Theory-of-Mind as a Guide for Self-Organised AI Behaviour

Michael S. Harré

An interesting question artificial intelligence (AI) confronts us with is how it might produce psychological phenomena such as a Theory of Mind (AI-ToM) [Har22]. While an AI-ToM has obvious aspirational goals, such as improved AI-to-AI or AI-to-human coordination, there is a need for a clear framing of a specific question it answers in order for us to engineer a solution. One way to frame its purpose is to enforce "social constraints" that guide an AI's interactions with other agents while leaving it as free as possible to carry out tasks that are independent of other agents. This approach would allow a ToM to provide high-level guidance while the AI can self-organise its behaviour for other tasks. A well-established approach that satisfies this framework is MaxEnt but the difficulty lies in inferring the internal states of other agents from their behaviour, which is where ToM comes into play. In this work I cover a new approach called Inverse Reinforcement Learning for ToM and connect it to constraints imposed on agents by other agents' behaviour.

The core of the idea is that there are two levels of constraints: the social and the individual and we see the social constraints (requiring a ToM in order to understand them) as taking priority over the individual constraints, not dissimilar to Asimov's 3 laws of robotics in which protecting humans (or human society if the zeroth law: "A robot may not harm humanity, or, by inaction, allow humanity to come to harm" is included) takes priority over the interests of the individual robot, as encoded in the 3rd law which is sub-ordinate to all the others. If instead of the three (+1) laws being cast as prohibitions on actions they are recast as behavioural constraints that need to be (approximately) satisfied then the Lagrangian approach used in the MaxEnt derivation of probability distributions over choices lets us vary the Lagrangian parameters depending on which constraints are more important than others: High-valued parameters imply a high priority constraint (relative to other constraints) while low-valued parameters imply a low priority constraint. MaxEnt is a consistent way of combining constraints of different priorities in order to provide the maximum degree of freedom (flattest distribution) in which its can (self-)organise its behaviour. An intriguing consequence of social-MaxEnt solutions is that they have symmetry breaking solutions where the system self-organises around one of the multiple stable states [WHO+12].

The particular approach I will work from as an example is developed from conventional reinforcement learning (RL), but instead of starting with a formal model of reward-based learning we start with an agent's behaviour and from there infer the internal states of the agent that gave rise to the behaviours. The inverse reinforcement learning (IRL) problem [NR+00] can be stated as follows. We are given:

- Measurements of an agent's behaviour over time and under a number of different circumstances,
- The agent's sensory input,
- A model of the environment.

Task: Determine the reward function being used by the agent.
This quite general description is then made specific in terms of a Markov Decision Process $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{T}, r\}$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space of the agent, $\mathcal{T}$ is the transition model from one state to another, and $r(s)$ is the reward, a function of the state s. The optimal policy $\pi : \mathcal{S} \mapsto \mathcal{A}$ for the agent is that π^* which maximises the expected accumulation of reward: $V^\pi(\mathbf{s}) = E(r(s_1) + \gamma r(s_2) + \gamma^2 r(s_3) + ...|\pi)$ for a discount factor $\gamma \in [0, 1]$. There is also a set of behavioural demonstrations of how the agent has acted in a given state: $\mathcal{D} = \{\xi_1, \xi_2, \xi_3, ...\}$ where each ξ_i is a sequential path through the state-action space $\mathcal{A} \times \mathcal{S}$: $\xi_i = \{(a_1, s_1), (a_2, s_2), (a_3, s_3), ...\}$. The goal of IRL is then, given the set of demonstrations $\mathcal{D}$, recover the reward function $r : \mathcal{S} \mapsto \mathbb{R}$.

The first algorithms to solve this problem were introduced by Ng and Russell [NR+00] in 2000 where early results were reported and this was seen as an imitation problem where one agent was to replicate the behaviour of another agent. In their work they derived a solution that found an a^* that optimises the expected value of the value function $E_t(V^\pi(s)|a^*) \geq E_t(V^\pi(s)|a)\ \forall\ a\ \in\ \mathcal{A}$ where $E_t(\cdot)$ is the expectation with respect to the transition model $\mathcal{T}$. This solution has a lot in common with economic decision theory where the single choice, i.e. no probability over choices, is made that has the highest expected value. As Ng and Russell mention, this suffers from the difficulty of exploring all possible values, particularly in realistic high dimensional or continuous spaces.

The next significant milestone was the introduction of a MaxEnt by Ziebart et al [ZMB+08] in 2008. In this approach they circumvented the problem of optimising over an inequality by using the expected value of the reward as a constraint in a MaxEnt analysis. They used a reward function that maps the features of each state space s_j,

given by $\mathbf{f}_{s_j}$, and a reward weight vector θ to a numerical value: $U(\mathbf{f}_{s_j}) = \sum_{s_j \in \xi} \theta^T \mathbf{f}_{s_j} = \theta^T \mathbf{f}_\xi$. They were then able to show that the optimal path through $\mathcal{S} \times \mathcal{A}$ space is a MaxEnt probability distribution given by:

$$P(\xi_i|\theta) = \mathcal{Z}^{-1}(\theta) \exp(\theta^T \mathbf{f}_{\xi_i}) \tag{1}$$

$$= \mathcal{Z}^{-1}(\theta) \exp\left(\sum_{s_j \in \xi_i} \theta^T \mathbf{f}_{s_j} \right) \tag{2}$$

This also circumnavigates the issue of point estimates of decisions and provides the highest diversity of choices subject to the constraints, a desirable feature when there is uncertainty in which constraint needs to be strictly satisfied, as in the case of Ng and Russell's initial solution.

Finally, in work by Jara-Ettinger [JE19] in 2019 the notion of IRL as an approach to modelling a theory of mind was introduced. Here Jara makes the point that being able to understand the cognitive model that generates a sequence of behaviours has a close relationship with the psychological notion of ToM. A person has a ToM if they can infer the mental states of another person: what does the other person know, believe, what are their goals etc. [FF05]. So developing a model of the reward function of an agent that accurately explains their behaviour has much in common with a ToM. However, as Jara points out, deep learning AIs, one of the most successful approaches to RL to date, requires a prohibitively large amount of data: "... state-of-the-art IRL through deep learning [RPS+18] requires 32 million training examples to perform goal-inference at the capacity of a six-month-old infant [Woo98]. If humans acquired Theory of Mind in a similar way, infants would need to receive almost 175,000 labeled goal-training episodes per day, every day."

With the above in mind, to date there has not yet been an analysis of the ToM interpretation of IRL that uses the MaxEnt approach *in the context of other decision-makers*, in the same way such analysis is carried out in other fields such as game theory and social-neuroscience. With the background to IRL given above I will not focus on the algorithmic details of IRL but rather the consequences of IRL in the context of other decision-makers, and how the joint action spaces combined with MaxEnt approaches (see for example [WHO+12]) provides a novel approach to jointly constrained behaviour. The overall outcome will be to connect ideas from psychology, AI, and economics [Har21] with a AI-ToM implementation that takes advantage of MaxEnt methods.

References

[FF05] Chris Frith and Uta Frith. Theory of mind. *Current biology*, 15(17):R644–R645, 2005.

[Har21] Michael S Harré. Information theory for agents in artificial intelligence, psychology, and economics. *Entropy*, 23(3):310, 2021.

[Har22] Michael S Harré. What can game theory tell us about an AI 'Theory of Mind'? *Games*, 13(3):46, 2022.

[JE19] Julian Jara-Ettinger. Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, 29:105–110, 2019.

[NR+00] Andrew Y Ng, Stuart Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

[RPS+18] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[WHO+12] David H Wolpert, Michael Harré, Eckehard Olbrich, Nils Bertschinger, and Juergen Jost. Hysteresis effects of changing the parameters of noncooperative games. *Physical Review E*, 85(3):036102, 2012.

[Woo98] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998.

[ZMB+08] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, Anind K Dey, et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.