Instrumental Properties of Social Testbeds

Javier Insa-Cabrera and José Hernández-Orallo

DSIC, Universitat Politècnica de València, Valencia, Spain
jinsa@dsic.upv.es jorallo@dsic.upv.es

Abstract. The evaluation of an ability or skill happens in some kind of testbed, and so does with social intelligence. Of course, not all testbeds are suitable for this matter. But, how can we be sure of their appropriateness? In this paper we identify the components that should be considered in order to measure social intelligence, and provide some instrumental properties in order to assess the suitability of a testbed.

Keywords: Social Intelligence · Multi-Agent Systems · Cooperation · Competition · Game Theory · Rewards · Universal Psychometrics.

1 Introduction

Evaluation tools are crucial in any discipline as a way to assess its progress and creations. There are some tools, benchmarks and contests, aimed at the measurement of humanoid intelligence or the performance in a particular set of tasks. However, the state of the art of artificial intelligence (AI) and artificial general intelligence is now more focussed towards social abilities, and here the measuring tools are still rather incipient. In the past two decades, the notion of agent and the area of multi-agent systems have shifted AI to problems and solutions where 'social' intelligence is more relevant (e.g., [1, 2]). This shift towards a more social-oriented AI is related to the modern view of human intelligence as highly social, actually one of the most distinctive features of human intelligence over other kinds of animal intelligence. Some significant questions that appear here are then whether we are able to properly evaluate social intelligence in general (not only in AI, but universally) and whether we can develop measurement tools that distinguish between social intelligence and general intelligence.

In this paper, we 1) identify the components that should be considered in order to assess social intelligence, and 2) provide some instrumental properties to help us determine the suitability of a testbed to be used as a social test (validity, reliability, efficiency, boundedness and team symmetry), while analyzing the influence that such components have on these properties. This helps us to pave the way for the analysis of whether many social environments, games and tests in the literature are useful for measuring social intelligence.

The paper is organized as follows. Section 2 provides the necessary background. Section 3 identifies the components that we should consider in order to measure social intelligence. Section 4 presents some instrumental properties to assess the suitability of a testbed to be used as a social intelligence test. Finally, Sect. 5 closes the paper with some discussion and future work.

2 Background

This section gives an introduction to the concepts and terminology of multi-agent environments and serves as a background for the following sections.

2.1 Multi-agent Environments

An environment is a world where an agent can interact through observations, actions and rewards. This general view of the interaction between an agent and an environment can be extended to various agents by letting them interact simultaneously with the environment.

A multi-agent environment is an interactive scenario with several agents. An environment accepting n agents defines n parameters (one for each agent) denoted as *agent slots*. We use $i = 1, \ldots, n$ to denote the slots. Each simultaneous interaction of the n agents is called a *time step*, where the order of events is always: observations, actions and rewards. $\mathcal{O}_i$ is the observation set that the agent in slot i can perceive, $\mathcal{A}_i$ is the action set that the agent in slot i can perform and $\mathcal{R}_i \subseteq \mathbb{Q}$ represents the possible rewards obtained by the agent in slot i. For each step k, the agent in slot i must perceive an observation $o_{i,k} \in \mathcal{O}_i$, perform an action $a_{i,k} \in \mathcal{A}_i$ and obtain a reward $r_{i,k} \in \mathcal{R}_i$. We use o_k, a_k and r_k respectively to denote the joint observation, joint action and joint reward profiles of the n agents at step k (i.e., $o_k = (o_{1,k}, \ldots, o_{n,k}) \in \mathcal{O}_1 \times \cdots \times \mathcal{O}_n$ represents the joint observation profile at step k, and similarly for actions and rewards). For example, a sequence of two steps in a multi-agent environment is then a string such as $o_1 a_1 r_1 o_2 a_2 r_2$ and the string $o_{1,1} a_{1,1} r_{1,1} o_{1,2} a_{1,2} r_{1,2}$ denotes the sequence of observations, actions and rewards for the agent in slot 1.

Both the agents and the environment are defined as probabilistic measures. At step k, the term $\pi(a_{i,k} | o_{i,1} a_{i,1} r_{i,1} \ldots o_{i,k}) \rightarrow [0,1]$ denotes the probability of the agent in slot i to perform action $a_{i,k}$ after the sequence of events $o_{i,1} a_{i,1} r_{i,1} \ldots o_{i,k}$. The observation provided by the environment at step k to the agent in slot i also has a probabilistic measure $\omega(o_{i,k} | o_1 a_1 r_1 \ldots o_{k-1} a_{k-1} r_{k-1}) \rightarrow [0,1]$. As with the observation, the reward at step k to the agent in slot i is provided depending on observations, actions and rewards on previous steps $\rho(r_{i,k} | o_1 a_1 r_1 \ldots o_k a_k) \rightarrow [0,1]$. Note that the rewards obtained by each agent depend on the joint observations, actions and rewards of all the agents interacting in the environment, and not only on their own.

2.2 Teams

It is important to determine the *roles* that agents take in the environment. The key issue is to establish whether the other agents goals and interests are compatible with one's goals. The concept is complex, as alliances can be created and broken even if no clear teams are established from the beginning (and this is an interesting property of social intelligence). These roles or alliances determine two major social behaviors: cooperation and competition.

We need to decide how the environment distributes rewards among the agents. An easy possibility would be to make each agent get its rewards without further constraints over other agents' rewards. With this configuration (e.g., general-sum games), both competition and cooperation may be completely useless for most environments, as the rewards are not limited or linked to the other agents. In contrast, if we set that the total set of rewards is limited in some way, we will foster competition, as happens in zero-sum games. But in any of these two cases cooperation will hardly take place. Alliances could arise sporadically between at least two agents in order to bother (or defend against) a third agent, but we need a way to make it more likely before any (sophisticated) alliance can emerge on its own. One possible answer is the use of teams, defined as follows:

Definition 1. *Agent slots i and j are in the same team iff $\forall k : r_{i,k} = r_{j,k}$, whatever the agents present in the environment.*

which means that all agents in a team receive exactly the same rewards. Note that teams are not alliances as usually understood in game theory. In fact, teams are fixed and cannot be changed by the agents. Also, we do not use any sophisticated mechanism to award rewards, related to the contribution of each agent in the team, as it is done with the Shapley Value [3]. We just set rewards uniformly.

2.3 Multi-agent Environments Using Teams

At this moment, we are ready to define an environment with parametrized agents by only specifying their slots and their team arrangement.

Definition 2. *A multi-agent environment μ accepting $N(\mu)$ agents (i.e., the number of slots in μ) is a tuple $\langle \mathcal{O}, \mathcal{A}, \mathcal{R}, \omega, \rho, \tau \rangle$, where $\mathcal{O}$, $\mathcal{A}$, $\mathcal{R}$ represent the observation sets, action sets and reward sets respectively (i.e., $\mathcal{O} = (\mathcal{O}_1, \ldots, \mathcal{O}_{N(\mu)})$, $\mathcal{A} = (\mathcal{A}_1, \ldots, \mathcal{A}_{N(\mu)})$ and $\mathcal{R} = (\mathcal{R}_1, \ldots, \mathcal{R}_{N(\mu)})$) and ω and ρ are the observation function and reward function respectively as defined in Sect. 2.1. τ is a partition on the set of slots $\{1, \ldots, N(\mu)\}$, where each set in τ represents a team.*

Note that with this definition the agents are not included in the environment. For instance, noughts and crosses could be defined as an environment μ_{nc} with two agents, where the partition set τ is defined as $\{\{1\}, \{2\}\}$, which represents that this game allows two teams, and one agent in each. Another example is RoboCup Soccer, whose τ would be $\{\{1, 2, 3, 4, 5\}, \{6, 7, 8, 9, 10\}\}$.

We now define an *instantiation* for a particular agent setup. Formally, an *agent line-up* l is a list of agents. For instance, if we have a set of agents $\Pi = \{\pi_1, \pi_2, \pi_3, \pi_4\}$, a line-up from this set could be $l_1 = (\pi_2, \pi_3)$. The use of the same agent twice is allowed, so $l_2 = (\pi_1, \pi_1)$ is also a line-up. We denote by $\mu[l]$ the instantiation of an environment μ with a line-up l, provided that the length of l is greater than or equal to the number of agents allowed by μ (if l has more agents, the excess is ignored). The slots of the environment are then matched with the corresponding elements of l following their order. For instance, for the noughts and crosses, an instantiation would be $\mu_{nc}[l_1]$. Note that different instantiations over the same environment would normally lead to different results. We use $L^n(\varPi)$ to specify the set of all the line-ups of length n with agents of $\varPi$.

We use the notation $R_i^K(\mu[l])$, which gives us the expected result of the ith agent in line-up l for environment μ (also in slot i) during K steps. If K is omitted, we assume $K = \infty$. In order to calculate an agent's result we make use of some kind of utility function (e.g., an average of rewards).

3 Components to Consider While Evaluating Social Intelligence

The components that we consider to measure social intelligence are:

- **Set of multi-agent environments M:** The environments we use to perform the evaluation.
- **Set of agents $\varPi$:** The agents that conform the line-ups.
- **Weights:** We give weights (non-negative numbers) to the environments, their slots and the line-ups. $w_M(\mu)$ denotes a weight to environment μ from a certain set M, $w_S(i,\mu)$ denotes a weight to slot i of a certain environment μ and $w_L(l)$ denotes a weight to a line-up l formed with agents from a certain set $\varPi$, giving weights to the agents in the line-up and their positions.
- **Definition of social intelligence $\varUpsilon$:** The actual definition that measures social intelligence. This definition should use sets M and $\varPi$ and weights over them, i.e., w_M, w_S and w_L, in some way to measure the social intelligence. As an example, we use the definition of social intelligence from [4, Sect. 3.3]:

$$\varUpsilon(\varPi, w_L, M, w_M, w_S) \triangleq \sum_{\mu\in M} w_M(\mu) \sum_{i=1}^{N(\mu)} w_S(i,\mu) \sum_{l\in L^{N(\mu)}(\varPi)} w_L(l) R_i(\mu[l]) \ . \tag{1}$$

- **Test of social intelligence $\hat{\varUpsilon}$:** The final test to measure social intelligence following a definition of social intelligence $\varUpsilon$. The test should consist of a set of exercises and some kind of procedure to sample them. As an example, we use the definition of social intelligence test from [4, Sect. 3.4]:

$$\hat{\varUpsilon}[p_\varPi, p_M, p_S, p_K, n_E](\varPi, w_L, M, w_M, w_S) \triangleq \eta_{\mathcal{E}} \sum_{\langle\mu,i,l\rangle\in\mathcal{E}} w_M(\mu) w_S(i,\mu) w_L(l) R_i^K(\mu[l]) \ . \tag{2}$$

where $\eta_{\mathcal{E}}$ normalizes the formula with $\eta_{\mathcal{E}} = \frac{1}{\sum_{\langle\mu,i,l\rangle\in\mathcal{E}} w_M(\mu) w_S(i,\mu) w_L(l)}$, K is chosen using probability distribution p_K and the exercises $\mathcal{E}$ are sampled as:

$$\mathcal{E} \sim^{n_E} \left[\bigcup_{\mu\in M} \bigcup_{i=1}^{N(\mu)} \left\{ \langle\mu,i,l\rangle : l \in L^{N(\mu)}(\varPi) \right\} \right]_{p_{\mathcal{E}}} \ .$$

with $S \sim^n [A]_p$ being a sample S of n elements from set A using probability distribution p, and $p_{\mathcal{E}}$ being a distribution on the set of triplets $\langle\mu,i,l\rangle$ based on p_M, p_S and $p_\varPi$.

4 Properties

In order to evaluate social intelligence and distinguish it from general intelligence, we need tests where social ability has to be used and, also, where we can perceive its consequences. This means that not every environment is useful for measuring social intelligence and not every subset of agents is also useful. We want tests such that agents must use their social intelligence to understand and/or have influence over other agents' policies in such a way that this is useful to accomplish their goals, but common general intelligence is not enough.

Hereafter, we investigate some instrumental properties for a testbed of multi-agent environments and agents to measure social intelligence.

4.1 Validity

Validity is the most important property of a cognitive test in psychometrics. In our context, the validity of a definition is that it accounts for the notion we expect it to grasp. For instance, if we say that a given definition of Υ measures social intelligence but it actually measures arithmetic abilities then the definition is not valid. Ultimately, this depends on the choice of Π and M in Υ, such as e.g., (1).

Poor validity may have two sources (or may appear in two different variants): a definition may be too specific (it does not account for all the abilities the notion is thought to consider) or it is too general (it includes some abilities that are not part of the notion to be measured). In other words, the measure should account for *all, but not more,* of the concept it tries to represent. We refer to these two issues of validity as the generality and the specificity of the measure. While validity is not usually seen as an instrumental property, we have to say that the choices of Π and M may both have generality and specificity, which eventually can compensate, but could lead to a test that is not very effective. That means that we should try to find proper choices such that they fit the concept we want to measure precisely.

Regarding generality, we should be careful about the use of very restrictive choices for Π and M. It could be possible to find a single environment that looks ideal to evaluate social intelligence. However, using just one environment is prone to specialisation, as usual in many AI benchmarks. For instance, if we use a particular maze, then we can have good scores by evaluating a very specialized agent for this situation, which may be unable to succeed in other mazes or problems. For instance, chess with current chess players is an example where a specialized system (e.g., Deep Blue) is able to score well, while it is clearly useless for other problems. A similar over-specialisation may happen if the agent class is too small. This is usual in biology, where some species specialize for predating (or establishing a symbiosis) with other species. Consequently, the environment class and the agent class must be general enough to avoid that some predefined or hardwired policies could be optimal for these classes. This is the key issue of a (social) *intelligence* test; it must be as general as possible. We need to choose a diverse environment class. One possibility is to consider all environments (as done by [5, 6]), and another is to find an environment class that is sufficiently representative (as attempted in [7]).

Similarly, we need to consider a class of agents that leads to a diversity in line-up. This class should incorporate many different types of agents: random agents, agents with some predetermined policies, agents that are able to learn, human agents, agents with low social intelligence, agents with high social intelligence, etc. The set of all possible agents (either artificial or biological) is known as *machine kingdom* in [8] and raises many questions about the feasibility of any test considering this astronomically large set. Also, there are doubts about what the weight for this universal set should be when including them into line-ups (i.e., w_L). Instead, some representative kinds of agents could be chosen. In this way, we could aim at social intelligence relative to a smaller (and well-defined) set of agents, possibly specializing the definition by limiting the resources, the program size or the intelligence of the agents.

Regarding specificity, it is equally important for a measurement to only include those environments and agents that really reflect what we want to measure. For instance, it is desirable that the evaluation of an ability is done in an environment where no other abilities are required, or in other words, we want that the environment evaluates the ability in isolation. Otherwise, it will not be clear which part of the result comes from the ability to be evaluated, and which part comes from other abilities. Although it is very difficult to avoid any contamination, the idea is to ensure that the role of these other abilities are minor, or are taken for granted for all agents. We are certainly not interested in non-social environments as this would contaminate the measure with other abilities. In fact, one of the recurrent issues in defining and measuring social intelligence is to be specific enough to distinguish it from general intelligence.

4.2 Reliability

Another key issue in psychometric tests is the notion of reliability, which means that the measurement is close to the actual value. Note that this is different to validity, which refers about the true identification or definition of the actual value. In other words, if we assume validity, i.e., that the definition is correct, reliability refers to the quality of the measurement with respect to the actual value. More technically, if the actual value of π for an ability ϕ is v then we want a test to give a value which is close to v. The cause of the divergence may be systematic (bias), non-systematic (variance) or both.

First, we need to realize that reliability applies to tests, such as e.g., (2). Reliability is then defined by considering that a test can be repeated many times, so becoming a random variable that we can compare to the true value. Formally:

Definition 3. *Given a definition of a cognitive ability Υ and a test over it $\hat{\Upsilon}$, the test error is given by:*

$$TE(\hat{\Upsilon}) \triangleq Mean((\hat{\Upsilon} - \Upsilon)^2) \ . \tag{3}$$

where the mean is calculated over the repeated application of the test (to one subject or more subjects).

The reason for defining test error as the mean *squared* error (and not an absolute error) is a customary choice in many measures of error, as we can decompose it into the squared bias $(Mean(\hat{\Upsilon}) - \Upsilon)^2$ and the variance of the error $Var(\hat{\Upsilon} - \Upsilon)$. If the bias is not zero this means that the procedure to sample the exercises and/or the number of steps is inappropriate. If there is a high variance, this suggests that the number of exercises is too small, or that the exercises run for a very short time.

The reliability $Rel(\hat{\Upsilon})$ can be defined as a decreasing function over $TE(\hat{\Upsilon})$, such as $Rel(\hat{\Upsilon}) = e^{-TE(\hat{\Upsilon})}$. The estimation of $TE(\hat{\Upsilon})$ or $Rel(\hat{\Upsilon})$ depends on knowing the true value of Υ. This is not possible in practice for most environments, so Υ will need to be estimated for large samples and compared with an actual test (working with a small sample).

4.3 Efficiency

This property refers to how efficient a test is in terms of the (computational) time required to get a reliable score. It is easy to see that efficiency and reliability are opposed. If we were able to perform an infinitely number of infinite exercises, then we would have $\hat{\Upsilon} = \Upsilon$, with perfect reliability, as we would exhaust Π and M. If done properly, it is usually the variance component of the reliability decomposition that is affected if we keep the bias close to 0 even with very low values for the number of exercises.

Efficiency can be defined as a ratio between the reliability and the time taken by the test.

Definition 4. *Given a definition of a cognitive ability Υ and a test over it $\hat{\Upsilon}$, the efficiency is given by:*

$$Eff(\hat{\Upsilon}) \triangleq Rel(\hat{\Upsilon})/Time(\hat{\Upsilon}) \ . \tag{4}$$

where $Time$ is the average time taken by test $\hat{\Upsilon}$. $Time$ can be measured as physical (real) time or as computational time (steps).

The issue is how to choose environments and agents such that a high efficiency is attained. Clearly, if the selected environments are insensitive to agents' actions or require too many actions to affect rewards, then this will negatively affect efficiency. As we are interested in social abilities, interactivity and non-neutralism between agents' rewards must be high, as otherwise most steps will be useless to get information about the agent to evaluate. This of course includes cases where the agents are stuck or bored because their opponents (or teammates) are too good or too bad, or the environment leads the agents to heaven or hell situations where actions are almost irrelevant. A way of making tests more efficient is by the use of adaptive tests [6], [8].

4.4 Boundedness

One desirable property is that rewards are bounded, otherwise the value of Υ (such as e.g., (1)) could diverge. Any arbitrary choice of upper and lower bounds can be scaled to any other choice so, without loss of generality, we can assume that all of them are bounded between -1 and 1, i.e., $\forall i, k : -1 \leq r_{i,k} \leq 1$. Note that they are bounded for every step. So, if we use a bounded function to calculate the agent's result, then $R_i^K(\cdot)$ is also bounded.

However, bounded expected results do not ensure that Υ is bounded. In order to ensure a bounded measurement of social intelligence, we also need to consider that weights are bounded, i.e., there are constants c_M, c_S and c_L such that:

$$\forall M : \sum_{\mu \in M} w_M(\mu) = c_M \quad . \tag{5}$$

$$\forall \mu : \sum_{i=1}^{N(\mu)} w_S(i, \mu) = c_S \quad . \tag{6}$$

$$\forall \mu, \Pi : \sum_{l \in L^{N(\mu)}(\Pi)} w_L(l) = c_L \quad . \tag{7}$$

A convenient choice is to have $c_M = c_S = c_L = 1$, and these weights would become unit measures (which should not be confused with the probabilities used to sample elements in a test). With these conditions Υ and $\hat{\Upsilon}$ are bounded.

An optional property that might be interesting occasionally is to consider environments whose reward sum is constant or zero, as zero-sum games in game theory, where $\forall k : \sum_{i=1}^{N(\mu)} r_{i,k} = 0$.

The above definition may be too strict when we have environments with an episode goal at the end, but we want some positive or negative rewards to be given while agents approach the goal. A more convenient version follows:

Definition 5. *An environment μ is zero-sum in the limit iff:*

$$\lim_{K \to \infty} \sum_{k=1}^{K} \sum_{i=1}^{N(\mu)} r_{i,k} = 0 \quad . \tag{8}$$

With teams, the previous definition could be changed in such a way that:

$$\lim_{K \to \infty} \sum_{k=1}^{K} \sum_{t \in \tau} \sum_{i \in t} r_{i,k} = 0 \quad . \tag{9}$$

So the sum of the agents' rewards in a team (or team's reward) does not need to be zero but the sum of all teams' rewards does. For instance, if we have a team with agents $\{1, 2\}$ and another team with agents $\{3, 4, 5\}$, then a reward (in the limit) of $1/4$ for agents 1 and 2 implies $-1/6$ for agents 3, 4 and 5. The zero-sum properties are appropriate for competition. In fact, if teams have only one agent then we have *pure competition*. We can have both competition and cooperation by using teams in a zero-sum game, where agents in a team cooperate and agents in different teams compete. If we want to evaluate *pure cooperation* (with one or more teams) then zero-sum games will not be appropriate.

4.5 Team Symmetry

In game theory, a symmetric game is a game where the payoffs for playing a particular strategy depend only on the other strategies employed by the rest of agents, not on who is playing them. This property is very useful for evaluating purposes, as the results would be independent of the position of the agent.

When using teams, this definition of symmetry must be reconsidered. The previous definition means that for each pair of line-ups with the same agents but in different order, the agents maintain their previous results. But with the inclusion of teams this definition is not appropriate. For example, using an environment with the partition of slots on teams $\tau = \{\{1,2\},\{3,4\}\}$ and line-up $l = (\pi_1, \pi_2, \pi_3, \pi_4)$, we have that agents π_1 and π_2 must both obtain the same result, as π_3 and π_4 as well. Following the definition and switching the positions of π_2 and π_3 we obtain line-up $l' = (\pi_1, \pi_3, \pi_2, \pi_4)$, which now means that agents π_1 and π_3 must have the same results (since they are now in the same team) while maintaining their previous results, as π_2 and π_4 as well. This situation can only occur when all slots (and therefore teams) obtain equal results.

Instead, we extend this definition of symmetry to include teams. First, we denote by $\sigma(l)$ the set of line-ups permuting the agent positions of line-up l. This set corresponds with the one used in game theory to define symmetry. To adapt this set to include teams, we must select a subset of line-ups from $\sigma(l)$ respecting the teams defined in τ. We denote this subset with $\sigma(l,\tau)$, where we only select line-ups from $\sigma(l)$ if original teams are maintained. Following the example, line-up l' is not included in $\sigma(l,\tau)$ since π_1 and π_3 from l' were not in the same team in l (as π_2 and π_4 as well). However, $l'' = (\pi_3, \pi_4, \pi_2, \pi_1)$ is included in $\sigma(l,\tau)$, since both pair of agents (π_1, π_2) and (π_3, π_4) are still in the same team. From here, we define team symmetry as follows:

Definition 6. *We say a multi-agent environment μ is* team symmetric *if and only if every team in τ has the same number of elements and:*

$$\forall i, K, \mathit{\Pi}, l \in L^{N(\mu)}(\mathit{\Pi}), l' \in \sigma(l,\tau) : R_i^K(\mu[l]) = R_{i'}^K(\mu[l']) \ . \tag{10}$$

where i' represents the slot of agent $l_{i:i}$ in l' and whatever the function used to calculate agents' results.

Note that we impose that every set in τ must have the same number of elements. This is because we only consider multi-agent environments to be team symmetric if we can evaluate an agent in every slot and obtain the same result. Having teams with different number of elements does not allow us to do this.

This definition now fits our goal of symmetry. But too few environments will fit this definition because it is too restrictive. We could particularize this definition of team symmetry into two parts depending on the relation between the slots, with a version known as intra-team symmetry and inter-team symmetry (for more details the reader is referred to [4, Sect. 4.6]).

Definition 6 corresponds with an Intra-Team and Total Inter-Team Symmetry, where every team of agents can be located in every set of τ and in different order, maintaining their performance expectation.

5 Conclusions

Social intelligence has been an important area of study in psychology, comparative cognition and economics for more than a century, and more recently, in artificial intelligence. In this paper we have identified the components to measure social intelligence, and analyzed how we must consider these components in some instrumental properties (i.e., validity, reliability, efficiency, boundedness and team symmetry) as a first insight about what we need to create social tests.

Of course, these properties are not enough to fully assess the suitability of a testbed to measure social intelligence. Indeed, more research is needed in order to better characterize these testbeds, such as analyzing the interaction between the agents, or how cooperative/competitive the multi-agent environments are.

Acknowledgements

This work was supported by the MEC projects EXPLORA-INGENIO TIN 2009-06078-E, CONSOLIDER-INGENIO 26706 and TINs 2010-21062-C02-02, 2013-45732-C4-1-P and GVA projects PROMETEO/2011/052 and PROMETEOII2015/013. Javier Insa-Cabrera was sponsored by Spanish MEC-FPU grant AP2010-4389.

References

1. Horling, B., Lesser, V.: A Survey of Multi-Agent Organizational Paradigms. The Knowledge Engineering Review, 19, 281-316 (2004)
2. Simao, J., Demazeau, Y.: On Social Reasoning in Multi-Agent Systems. Inteligencia Artificial, 5(13), 68-84 (2001)
3. Roth, A.E.: The Shapley Value: Essays in Honor of Lloyd S. Shapley. Cambridge University Press (1988)
4. Insa-Cabrera, J., Hernández-Orallo, J.: Definition and properties to assess multi-agent environments as social intelligence tests. Technical report, CoRR (2014)
5. Legg, S., Hutter, M.: Universal Intelligence: A Definition of Machine Intelligence. Minds and Machines. 17(4), 391-444 (2007)
6. Hernández-Orallo, J., Dowe, D.L.: Measuring universal intelligence: Towards an anytime intelligence test. Artificial Intelligence. 174(18), 1508-1539 (2010)
7. Hernández-Orallo, J.: A (hopefully) Unbiased Universal Environment Class for Measuring Intelligence of Biological and Artificial Systems. In: 3rd Conference on Artificial General Intelligence, pp. 182-183, (2010)
8. Hernández-Orallo, J., Dowe, D.L., Hernández-Lloreda, M.V.: Universal psychometrics: Measuring cognitive abilities in the machine kingdom. Cognitive Systems Research, 27, 50-74 (2014)