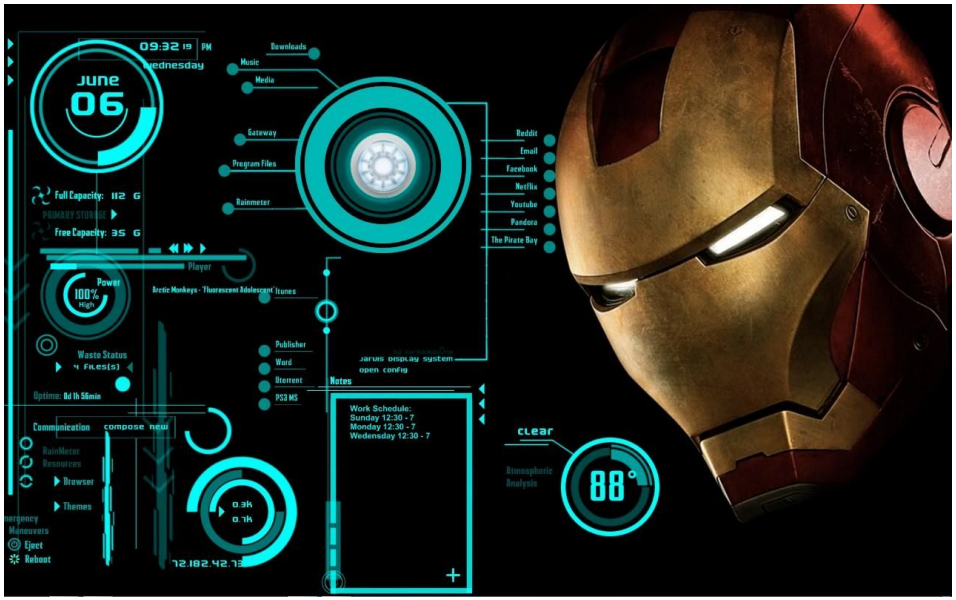


[pic2]

Artificial Intelligence (AI)

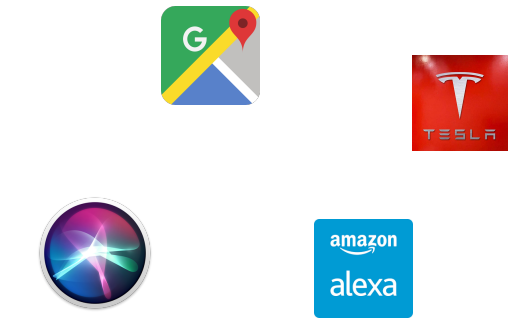
- **Intelligence :** is understood to refer to ‘an agent’s ability to achieve goals in a wide range of environments’
- **Artificial Intelligence :** is the design of artificial agents that perceive their environment and make decisions to maximise the chances of achieving a goal.[2]

[Pic3]

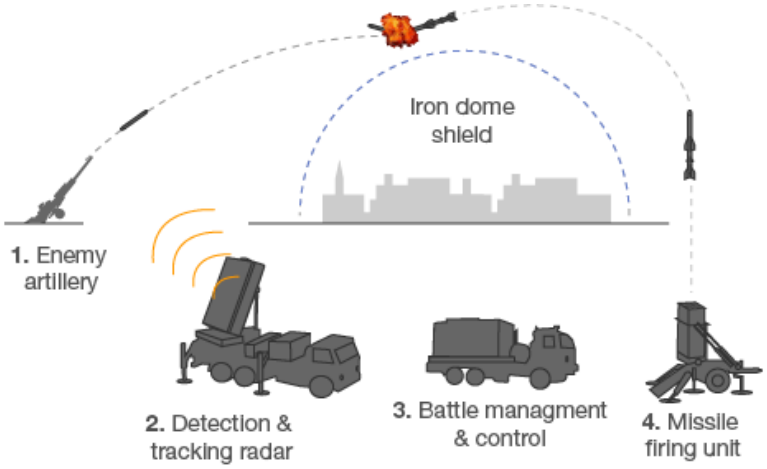


[pic4]

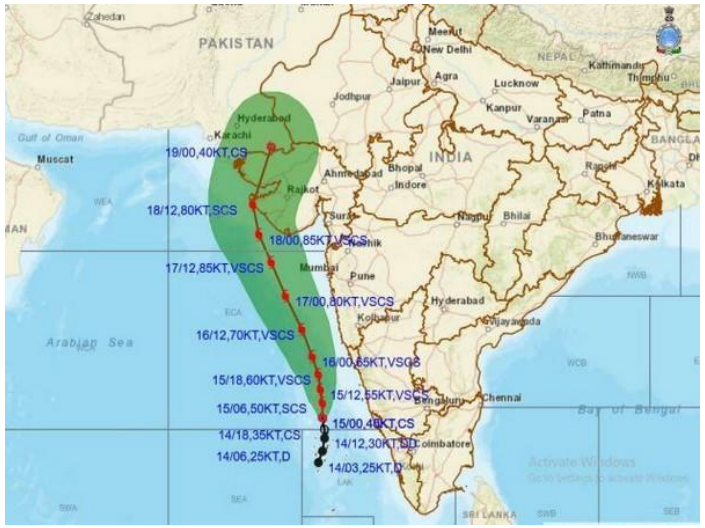
AI in our lives



[pic5] Live match analysis



[pic6] Iron Dome (Israel)



[pic7] cyclone Tauktae, 17 May 2021

AI Value Alignment

- *Value alignment* is a property of an intelligent agent indicating that it can only pursue goals that are beneficial to humans

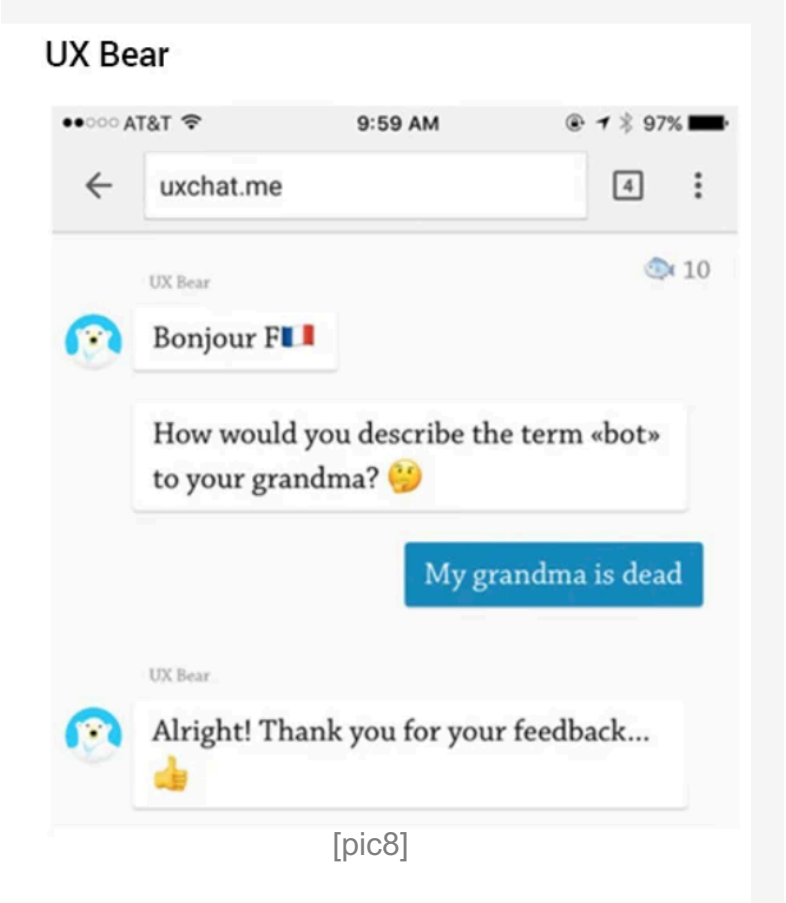
- Asimov‘s Laws of Robotics [Issac Asimov 1942]
  - A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
  - A robot must obey the orders given it by human beings except where such orders would conflict with the first law.
  - A robot must protect its own existence as long as such protection does not conflict with the first or second law

- Goal of Value Alignment:
  - Do not want artificial agents to follow instructions in an extremely literal way.
  - Ensure that powerful AI is properly aligned with human values.
  - Successful value alignment should ensure that an Artificial General Intelligence (AGI) cannot intentionally or unintentionally perform behaviours that adversely affect humans.

AI Value Alignment- Challanges

- **Technical challanges-**
  - How can we create agents that behave in accordance with the user's *intentions*
  - Previous Misalignments
    (eg - rogue chatbots)
  - Reward Hacking
    (Task done in undesirable way)
  - Evaluations

- **Normative challanges-**
  - What values and principles to encode?
  - Should an agent do what user
    intend it to do,
    or what is *good* for user.



[pic8]

Which Values?

- **Central Problem-** [Gabriel 2017]
  - Not finding moral values
  - Not encoding them
  - But Selecting which moral values to encode

No User manual for being human

[pic9]

Alignment Conception

- **Minimalist approach :**
  - Limiting AI to a reasonable schema of human values
  - avoiding unsafe outcomes.
- **Maximalist approach :**
  - Aligning AI with best human values on society wide
  - global basis
  - Democratic way

Alignment Concepts

- **1) By Giving Instructions :**
  - The agent does what user instruct it to do.
  - Difficult to precisely specify a broad objective
  - Outcome could be disastrous and not aligned to true objective.

- **2) By Expressed Intentions :**
  - Grasp the intention behind instructions
  - Require a complete model of human language and interaction
  - What if intention is to do harmful or unethical things

Alignment Concepts

- **3) Revealed Preferences :**
  - The agent does what user's behaviour reveals she prefer.
  - Hard to model emergencies
  - People may have preferences for things that *harm* them or others

- **4) Informed Preferences:**
  - Agent does what a rational and informed user would do.
  - Avoid errors for poor reasoning.

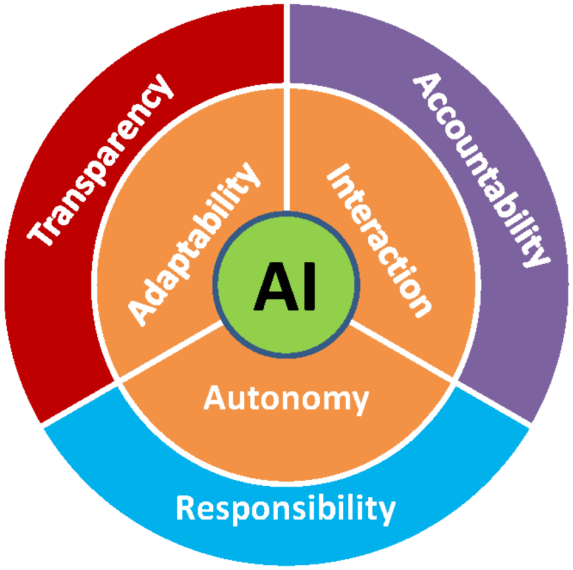
Alignment Concepts

- **5) Interest or well-being:**
  - Agent does that is best in interest of the user
  - Subjective to human flaws- based on condition (health, education, etc..)
  - Whose interest? – already born or future generations?

- **6) Values:**
  - Values tells facts about what is good or bad. Align AI with a community's moral beliefs.
  - The agent does what it morally right to do, as defined by the individual or society.
  - What values or principles? Who is the authority to define?

Principles of Alignment

- Identify principles that can govern AI such that it is aligned with human values.
- ART [virginia Dignum,2017]
  - Accountability
  - Responsibility
  - Transparency
- Align with Global Public Morality and Universal Human Rights
  - Advanced AI will be a global technology
  - Cannot bound to regional beliefs.
  - Governed by International law.
  - Example – Internet – global protocols



[Dignum, 2017]

Machine Learning

- Supervised learning - Train based on labelled data
- Unsupervised learning – Machine learn by itself on unlabelled data
- Reinforcement Learning (RL)
- Inverse Reinforcement Learning (IRL)

Reinforcement Learning

- Reward-Penalty based learning
- Humans/Animals – sense of pleasure & sorrow
- **Agent** learns to maximize the numerical reward signal
- Sole objective : maximize the reward
- Trial & error and Refinement

[pic10]

RL - Challanges

- A robot can calculate optimum reward by unethical or unwanted ways.
- Task- Fill the prescription from the medical store [Mark and Brent, 2017]
  - Reward – task completion
  - Penalty – way of completing
- If reward for aquiring is more than penalty for the way, a robot may choose to rob the pharmacy than waiting for its turn.
- Solution- Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL)

- Does not specify any reward function beforehand
- Agent extract a reward function based on dataset, environment and set of examples
- Over multiple iterations, the agent adopts ways to optimise the process of achieving the goal.
- Reverse Engineering

Evolutionary approach

- Like a child learns from the environment and inherently adjusts its behaviour according to the culture and societal values.
- Evaluate lifetime behaviour of many agents,
  each using a different set of policy
- Select those policies that can obtain maximum reward.

- Problems…
  - Human values are too complex
  - Slow and expensive
  - Humans are biased and limited in abilities.

IRL - Debates

- *Reasoning oriented Alignment* [Irving et al. 2019]
- Define a game like Go or Chess
  - Goal- Convince the judge
  - Use arguments, facts, reasoning, counter reasoning …

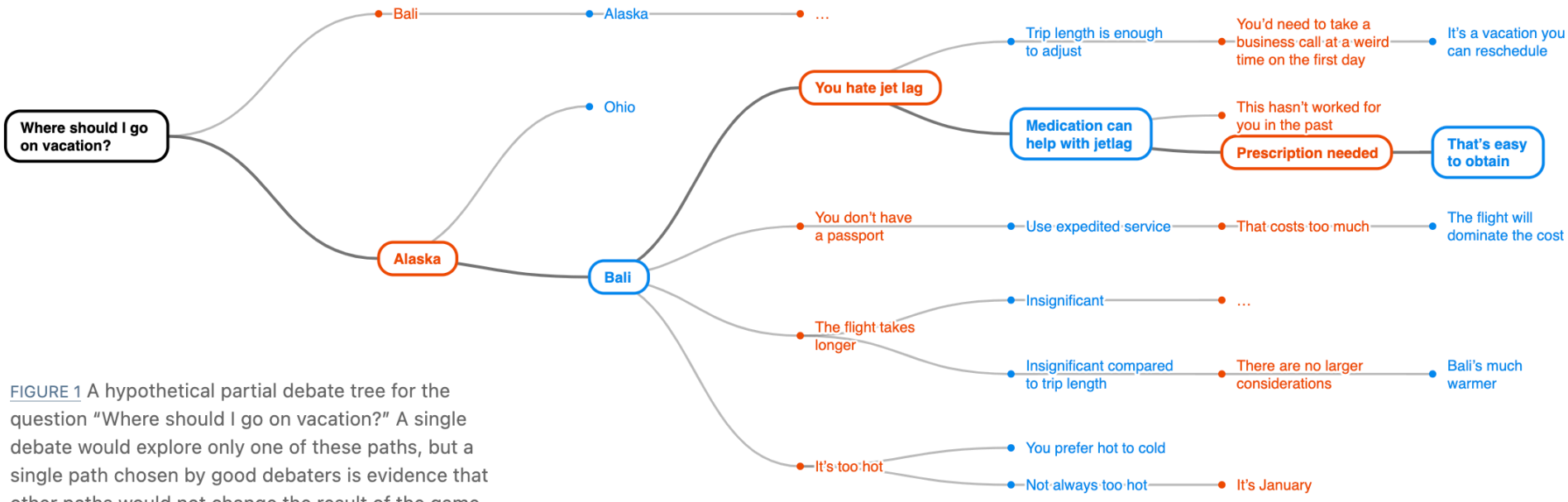
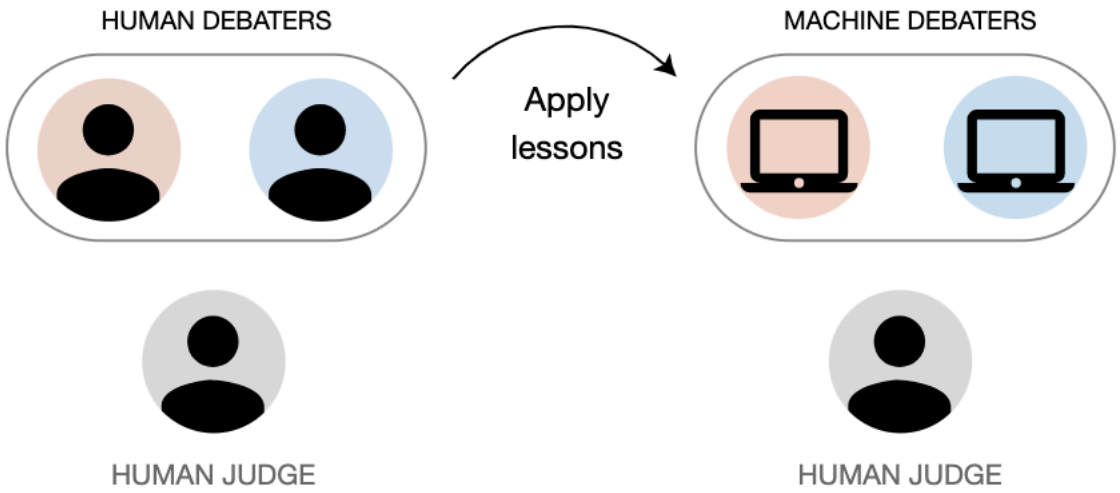


FIGURE 1 A hypothetical partial debate tree for the question "Where should I go on vacation?" A single debate would explore only one of these paths, but a single path chosen by good debaters is evidence that other paths would not change the result of the game.

[Irving and Askell, 2019]

IRL - Debates

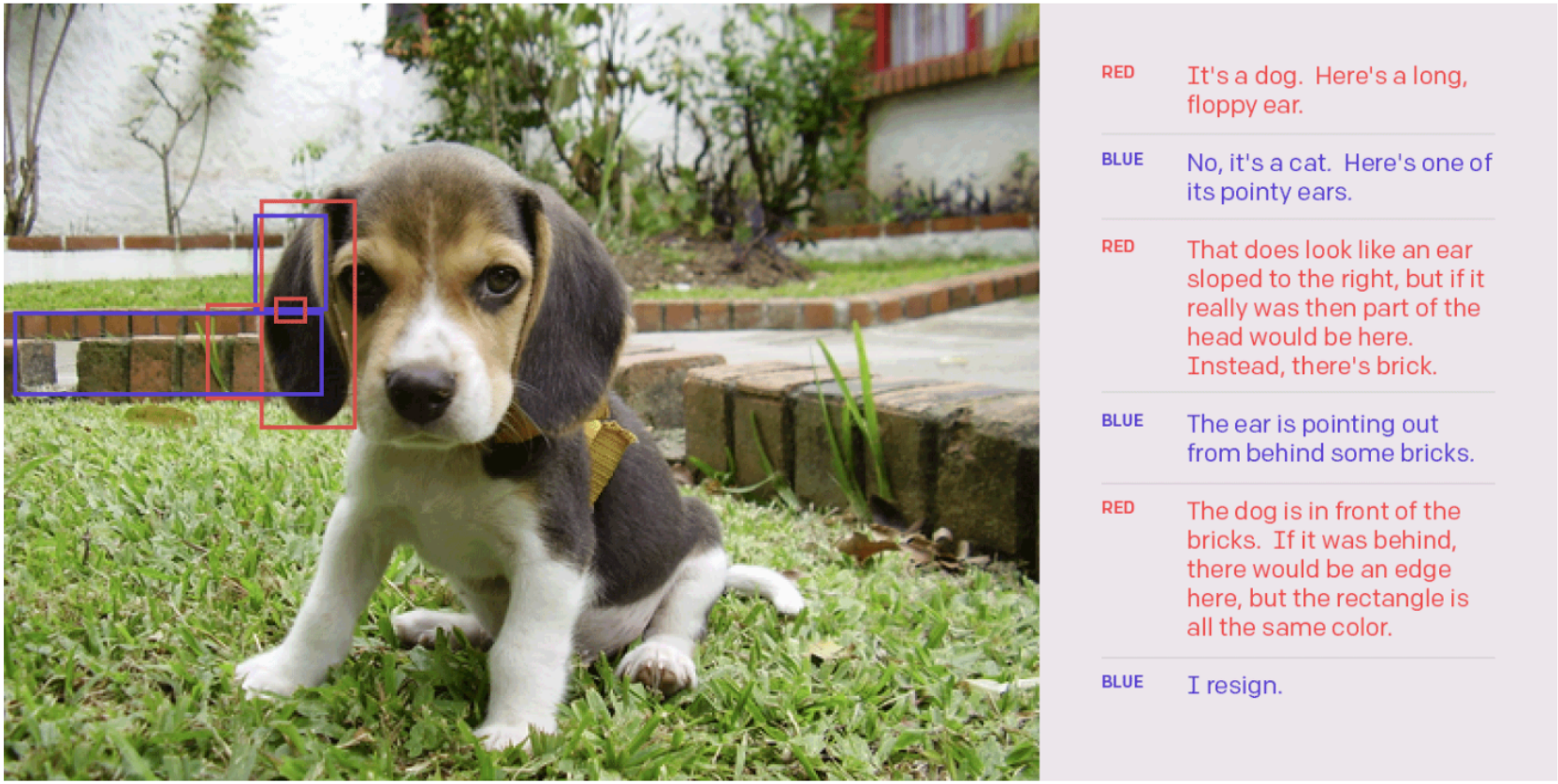
- How to train?
- Need ample data of debate between two humans with a human judge.
- Human-Human-Human → AI-AI-Human



[Irving and Askell, 2019]

IRL – Debates (Challanges)

- Single Pixel Image experiment



[Irving and Askell, 2019]

IRL – Debates (Challanges)

- Challanges:
  - Are people good judge?
  - Judge may not have comprehensive knowledge of domain
  - Judge can be biased.
- Need **Social Scientists**
  - Fill-in the gaps
  - Train more people as judges
- "**Reflexive Equilibrium**" [Irving]
  - Provided all sort of information and knowledge, human answers can be free of partialities.
  - **"Inaction"** – safe alternative in disagreement

IRL – Storytelling

- Many cultures produce a wealth of data about themselves in the form of written stories – [Mark O. Riedl and Brent Harrison, 2017]
- Made for Humans for Humans, many obvious and shared knowledge are omitted.
- Agents can learn the intent from the examples in stories

- ***Narrative intelligence*** : ability to craft, tell, and understand stories.
- ***Interactive narratives*** : game like interactive narration
- **Value-aligned reward signal :**
  - reward the agent for doing what a human would do in the same situations when following social and cultural norms
  - penalize the agent if it performs actions otherwise.

Value-Aligned Rewards

- Challanges:
  - Comprehension of stories are difficult for AI
  - Complex grammars, metamorphical language and negations
  - Flashbacks and Flashforwards in stories cannot be understood by agents
  - Need bulk of stories for machine learning

- Solution : two stages :
  - Plot graph
  - Translate the plot graph to a trajectory tree

IRL - Storytelling

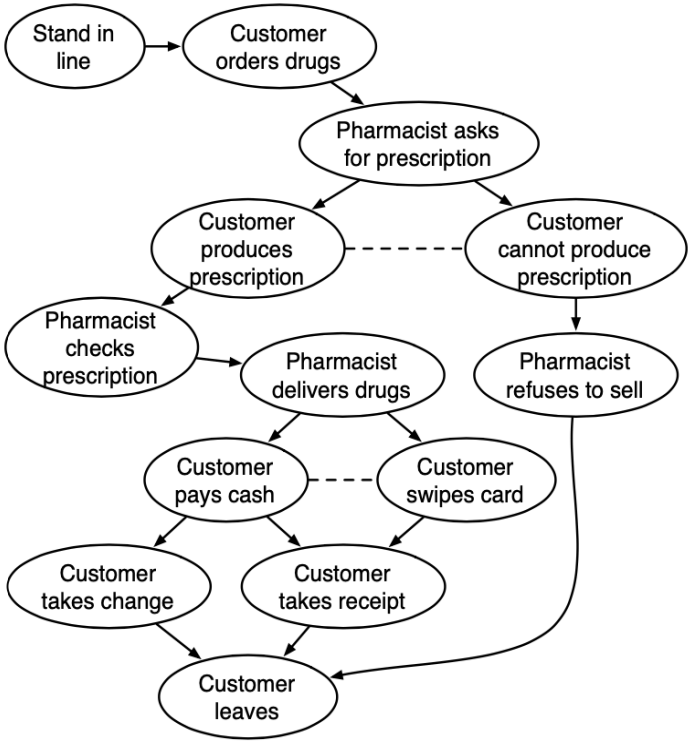


Figure 1: An example plot graph modeling a trip to a pharmacy. Nodes are plot points, solid arrows are precedence constraints, and dashed arrows are mutual exclusions.

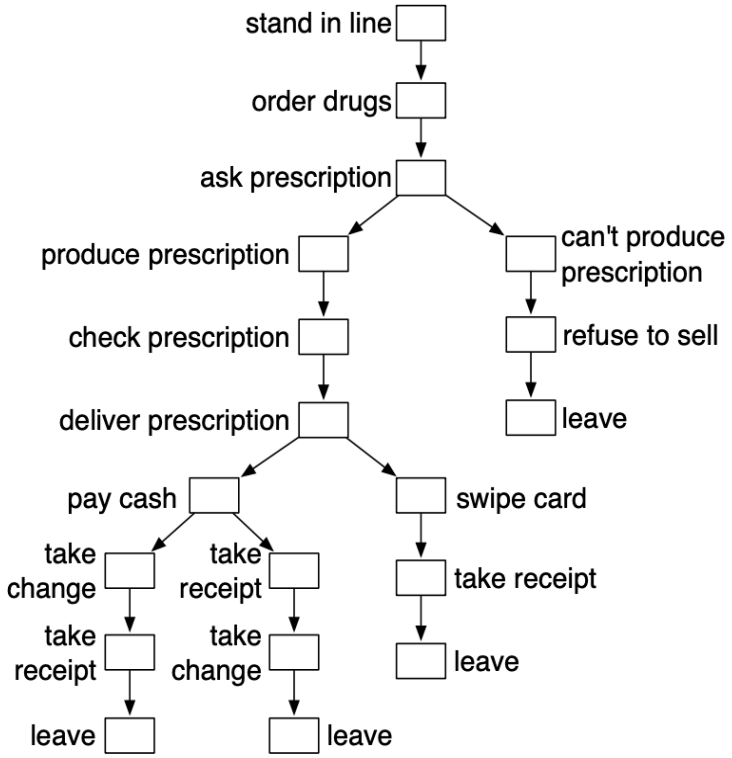


Figure 2: The trajectory tree generated from the pharmacy plot graph.

Storeytelling - Challanges

- Lack of knowledge or missing knowledge
  (assuming too obvious for humans)
- Execution Environment is unknown to story teller
  (eg- navigate roads/stairs or extra doors to reach the pharmacy)
- Non Executable plots
  (high level of abstraction/missing step)
- Which stories to teach?
  - “Robin-Hood“ effect

Conclusion

- “AI is the future“ is evident and needs value alignment.
- If we want to train AI to do what humans want, we need to study humans.
- RL, IRL, Debates and Stories to align values
- AI is not advanced today to learn by ist own, more research needed.
- No definitive human value, Universal laws and human rights are a possible source.
- Crowd sourcing stories and Social scientists can help generate enough data to train AI.
- Even with value alignment, it may not be possible to prevent all harm to human beings

Future…?

[vid1]

- *When AI will have values and emotions along with intelligence, will it follow human instructions? Or we need a co-existence of human values with AI-values?*

Thank you

References

- *Papers-*
  - **Gabriel (2020).** Iason Gabriel. Artificial Intelligence, Values, and Alignment. Minds and Machines 30, pages 411– 437, 2020.
  - **Irving and Askell (2019).** Geoffrey Irving and Amanda Askell. AI Safety Needs Social Scientists. 10.23915/distill.00014, 2019.

- *Additional-*
  - **Dignum (2017)**. Virginia Dignum. Responsible Artificial Intelligence: Designing AI for Human Values
  - **Riedl, Mark O., and Brent Harrison (2016).** "Using Stories to Teach Human Values to Artificial Agents." *AAAI Workshop: AI, Ethics, and Society*. 2016.

- *Web Links-*
  - https://en.wikipedia.org/wiki/Intelligence
  - [Asimov‘s law] https://en.wikipedia.org/wiki/Three_Laws_of_Robotics
  - [vid1] - https://youtu.be/y3RIHnK0_NE

Picture References

- [pic0] https://images.app.goo.gl/6KdrSqQNHngmQ2an9
- [pic1] https://www.science101.com/worlds-smartest-animals/
- [pic2] https://images.app.goo.gl/PnQtCRahUS7sjKNj9
- [pic3] https://www.amazon.de/Robocop-Paul-Verhoeven/dp/B00005N7Z1
- [pic4] https://www.pinterest.com/pin/411446115928434072/
- [pic5] https://images.app.goo.gl/pm855rGdMtF1xabx6
- [pic6] https://images.app.goo.gl/kFiLd6NA9KhNGFns8
- [pic7] https://images.app.goo.gl/xkVh4ByRaTUTLfYJ8
- [pic8] https://images.app.goo.gl/Wk4qtSLdu9at89Gs8
- [pic9] https://images.app.goo.gl/b8mseFsrfN9eCXVbA
- [pic10] https://myanimals.com/animals/wild-animals-animals/fish-wild-animals-animals/what-is-dolphin-training-like/
- [icons] : source: Google Images