Friendly AI (FAI)

Ensuring ethical behaviour of intelligent agents

Vadim Tschernezki

Institute of Computer Science
Heidelberg University

Seminar: Is Artificial Intelligence dangerous?
19.07.17

Table of Contents

1. Definitions

2. Design
   - Agents
   - Reinforcement Learning
   - Value Learning
   - Inverse Reinforcement Learning (IRL)
   - Value Learning with Storytelling and IRL

3. Problems
   - Hidden Complexity of Wishes
   - Fragility of Values

4. Conclusion and Outlook

Definitions [1]

- Moral: standards of behaviour; principles of right and wrong
- Ethics: the branch of knowledge that deals with moral principles
- Intelligence: the ability to acquire and apply knowledge and skills
- Friendly Artificial Intelligence: AI that decides morally/ethically

Agents

- Traditional agents interact with environment cyclically [2]
  - k : cycle
  - y_k : action in k
  - x_k : observation in k
- $y_1x_1y_2x_2 \ldots y_mx_m$: interaction history of agent with lifespan m [3]
  - Also written $yx_{1:m}$ or $yx_{\leq m}$
- Agent function: mapping from $yx_{<k}$ to y_k
- Agent implementation: physical structure, implements agent function
- Why study agent implementations?

Reinforcement Learning

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} (r_k + \cdots + r_m) P(y x_{\leq m} \mid y x_{<k} y_k)$$

- Additionally: concept of scalar reward r_k for each x_k
- For simplicity: using Hutter's AIXI optimality notion for RL [2] [5]
  - Approximate full search of all possible future interaction histories $yx_{k:m}$
  - Find probability of each history
  - Take action with highest expected total reward
- AIXI too abstract? approximation $\rightarrow$ working for games such as Pac-Man, Snake [6]

Wireheading

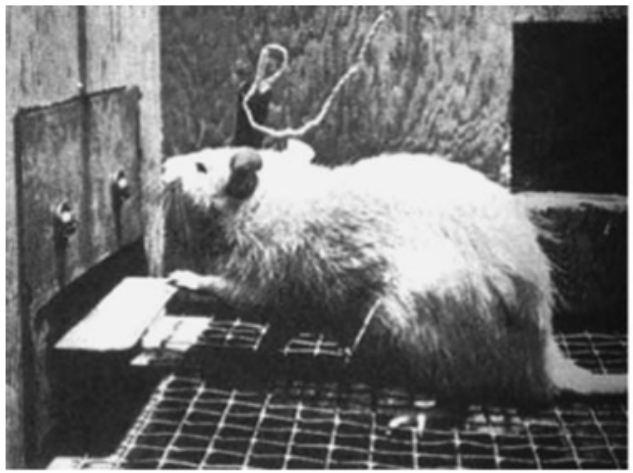

Figure 1: Headwired rat presses button for releasing reward to itself. [7]

- Self-stimulation experiments, J. Olds, P. Milner on rats, 1950s [8]
- Rats would continue to self-stimulate without rest
- Self-stimulation behaviour completely displaced all other interests
- What about humans, or AI-RL?

Value Learning (1)

- Proposed method for incorporating human values in an AI [2] [9]
- Assumption: humans' goals would not naturally occur in an artificial agent and should be enforced in it [3]
- Creation of an artificial learner whose actions consider many possible set of values and preferences, weighted by their likelihood
- Could prevent an AI of having goals detrimental to human values

Value Learning (2)

- U: observation-utility function, maps $yx_{\leq m}$ to scalar utility
- Uncertainty over utility functions: agent has many of these utility functions
- Assign probability to utility given interaction history: $P(U \mid yx_{\leq m})$
- Now possible, expected value over possible utility functions:

$$\sum_U U(yx_{\leq m})P(U \mid x_{\leq m}))$$

- Optimality notion [2]:

$$y_k = \arg\max_{y_k} \sum_{x_k y x_{k+1:m}} P_o(yx_{\leq m} \mid yx_{<k}y_k) \sum_U U(yx_{\leq m})P_u(U \mid yx_{\leq m})$$

Inverse Reinforcement Learning (IRL)

- Also: Apprenticeship Learning via Inverse Reinforcement Learning [11]
- Reconstruct reward function of some other agent by observing actions
- IRL has been proposed as a potential means of value learning (Russell, Dewey, and Tegmark 2015)
- Example application: teach an agent helicopter tricks [12]

Value Learning with Storytelling and IRL (1)

- "Using Stories to Teach Human Values to Artificial Agents", M. Riedl, B. Harrison (2015) [10]
- Extract sociocultural values from narratives and construct a value-aligned reward signal
- Example task: "Get medicine from pharmacy."

1. Create a graph by crowd-sourcing stories for this specific task
2. Translate plot graph into trajectory
3. Use trajectory for reward
   - Reward: perform action in environment which is successor to current node
   - Negative reward: if not successor of current node

Value Learning with Storytelling and IRL (2)

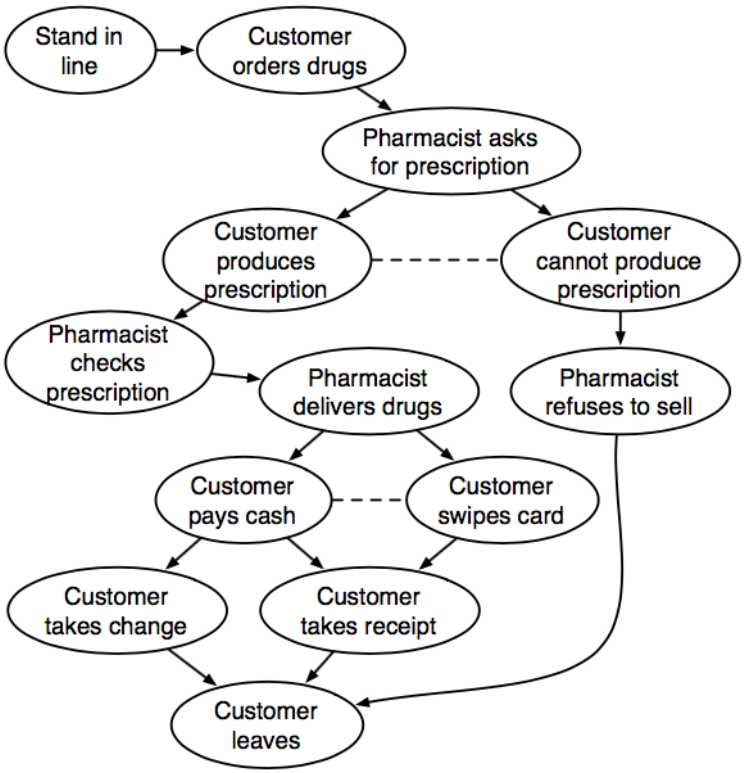

Figure 2: Plot graph for problem: getting medicine in pharmacy [10]

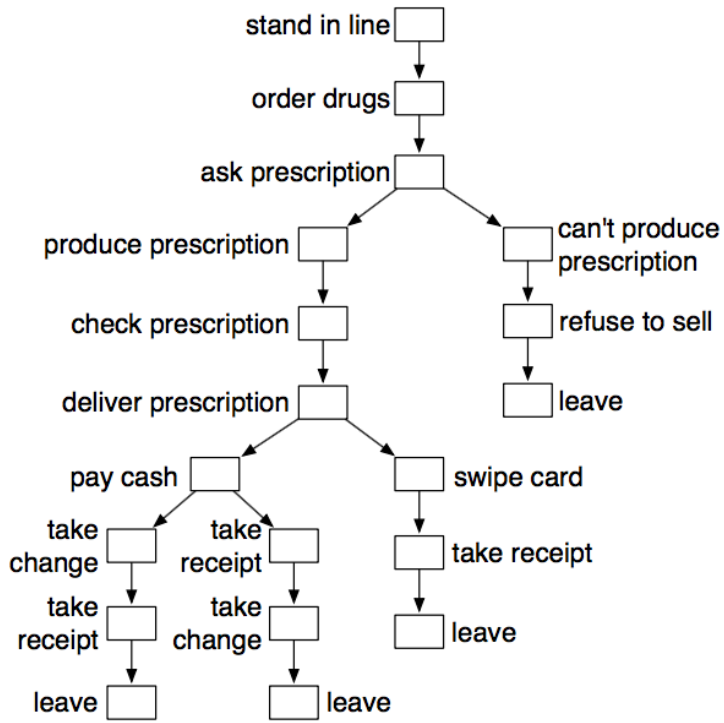

Figure 3: Trajectory of different plot graphs [10]

Value Learning with Storytelling and IRL (3)

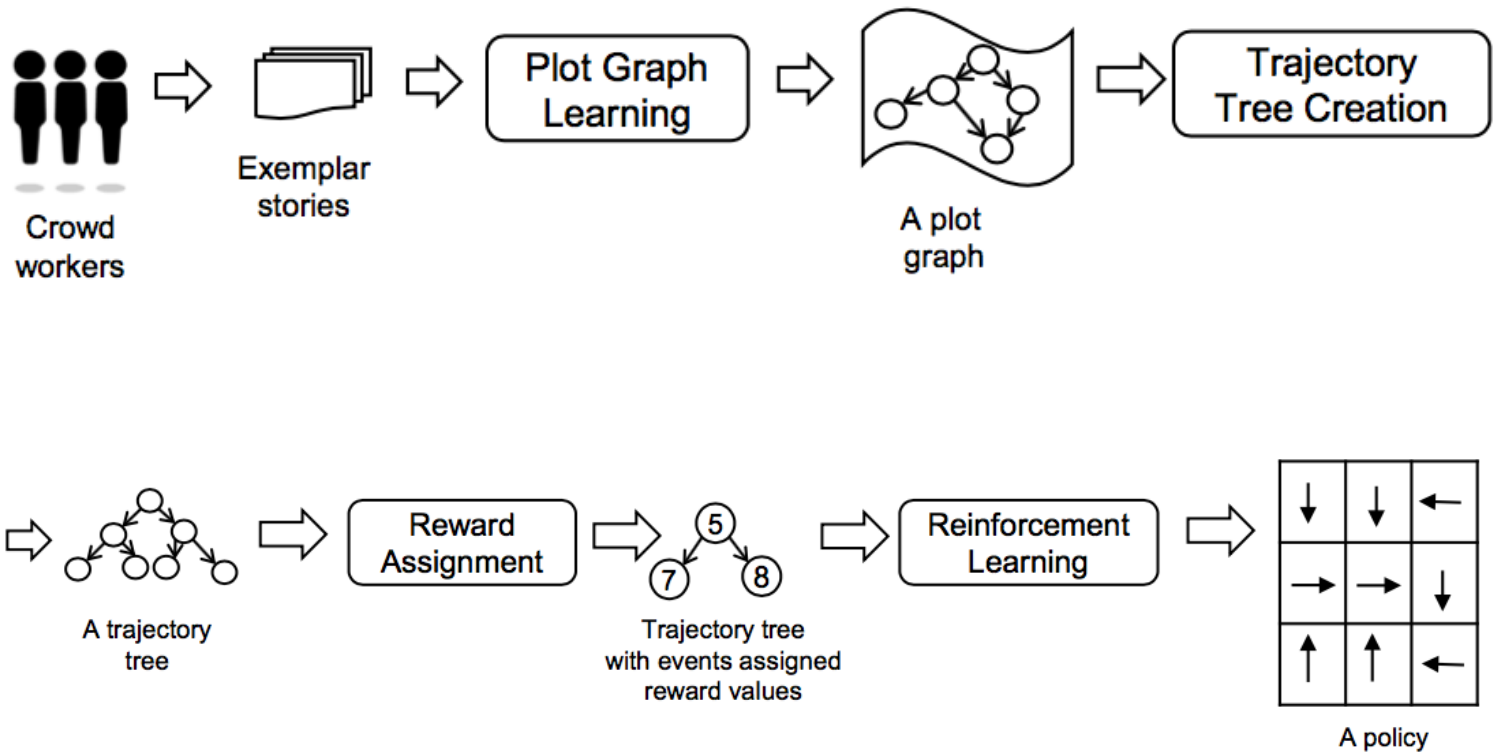

Figure 4: Summarized process for teaching an agent values with storytelling and IRL [10].

Hidden Complexity of Wishes

- So far: considered relatively easy task
- Open-Source Wish Project V1.1: "This wish's intent is to allow someone to live for as long as they want to" [14]

I wish to live in the locations of my choice, in a physically healthy, uninjured, and apparently normal version of my current body containing my current mental state, [...]

- Hostile wish-granter: *sure, but you won't learn anything anymore*
- How to know when task is too complex? after something bad happens? [2]
- Any attempts at "exact wording" written in natural language dominated by properties of mind

Fragility of Values

- Excerpt of list of terminal values (W. Frankena, 1973) [2]:
  - Life, consciousness, and activity; health and strength; pleasure and satisfactions of all or certain kinds; happiness, beautitude, ...
- Suppose one more values left out, e.g. digit of a phone number
- Does an agent that lacks a value have a net neutral impact on reality?
- Counter example: evolution as optimization process

A system that is optimizing a function of n variables, where the objective depends on a subset of size $k < n$, will often set the remaining unconstrained variables to extreme values; if one of those unconstrained variables is actually something we care about, the solution found may be highly undesirable. - Stuart Russel (2014) [4] [13]

Conclusion and Outlook

- Step forward in achieving AI that can pursue their own goals in a way that limits adverse effects
- Not possible to exclude all harm, but AI has been "encultured"
- Root of the problem: what is "friendly"?
- Too be friendly: empathy? the ability to understand and share the feelings of others
- Hypothetically: you are super intelligent, empathic AI (friendly wish-granter), how would you interpret "current mental state"?
- Best way maybe to make AI feel; become aware of itself?

References I

[1] Oxford Dictionary. https://en.oxforddictionaries.com.

[2] D. Dewey. Learning What to Value. AGI, August 2011.

[3] T. Everitt, M. Hutter. Avoiding Wireheading with Value Reinforcement Learning.

[4] S. Russell. Of myths and moonshine. Edge, 2014.

[5] M. Hutter. Universal Algorithmic Intelligence: A mathematical top-down approach. AGI, January 2007.

[6] M. Hutter et al. Reinforcement Learning via AIXI Approximation. July 2010.

[7] R. Yampolskiy. Utility function security in artificially intelligent agents. Journal of Experimental & Theoretical Artificial Intelligence, 2014.

References II

[8] J. Olds, P. Milner. Positive reinforcement produced by electrical stimulation of septal area and other regions of rat brain. Journal of Comparative & Physiological Psychology, 47, 419–427, 1954.

[9] E. Yudkowsky. Coherent Extrapolated Volition. Machine Learning Research Institute, 2004.

[10] M. Riedl, B. Harrison. Using Stories to Teach Human Values to Artificial Agents. Association for the Advancement of Artificial Intelligence, 2015.

[11] P. Abbeel, A. Ng. Apprenticeship Learning via Inverse Reinforcement Learning. International Conference on Machine Learning, 2004.

[12] Stanford University (A. Ng). Autonomous Helicopter. http://heli.stanford.edu, 10.7.2017.

References III

[13] N. Soares. The Value Learning Problem. Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence, July 2016.

[14] Open-Source-Wish-Project, https://www.reddit.com/r/RedditDayOf/comments/6c4hcn/the_opensource_wish_project_wish_for_immortality/, 1.7.17.