Knowing me, knowing you: On the relevance of a mind reading test for general testing of intelligence

Elpida S. Tzafestas [1]

Abstract. This short article presents a discussion of the relevance of mind reading tests to general testing of intelligence and an example of mind-reader behaviours for the IPD. It is discussed how a mind reading capacity may allow intricate emotional behaviours to emerge and how these relate to a broader developmental context.

1 INTRODUCTION

The human quest for the design of an artificial brother is almost as old as humanity. In ancient Greek mythology, lame god Hephaestus had created himself artificial slaves to help him in his smithy. Other stories and fantasies abound ever since. The quest to design an artifact as intelligent as ourselves or even more intelligent was formulated more recently and is the central goal of artificial intelligence. Alan Turing is considered a precursor of modern AI by having provided a relatively formal answer to the question “Can a machine think?” in the form of a certain “test” that the machine should pass in order to be considered intelligent [1]. The test consists in a verbal interaction with a human that should be considered as natural by a third party in the sense that the observer should not be able to distinguish between the human and the machine. This procedure, labeled “imitation game” by Turing, has been generalised to encompass *any* sort of problem and *any* sort of behaviour that can be considered as human by an external observer. Turing's original claim for adequacy of this test has been heavily criticised, reformulated and defended throughout the years (for a not too recent overview see [2]).

In almost all meaningful interactions between two humans, the full human potential for intentionality and consciousness is exhibited. For the purpose of the imitation game, in corresponding interactions between a human and a machine, some degree of machine intentionality or consciousness should be perceivable from the outside, by an external observer, otherwise the machine will be, somewhat fuzzily, considered as a “robot” rather than as a human replica. One such social feature characteristic of human nature and not of other primates or lower animals is mind reading, i.e. the capacity to “read” another person's mind and understand its intentions.

Mind reading is generally implicit baggage for any social activity and corresponding deficits to correct mind reading will lead to what will be externally perceived as lower social intelligence (see for example discussions on autistic intelligence, [3]). Because mind reading is considered unique to humans, it is also associated, although not necessarily causally, with other human monopolies, namely with intentionality, human-level imitation, language, empathy and more (see several chapters in [4][5]).

In the next section, we explain more in depth the connection between mind reading and intelligence and between testing for mind reading and testing for intelligence while in section 3 we give an example on the well known prisoner's dilemma. Section 4 discusses the implications of this approach and concludes.

2 FROM TURING TO MIND READING

The connection between mind reading and a test for intelligence is not as far fetched as one might initially think. Turing himself, in the original formulation of his imitation game [6, last section] wrote:

> *“The extent to which we regard something as behaving in an intelligent manner is determined as much by our own state of mind and training as by the properties of the object under consideration. If we are able to explain and predict its behaviour or if there seems to be little underlying plan, we have little temptation to imagine intelligence. With the same object therefore, it is possible that one man would consider it as intelligent and another would not; the second man would have found out the rules of its behaviour [...]”*

Not to be underestimated is the fact that Turing gave to this short last section the title *“Intelligence as an emotional concept”*. One can speculate that for a human observer of another human or machine acting, attribution of a degree of intelligence to it may have a number of emotional consequences, such as admiration, envy, anger with oneself, perseverence to own effort etc. From our everyday experience we know that such feelings lead almost unavoidably to alteration of one's behaviour, especially in cases where some competition is involved, implicitly or explicitly. For example, we may imagine a school child realizing that its fellow pupil solves

[1] Cognitive Science Laboratory, Department of Philosophy and History of Science, Univ. of Athens, Ano Ilisia 15771, Athens, GREECE. Email: **etzafestas@phs.uoa.gr**.

difficult problems in arithmetic that he cannot solve. Apart from the birth of feelings such as the above, the school child may eventually devise ways to achieve the same performance by either making friends with the fellow or by plainly stealing his work. In the latter case, and provided that the child is not caught, an external observer will attribute to it the same level of mathematical intelligence as its fellow, based solely on performance.

A human interacting with a machine, for example in the context of a two-party game, will impulsively make use of such mind reading abilities, that generally give him an advantage against the machine. Furthermore, the machine may not be externally regarded as intelligent, if one cannot attribute to it *any* notion of intentionality, which involves in most cases a degree of other or self understanding. On the contrary, a human observer acquiring the feeling that he is being watched and partly understood by the machine will be ready to assign intelligence to the machine and probably become frightened and tempted to quit the interaction.

We depart from Meltzoff's “like-me hypothesis” [7] that connects imitation and mind reading by assuming that when infants see others acting similarly to how they have acted in the past, they project onto others the mental state that regularly goes with that behaviour. However, we do not claim that people ordinarily *think* that others are like them, rather we adopt the more modest view that what a machine can actually do is discover whether an observed entity is “like it” in the sense that it acts in the same way. To paraphrase Metzoff ([7], p. 75) intelligent artifacts may use their own intentional actions as a framework for interpreting the intentional actions of others.

An artificial agent endowed with some minimal mind reading possibility of this kind has the potential of being perceived by a human observer (or adversary in a two-party interaction) as a human being, thus it has the potential to pass some version of the Turing test. The usual objections still apply, for example that the Turing test is too easy, and that there is always one way to pass the test without any “intelligence” [2] –so maybe we found just another way to do so -, but Harnad [8] sustains that all mind-reading is Turing test passing. The middle view that is easier to accept, is that mind reading is a recommended, if not required, component of a Turing test passer.

3 AN EXAMPLE

We have implemented a mind reader version of Iterated Prisoner's Dilemma (IPD) players. IPD is often used as a benchmark for the study of cooperative and altruistic behaviour. In general, the cooperation problem between two (or more) agents states that each agent has a strong personal incentive to defect, while the joint best behaviour would be to cooperate. This problem is traditionally modeled as a special two-party game, the Iterated Prisoner’s Dilemma (IPD).

At each cycle of a long interaction process, the agents play the Prisoner’s Dilemma. Each of the two may either cooperate (C) or defect (D) and is assigned a payoff defined by table 1.

AGENT	OPPONENT	PAYOFF
C	C	3 (= Reward)
C	D	0 (= Sucker)
D	C	5 (= Temptation)
D	D	1 (= Punishment)

Table 1. Iterated Prisoner's Dilemma score matrix

The first notable behaviour for the IPD designed and studied by Axelrod [9] is the Tit For Tat behaviour (TFT, in short):

- Start by cooperating,
- From there on return the opponent’s previous move

This behaviour has achieved the highest scores in early tournaments and has been found to be fairly stable in ecological settings. TFT demonstrates three important properties, shared by most high scoring behaviours in IPD experiments.

- It is good (it starts by cooperating)
- It is retaliating (it returns the opponent’s defection)
- It is generous (it forgets the past if the defecting opponent cooperates again).

In the literature we may also find stochastic strategies [10], studies in a purely evolutionary perspective ([11]), theoretical or applied biological studies ([12]) and studies of modified IPD versions ([13]). Noise is introduced as either mis-perception (a move may occasionally be perceived as the opposite, *COOPERATE* instead of *DEFECT* or vice versa) or, more often, mis-implementation (a move may occasionally be switched from *COOPERATE* to *DEFECT* or vice versa). It has been shown that retaliating strategies such as TFT can score quite badly in the presence of noise, despite their superiority in the non-noisy domain [14][15]. This happens because even accidental defections may lead to a persistent series of mutual defections by both players, thus breaking cooperation. The usual approach is to introduce some degree of explicit generosity to account for opponent’s misbehaviours or to attempt opponent modeling.

In earlier work of ours we have shown how a fundamental TFT-like behaviour (called Adaptive TFT) with the possibility to adapt itself to its opponent’s friendliness may achieve very high scores against all kinds of strategies, including suspicious, random and periodic behaviours [16]. We have also shown [17] how an additional mechanism of attraction may induce highly cooperative behaviour even if one of the agents is normally spiteful or in the presence of noise. The difficulty of tackling an arbitrary opponent of unknown behaviour, especially in the presence of noise, is to understand whether perceived defections of his are intentional or a result of mis-perception/mis-implementation or inertia. As a simple example, two Suspicious TFT (STFT) agents, that are TFTs that initially defect, are unable to converge to mutual cooperation. An Adaptive TFT agent can solve this problem against STFT but not against another Adaptive TFT which is defective at the moment due to prolonged unhappy interactions.

A mind reader version of an arbitrary behaviour for the IPD is simply a behaviour that continuously examines whether its opponent uses the same behavioural model as itself. This is implemented as follows:

```
t = actual time (IPD game round);
for i=0 to min(t-1,w)
{
     theT = t-i-1;
     if (simHist[i] == oppHist[theT]))
            opp_like_me ++;
}
if (opp_like_me >= T) COOPERATE;
else generate_behavior(SELF);
// Look ahead
int sim_move = generate_behavior(MIRROR);
pushSimHist(sim_move);

w = mind reading window
T = mind reading threshold
simHist = simulated opponent's history
(Array of w elements)
oppHist = actual opponent's history
(Array of w elements)
```

Simply stated, this behaviour examines whether within a fixed backward looking window the opponent does what the agent would have done in its place, by simulating a copy of itself against its actual self. The function `generate_behaviour` does what the agent normally does (for example it is a TFT behaviour) and takes an argument that shows whether the inputs correspond to what the agent sees (SELF) or to what its opponent sees (MIRROR).[2]

This simple mind reading facility has led retaliating, even suspicious STFT-like behaviours, to converge to mutual cooperation with other agents of the same kind. Furthermore, for low values of noise (up to 10%) it has practically solved the mis-perception/mis-interpretation problem.

The condition `(opp_like_me >= T)` may be translated as `(opponent is like me)`. It is straightforward to think that what looks as a reasonable behaviour

```
if (like me) cooperate
else play as usually (reason)
```

may not be and is usually not the case. The cooperation problem may thus be defined at a meta level as:

```
if (like me) do something
else do something else
```

For example, it is not uncommon to meet people who are more cooperative with unsimilar ones (as a precaution) than with similar ones (which might be a selfish reaction). Another often encountered feature is of lower attentiveness in case of interaction with similar ones, so that signs of defection or cheating may go unnoticed for a long period.

Now imagine a human playing the IPD against our artificial player. He will be surprised to find that the machine player occasionally behaves generously, but that if the human persists in defecting, the artificial player will revert to continuous retaliating, as if it could bypass mistakes and misundestandings but perceive consistently vicious behavior and punish accordingly, just as we humans do. The minimal mind-reading ability is therefore believed to add to an artificial player's potential to pass the Turing test. Moreover, if this ability is combined with meta-strategies as described above, then a rich repertoire of behaviours may be produced, further convincing a human interacting with variants of the player, that he plays against a team of other human-like beings with different characteristics, personalities and temperaments.

[2] The situation is slightly more complicated, because an agent cannot know what exactly its opponent sees but can only judge based on what he thinks its opponent sees. Generally, it is safe to assume that the opponent sees the agent's actual move, altough if misperception or misimplementation are allowed the opponent will occasionally see the opposite.

4 CONCLUSION

Testing for mind reading to test for intelligence may have a number of assorted consequences. First, the opponent simulation part of the previous section may not be available at birth. Indeed, infants have been found to somehow develop “like-me” behaviour, so normally we should endow our agents with a limited simulation possibility that is enriched in the process. Because, normally, an intelligent artifact continuously develops its own behaviour further, it makes sense to allow the mind-reading behaviour to try to copy regular behaviour, always staying behind in complexity and performance. Second, partly due to the previous reason, mind reading is expected to be less developed than its regular counterpart. Thus externally perceived atomic intelligence may differ substantially from externally perceived social intelligence. Finally, differentially organized mind reading apparatus may be externally perceived as defective, thus giving room to the design of artificially defective agents.

In the same vein and inspired from the original Turing test, mind reading considerations bring us to consider human-machine interaction where the human will assign intelligence to the machine, but, say, childish intelligence or schizophrenic intelligence (for an old account of the latter idea see [18]). In sum, a mind reading test for intelligence allows a broadening of the scope of intelligence tests, so as to also encompass developmentally immature, defective or perceptibly distorted intelligence.

REFERENCES

[1] A. Turing, “Computing machinery and intelligence”, *Mind*, 39:433-460, 1950.

[2] A.P. Saygin, I. Cicekli, V. Akman, “Turing Test: 50 years later”, *Minds and Machines*, 10:463-518, 2000.

[3] A. Klin, W. Jones, R. Schultz, F. Volkmar, “The enactive mind, or from actions to cognition: Lessons from autism”, in U. Frith and E. Hill, Eds., “Autism: Mind and brain”, Oxford University Press, 2003.

[4] S. Hurley, N. Chater, Eds., *Perspectives on imitation: From neuroscience to social science*, MIT Press, 2005.

[5] K.R. Stueber, *Rediscovering empathy: Agency, fold psychology and the human sciences*, MIT/Bradford Books, 2006.

[6] A. Turing, “Intelligent machinery”, National Physical Laboratory Report, 1948.

[7] A.N. Meltzoff, “Imitation and other minds: The 'Like Me' Hypothesis”, in S. Hurley, N. Chater, Eds., *Perspectives on imitation: From neuroscience to social science*, MIT Press, 2005.

[8] S. Harnad, Can a machine be conscious? How?, *Journal of Consciousness Studies*, 2003.

[9] R. Axelrod, *The evolution of cooperation*. Basic Books, 1984.

[10] M. A. Nowak and K. Sigmund, “Tit-for-tat in heterogeneous populations”, *Nature*, Vol. 355, pp. 250-53, 1992.

[11] D. Fogel, “Evolving behaviors in the iterated prisoner’s dilemma”, *Evolutionary Computation*, Vol. 1, pp. 77-97, 1987.

[12] M. W. Feldman and E. A. C. Thomas, “Behavior-dependent contexts for repeated plays of the prisoner’s dilemma II: Dynamical aspects of the evolution of cooperation”, *Journal of Theoretical Biology*, Vol 128, pp. 297-315, 1987.

[13] E. A. Stanley, D. Ashlock and L. Tesfatsion, “Iterated prisoner’s dilemma with choice and refusal of partners”, in *Artificial Life III*, Addison-Wesley, 1994.

[14] D. Kraines and V. Kraines, “Evolution of learning among Pavlov strategies in a competitive environment with noise”, *Journal of Conflict Resolution*, Vol. 39, Issue 3, pp. 439-466, 1995.

[15] P. Molander, “The optimal level of generosity in a selfish, uncertain environment”, *Journal of Conflict Resolution*, Vol. 31, Issue 4, pp. 692-724, 1987.

[16] E. Tzafestas, “Toward adaptive cooperative behavior”, Proceedings of the Simulation of Adaptive Behavior Conference, Paris, September 2000.

[17] E. Tzafestas, “Attraction and cooperation in space”, Proceedings 2007 Conference on Evolutionary Computation.

[18] K.M. Colby, F.D. Hilf, S. Weber, H. Kraemer, “Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes”, *Artificial Intelligence*, 3:199-221, 1972..