Int J of Soc Robotics (2015) 7:825–840
DOI 10.1007/s12369-015-0300-4

Institutional Robotics

Institutions for Social Robots

Porfírio Silva[1] · **José N. Pereira**[2] · **Pedro U. Lima**[1]

Accepted: 26 April 2015 / Published online: 8 May 2015

Abstract The way human beings engage with material things in our environment is experiencing rapid modification. Human and non-human, natural and artificial creatures are on the verge of building unprecedented relations of sociability. This paper takes this process as a horizon for Social Robotics, advancing a new approach to coordinate systems of multiple robots within social spaces durably shared by humans and machines. Given the fact that institutions are the tools in use within human societies to shape social action over long periods of time, we use human-inspired institutions to deal with scenarios involving many-to-many human-robot lasting interactions. Our approach, Institutional Robotics, is inspired by leading economists and philosophers having dedicated sustained efforts to the understanding of social institutions. This paper: (1) advocates the importance of an institution-based approach for multi-robot systems (Institutional Robotics) in real-world human-populated environments, where many-to-many social interactions among robots and humans must be considered; (2) reviews experiments conducted (including novel experimental work) and methodologies used in the process of advancing Institutional Robotics. Both contributions pave the way for a new institution-based methodology to coordinate robot collectives, which stems from an inter-disciplinary approach based on robotics, social sciences and philosophy.

Keywords Institutional Robotics · Philosophical foundations of social robotics · Heterogeneous multi-robot systems · Humans–robots interactions · Mediated interaction

✉ Porfírio Silva
porfiriosilva@isr.ist.utl.pt

José N. Pereira
jose.pereira@epfl.ch

Pedro U. Lima
pal@isr.ist.utl.pt

[1] Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001 Lisbon, Portugal

[2] Distributed Intelligent Systems and Algorithms, EPFL ENAC IIE DISAL, Station 2, 105 Lausanne, Switzerland

1 Introduction

With the aim of expanding the horizon of Social Robotics, this paper uses contributions from Economics and Philosophy about human institutions to introduce an approach to institutional environments shared by humans and social robots.

There is neither a single accepted definition of (human) institutions nor a consensual list of their essential features. Furthermore, it is usually difficult to use concepts originated in human institutions directly in robotic experiments. A few years ago we began using an informal definition of institutions that serves as a bridge between inspiration coming from disciplines studying human societies and formalization requirements dictated by experimentation in robotics. This definition is as follows: "Institutions are cumulative sets of persistent artificial modifications made to the environment or to the internal mechanisms of a subset of agents, thought to be functional to the collective order". (More on this definition in Sect. 5 below.) Our approach, Institutional Robotics, aims at providing a comprehensive strategy for specifying social interactions among robots and humans where natural and artificial creatures are situated not only in a physical but also in an institutional environment, their interactions being guided by human-inspired institutions (e.g., norms, roles, hierarchies). The rest of this introduction will be devoted to motivate our endeavor.

Social Robotics is the study of robots (including humanoid robots) that are situated in a social environment and interact with humans and other robots sharing the same social space, possibly in a human-like, anthropomorphic style. Fundamental questions have been raised about Social Robotics, though not always with a noticeable impact on current research. For example, [9] evoke the following question: will the development of intelligent affective human-like robots cast doubts about the taxonomic legitimacy of the classification "human", and shall we prepare to regard robots as a new separate species developing their own sociality? Some ethnographic studies have shown the importance of a deeper understanding of the dynamics triggered in cases where robots become part of an institutional setting. For example, [35], on the Mars Exploration Rover mission, shows how a pair of robots can play a role in producing and maintaining a local order within a group or an organization, while [19], on the integration of robots into a hospital, shows how this technology can cause conflicts of interest within an organization due to, e.g., adjustments of the workflow.

What we propose in this paper is to deepen the philosophical foundations of Social Robotics in another direction. Most, though not all, research on social robotics focuses on one-to-one human-robot interaction. This approach stems from a certain conception of sociability, dominated by direct personal (peer-to-peer) relationships or social relationships within relatively small groups whose cement is a history all its members share (at least partially). We can call this "proximity sociability". What we propose is to advance the understanding of another type of sociality: many-to-many links within large populations where anonymous social relations prevail. When we realize that a robot does not have to be a creature with a physical configuration resembling a human or an animal, that a robot system can be a network of sensors and actuators, including mobile components distributed in space in very diverse configurations, we realize how productive it can be to extend Social Robotics to deal with scenarios involving anonymous many-to-many human-robot complex relationships.

There are already certain domains of social interaction being deeply transformed by the massive presence of anonymous agents. The point of anonymity is that it prevents us from easily determining whether we are dealing with humans or machines. One example is automated trading: the use of computers for very short-term trading of financial instruments in electronic markets, raising a controversy on the difficult regulation of such operations [6]. Other aspects of social interaction at a global level can be impacted by possibilities opened by similar technologies and practices.

The wider context of a research on "artificial societies" is the ongoing "metamorphosis of objects". One major scenario here is the Internet of Things. Within 20 years, a huge global network of billions and billions of "smart things" interspersed in the interactions between humans will have the potential to dramatically change the way human beings engage with material things in our environment. Some authors are writing about "autonomous objects", a "new actualization of subject-object relationships", and saying that "things will become social actors in a networked environment" (see e.g. [32,34]).

To understand the role Social Robotics can play in this scenario of metamorphosis of objects we cannot stay confined to "proximity sociability". We need to explore the many-to-many human-machine (partly) anonymous relationships scenario in which we will possibly find ourselves in a not so distant future. The relationship with a robotic system which is a network of sensors and actuators distributed in space may not be necessarily focused, say, on verbal and gestural communication skills, but rather focused, for instance, on understanding the intentions of artificial agents or robots where they are expressed by means unusual for human habits. Our vision is a scenario where this type of social interaction is characterized, from the point of view of humans, by informality, where humans no longer need any specific training to interact with robots, because they adopt when dealing with robots the same attitudes in use to deal with humans. Within the framework of many-to-many (partly) anonymous relationships, which is our scenario, the range of problems differs from those studied in the case of one-to-one direct personal, or within small groups, relationships. Given the fact that institutions are the tool human societies use to deal with this kind of sociability, this paper intends to explore how an institutional approach can contribute to an enlarged vision for Social Robotics, and enrich the toolbox for robot collectives' coordination methods. Since Economics studies this kind of social relationships, we will look at this discipline on search of inspiration to our approach.

To sum up: to meet the challenges presented by highly networked environments of objects and people, which open prospects for a deep modification of human sociability at population level, our proposal is to use human-inspired institutions to coordinate systems of multiple robots embedded in environments shared with humans. Having already motivated our endeavor, the next section introduces and problematizes the concept of institution (Sect. 2). Then, we give a brief account of our early experiments in Institutional Robotics: systems of multiple robots coordinated by institutions (Sect. 3). In Sect. 4, central aspects of the Institutional Economics' approach are further detailed. A more recent contribution in Institutional Robotics, which could pave the way for the implementation of institutional environments shared by humans and robots, is analyzed in Sect. 5. In Sect. 6, a new experimental scenario, with a more socially complex application of Institutional Robotics, is introduced and first results are reported. In Sect. 7 we conclude and reflect on some prospects for future work.

2 What are Institutions?

It may seem simple to say what institutions are. However, dealing with such a complex reality, this apparent simplicity has to be misleading. In this section we will firstly consider some definitions proposed by social scientists, mainly economists taking Economics as a branch of Social Sciences rather than a branch of Mathematics. Secondly, the question will be addressed from a more fundamental, ontological perspective.

2.1 Economics' Definitions of Institutions

Several leading economists have provided examples and underlined fundamental features of social institutions. For Douglass North, 1993 Nobel Prize in Economic Sciences, institutions "consist of both informal constraints (sanctions, taboos, customs, traditions, and codes of conduct), and formal rules (constitutions, laws, property rights)" [20]. Elinor Ostrom, 2009 Nobel Prize in Economic Sciences, takes institutions as sets of rules containing "prescriptions that forbid, permit, or require some action or some outcome", which are used "to determine who is eligible to make decisions in some arena, what actions are allowed or constrained, what aggregation rules will be used, what procedures must be followed, what information must or must not be provided, and what payoffs will be assigned to individuals dependent on their actions" [22]. Ménard and Shirley [15] say that institutions are the written and unwritten rules, norms and constraints used within human societies, including constitutions, laws, unwritten codes of conduct, norms of behaviour, and beliefs. For Geoffrey Hodgson, one of the leading heirs of "Old Institutionalism", institutions are "systems of established and prevalent social rules that structure social interactions", like "language, money, law, systems of weights and measures, table manners, and firms (and other organizations)", as well as "the informal basis of all structured and durable behaviour", informal basis that requires the presence of non-deliberative mechanisms like habits and routines [14].

Besides giving examples, these economists explain the functions fulfilled by institutions: to connect the past of a human society with its present and the future, by evolving incrementally [20]; to organize repetitive and structured aspects of interaction in human life [23]; to reduce uncertainty and control the environment [15]. Ostrom points out a crucial aspect of institutional dynamics when she says that not all rules are relevant; only working rules, "those actually used, monitored, and enforced when individuals make choices about the actions they will take (...)." [23]

Collectively, these contributions underline some fundamental features of social institutions. However, they lack a clear indication of the fundamental mechanisms underlying institutions in human societies. To overcome this lack, we need to understand the fundamental ontology of institutional reality.

2.2 A Fundamental Ontology of Institutions

We need an understanding of the basic structure of institutional reality in order to capture the essential mechanisms beneath the workings of social and economic institutions. To this effect, we need to raise a basic point of ontology: what are the most fundamental foundations of institutions? What are institutions from an ontological point of view?

John Searle's research on the construction of social reality (starting with [27]) contributes decisively to answer that question; in [28] he presents a compact and updated systematization of his approach, based on three elements.

First, collective intentionality Collective intentionality is a capacity of human beings (and of many other species) to engage in cooperative behaviour and sharing of attitudes with conspecifics. Collective intentionality can described by forms such as "we desire", "we believe", "we intend", and can take the form of intentional collective action.

Second, status functions Humans, and some animals, have the capacity to assign functions to objects. If an individual can use a stump as a chair, a group can use a log as a bench. Here, the assignment of function is supported on physical features of objects. Humans have the capacity to assign functions to objects where the physical features of the objects are largely irrelevant to the assigned function. In this case we speak of status functions. Money, as a function, does not depend on the material chosen for banknotes or coins (although material has some practical relevance, related, e.g. to easy transportation and difficult counterfeiting). Money, as well as many other institutions and institutional facts, are created and exist thanks to acts of collective intentionality: collective assignment and recognition of status functions.

Third, deontic powers Status functions are vehicles of power in human society. We accept status functions and in so accepting, we accept a series of obligations, rights, responsibilities, duties, permissions, and so on. All these are deontic powers. If I have a property, I have a certain authority over it, and I have an obligation to pay some taxes. In human societies, we have a set of deontic power relations. Obligations and permissions are reasons for action, if we can recognize them. And, importantly, deontic relationships provide reasons for action that are independent of desires. To recognize that I am the owner of this site gives people some reason to act in a certain way, those reasons not being based on any of their desires.

To sum up: on this account, institutions are all a matter of the assignment of status functions by collective intentional acts, so creating deontic powers representing reasons for action that are independent of desires. The experiment described in Sect. 3 below includes the assignment of status

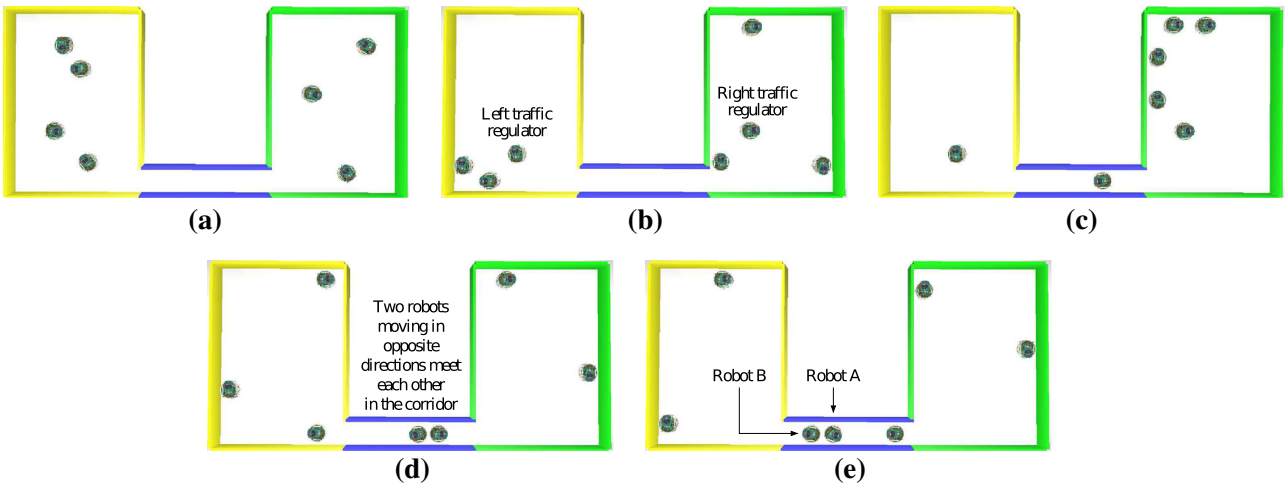

Fig. 1 Screenshots from Webots simulations: **a** initial deployment of robots in the rooms; **b** regulators in their final positions at each entrance of the corridor; **c** queue formed behind right traffic regulator, while robot moves in the corridor. **d** two robots encounter each other in the corridor; **e** after adopting the role, robot A switches the role with robot B

functions among robots. In Sect. 3.2. we discuss the prospects of implementing collective intentionality and deontic powers also in systems of multiple robots.

3 A First Experiment on Institutional Robotics

How should we use the definitions and concepts introduced in the previous section in order to design and conduct experiments with robotic systems? In subsequent sections we will begin the work of formalizing institutional concepts. In this section we start with a first attempt to experiment with some aspects of the institutional framework.

In [30] we suggested a first, broad and somehow loose definition of institutions for robots: "Institutions are coordination artefacts and come in many forms: organizations, teams, hierarchies, conventions, norms, roles played by some robots, behavioural routines, stereotyped ways of sensing and interpreting certain situations, material artefacts, some material organization of the world. A particular institution can be a composite of several institutional forms." Initial experiments with Institutional Robotics were inspired by this definition of institutions and focused on one specific institutional form: the *institutional role*. In this section a report of a first experiment on Institutional Robotics will be given.

3.1 Task Description

Our experiments were designed to be conducted on the e-puck robots [18]. The e-pucks are small (7 cm diameter) wheeled robots designed at the École Polytechnique Fédérale de Lausanne (EPFL) for use in educational and research experiments. The robots are simple, relatively inexpensive and robust, which makes them suitable for experiments in collective robotics. In our experiments, we use the e-pucks' proximity sensors, differential drive system, and camera. Local communication is achieved using the active infrared sensors on the e-pucks. We use the Webots simulation platform, a 3D, kinematic, sensor-based simulator with models of the e-puck robot defined.

The overall scenario is as follows. A collective of robots is situated in an environment containing two rooms connected by a narrow corridor (see Fig. 1a). The robots must continuously transport virtual items between the two rooms. The robots pick up the virtual items in the left room. They must then navigate through the corridor and deploy the items in the right room. The corridor connecting the rooms is too narrow for two robots moving in opposite directions to pass one another. In order to avoid congestion in the corridor, the traffic between the two rooms must be regulated so that robots only attempt to traverse the corridor in one direction at a time. In order to facilitate coordination, we let a subset of the robots adopt the *institutional role* of "traffic regulators" to control the circulation of the remaining robots in the collective.

At the beginning of each experiment, robots are placed randomly in the two rooms. Transporting robots combine the use of proximity sensors and cameras to recognize their location (left room, right room, corridor), perform wall following and get to the opposite room trough the corridor, and try to avoid conflicts with other robots and collisions with obstacles (walls). When two robots moving in opposite directions encounter one another in the corridor, a need of traffic regulation becomes patent due to this conflict. The two robots involved in this local conflicting situation place themselves each one at one of the opposite ends of the corridor and jointly assume the *institutional role* of traffic regulators (see

Fig. 1b). Each regulator will now control the flow of transporting robots entering the corridor from one of the rooms.

The regulator robots are synchronized to ensure that only one of them will let transporting robots enter the corridor from their respective rooms at any one time. Using their active infrared sensors, regulators emit messages to guide the transporting robots trying to enter the corridor. A traffic regulator periodically emits "stop" messages when it has to prevent transporting robots from entering the corridor from the room in which it is placed. Transporting robots have to be relatively close to one of the regulators (within 15 cm) to receive messages. If a transporting robot receives a message to stop, it will stop and begin to relay the stop message so other transporting robots behind it will stop too. As a result, the transporting robots will form a queue, see Fig. 1c. When the first robot in the queue receives a "go" message from the regulator to proceed, it forwards the message to any robots that may be behind it, and the queued up robots will start to move. We will discuss further ahead how the assignment of the traffic regulator role is performed.

Before turning to an evaluation of this experiment in Sect. 3.3., it will be discussed from an institutional perspective in the next subsection.

3.2 Interpreting Experiments Under an Institutional Approach

It may be difficult in many cases to translate institutional concepts into a robotic implementation, but such an exercise helps to refine concepts that were somehow vague at its roots in other disciplines. In our example, this translation exercise was needed for the concept of *institutional role*. We take a role as a behaviour specific to a subset of all robots, where it can be seen (by an internal or an external observer) as functional to some collective task or activity, and where it depends on the behaviour of other robots (in the sense that others must recognize and/or permit such role playing by particular robots).

The experiment we are reporting addresses three crucial issues for institutional roles: role assignment (how some robots start playing a role), role recognition (how robots recognize that some others are playing a role), and role permission (how robots permit other robots to play a role and behave accordingly). These three issues specify one of three elements involved in the ontology of institutional reality (according to Searle, see above, Sect. 2.2.): the assignment of status functions. The two others elements are collective intentionality and deontic powers.

Consider collective intentionality first. The robots from our experiment, are they endowed with genuine collective intentionality? Probably not. The next useful question is: would it be possible to implement phenomena of collective intentionality in robots? Could a group of robots, not only do something collectively, but also have a representation of that collective action? Could, for example, a team of robotic soccer, not just perform coordinated movements that look like a soccer match, but also understand what a soccer match is and have a collective intention to win the match? This capacity for collective intentionality could greatly improve the performance of the team. Some believe that this type of intentionality is unique to humans and depends on our sophisticated language. Let us see.

The intentionality of human languages basically consists of linguistic elements mapping aspects of the non-linguistic world (the world that is beyond language). However, this mapping clearly exists in other quadrants of the natural world. An example is the "dance language" of honey bees [36]. Ruth Millikan [16,17], drawing on biological functions resulting from natural evolution, considers the bee dance as an example of intentionality. This variety of intentionality lacks the power of representation that human languages have: we know what the bee dances are about, they don't. Bees just perform the dance and react to it appropriately. Yet, that variety of intentionality is very effective: it maps a region of the world and recruits other individuals to specific actions in the external world. We can therefore accept that intentionality can come in degrees. For Millikan, essential in establishing the intentionality of bees dance is the historical nature of the evolutionary process. Now, what we want to stress is that the historical character of intentionality can be implemented also in systems of multiple robots, using for example, artificial evolutionary processes or collective reinforcement learning. Future experiments on Institutional Robotics should explore these possibilities.

Deontic powers (the use of obligations, rights, and permissions) are the other ontological constituent of institutional reality. Deontology deserves some consideration here, because it will be put into use by the Institutional Agent Controller methodology to be described in Sect. 5 below. The use of institutional norms among humans usually provides a variable degree of conformity to such norms without disrupting the collective. For example, a norm can have legal force and yet be considered illegitimate, which could justify disobedience, contestation and even repealing the norm. We easily accept that our current robots cannot fully recognize a deontology, but this does not exclude the progressive acquisition of this capability in the future. In the same vein of our previous statement about intentionality: we can accept that deontology can come in degrees. The well-known concepts of learning and adaptation in Robotics can play a role here. In stochastic learning algorithms, such as reinforcement learning, the policy is often probabilistic (to balance exploration and exploitation) until the algorithm converges to a situation where the optimal decision is always taken with probability one. Therefore, one can conceive a similar process that learns from scratch an institution (i.e., the opti-

mal set of rules - though the institution is not limited to a set of rules) that, after a learning process, will become the decision rules without further need for formulating hypotheses and trying different alternatives. This (learning an institution) is a crucial step, certainly hard to implement, but essential to the process of coordinating a collective based on institutions, where the complexity of decision-making decreases as more (possibly hierarchically organized) institutions are created, reducing the uncertainty and time involved in collective decision-making.

Another aspect we want to emphasize in this experiment concerns the distinction between "role" and "individual." A practical difficulty in mounting the experience led to the process we call "role propagation" and this is directly linked to the distinction between role and individual in an institutional environment.

The assignment of the traffic regulator role happens when two robots moving in opposite directions get stuck in the corridor (see Fig. 1d). Both robots check if there is already a regulator in the room they came from (by communicating with a workstation that keeps track of regulator status) and, if no regulator is present, assume the traffic regulator role, retreat to that room, and place themselves at the entrance of the corridor. However, after two robots moving in opposite directions inside the corridor have assumed the role as regulators, other robots may already have entered the corridor behind them and prevent them from navigating back to the entrances, see Fig. 1e. Role propagation takes place at that juncture to speed up conflict resolution: the role is propagated to the last robot that entered the corridor from a given direction. In this instance, although originally robot A assumed the regulator role, upon their encounter in the corridor, robot B will assume the role and robot A will become a transporting robot again. After a certain number of interactions, both regulators abandon the role and the system goes back to the initial state, giving other robots the chance to take the role. This means: no robot is specifically designed to play any particular role. In principle, any individual can play any role. Playing a role is something justified because of a collective need, not as a right or an inherent feature of any individual.

In more general terms, not restricted to the previously described experiment, this distinction between "role" and "individual" can help to address robustness issues for systems of multiple robots. Let us state, within this framework, that the property of robustness is about at what extent a collective is able to respond, at a structural level, to a perturbation resulting from removing/adding individuals with specific roles from/to the collective. Now, within an institutional framework, roles must be distinguished from particular individual robots. Robots are heterogeneous with respect to some features, but fully interchangeable with respect to some other (basic) features. This makes any robots in principle able to play any role, even if some learning can be required to attain full mastery. Robots are redundant in relation to roles. To this effect, different institutional roles must not be allocated by fixed, once for all, mechanisms (e.g. "genetic" mechanisms) but, instead, by institutional assignment of status functions. If this can be implemented, removing specific individuals from the collective does not amount to renounce to specific roles. On the other side, the adding of individuals with malevolent roles can be countered by a specific feature of institutional roles: for an individual to play a role, other participants must recognize that role as part on the institutional setting, and accept to behave accordingly. The refusal to accept an individual playing a role (because the role is not part of the institutional setting) can be a mechanism to prevent the intrusion of malevolent roles.

3.3 Results

We prepared different setups in order to evaluate how parameters such as the size of the robotic collective and the length of corridor affect the performance. Three different corridor lengths (50 cm, 100 cm and 200 cm) were considered. For each corridor length, we ran experiments with different numbers of robots (7, 15 and 20 robots). We also implemented a different solution to this task, based on the principles of swarm robotics, (for details and examples see [2,3]), where no explicit group-level coordination mechanism is at work, individual robots are steered by simple rules and have access to local information only, and the robots rely exclusively on self-organization to solve the task. For each of the nine resulting setups, we repeated the experiment 30 times (with a duration of 15 minutes for each) both for our proposed institutional robotics approach and for the swarm robotics inspired approach. While results with different corridor lengths show how both approaches perform under different levels of task complexity the focus is, in these experiments, on how effectively collectives of different sizes are able to allocate resources to coordination (by allocating roles) or use those resources following a self-organization approach. Consequently, in Fig. 2 we show boxplot distributions of number of successfully transported items (our performance metric) for different sizes of collectives and both approaches, while considering a 100 cm corridor. The comparison of these two approaches was not mainly targeted at determining which one is the best approach, but rather to better understand how different features influence the effectiveness of the institutional device in use. While further research is needed on that direction, first results from the series of experiments we are reporting here are already enlightening. We observe that larger collectives have a greater need for regulation than smaller ones, as they are more prone to coordination failures (getting stuck in the corridor). This can be seen in the increase in performance when we use 15 robots instead of 7; with a more coordinated approach (insti-

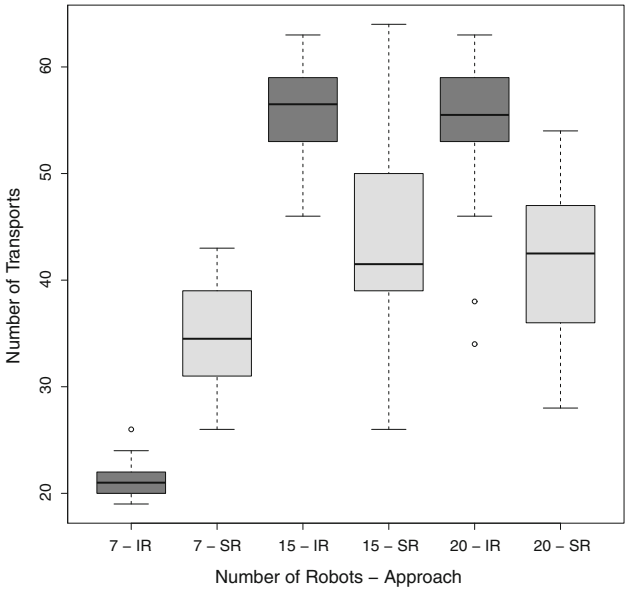

Fig. 2 Distribution of number of successful transports for different sizes of collectives (7, 15, 20) and different approaches (*IR* institutional robotics—*dark grey*, *SR* swarm robotics— *light grey*)

tutional) we have a much larger increase than with a more self-organized approach (swarm). A clear point is that using institutional roles to coordination purposes represents a (proportionally) heavier effort for smaller collectives (e.g. using two robots to improve coordination of the collective represents a different effort for collectives of size 7 or 20). This is clearly visible in Fig. 2 when we compare the performance of the institutional approach for different sizes of collectives. Several other results also appear: apparently uninteresting features of the scenario seriously impact the importance of coordinating (the longer the corridor, the heavier the negative impact of conflicts occurring at that place on the performance of the collective); the need to prevent conflicts can be more pressing to some systems than for others (e.g. due to different hardware in use). This shows that the fine tuning of an institutional form, so that it fits certain concrete circumstances, can be decisive for its practical operation. Also human institutions have to adjust to the historical, cultural and social context. This could imply that robots embedded in institutional environments shared with humans should have ways to be part (in active or passive mode) of that same process.

In itself, the case study of this section could not justify an institutional approach (relatively simple problem; few robots). However, it served to, for the first time, implement with robots the concept of institutional role. This scenario is not human-machine, but it can be extended in that direction: it is a plausible real life scenario for humans and robots sharing a task. Solutions developed in this case study may be a first step towards an institutional approach to scenarios of this kind. The definition of institutions introduced in this section has to be progressively improved and formalized. (In Sect. 5 we introduce a formal model of institutional controllers for the coordination of robot collectives, using Petri nets as a tool.) However, informal definitions cannot be abandoned too hastily when we need to keep alive a connection to other disciplines providing inspiration for new approaches. We are well aware of the many limitations of this first experience in Institutional Robotics. These limitations become even more apparent when we consider in depth what we can learn from Economics about the dynamics of institutions. We devote the next section to some basic institutional concepts from Economics with the purpose of highlighting their potential impact in multi-robot systems designed using Institutional Robotics concepts.

4 Human Institutions for Robots

The key point requiring an institutional approach to many-to-many social interaction within populations of anonymous individuals is as follows: under these conditions "proximity sociability" cannot be the only cement of the collective. It is not feasible to ground this kind of social relationship only on face to face, personal relationships between individuals who know each other. For example, respect for social norms cannot be guaranteed only by emotional ties between close friends or by personal respect between people meeting in social spaces. There is a need for a more abstract respect for social norms as a contribution to common good. Coordination and cooperation in a complex society cannot depend only on individuals' interior motivations; they have to be supported also by the structure of the institutional environment. Thus, we will adopt here an approach that, instead of modifying the original robot behaviours, rather builds institutions within institutional environments that constrain them.

The structure of a situation, not the internal motivations or capabilities of the individuals, can be the main factor causing the observed actions of a system. In a series of computational economics experiments, Bosch, Gode and their colleagues ([4,11–13]) have shown that situations previously explained, under the substantive rationality paradigm, as a result of individual rational behaviour, can be explained by the institutional setup itself. Specifically, they have shown that Pareto efficient outcomes are achieved within double auction contexts by "zero-intelligence" (ZI) traders.

ZI traders are software agents whose decision rules fall far short of utility maximization. ZI traders are not endowed with any kind of high level intelligence, motivation or learning of the kind human individuals are supposed to enjoy. They just submit random bids and offers, under some imposed simple constraints (like not permitting traders to sell below their costs or buy above their values). On the institutional side, double auction markets are very specific contexts. An auction

is a market institution where messages from traders include some price information (e.g. an order to buy or to sell at a given price) and which gives priority to higher bids and lower asks. A double auction is an auction where both buyers and sellers are allowed to make orders. Now, those experiments show that rules of specific market institutions, and not necessarily the maximizing capabilities of individuals, can be responsible for efficient aggregate outcomes. "Performance of an economy is the joint result of its institutional structure, market environment, and agent behaviour" [12]. Interpreting some of the earlier experiments of Gode and Sunder, Denzau and North [8] say that, in such kind of context, the difference in institutions alone explains the main differences between diverse aggregate outcomes.

One important aspect of understanding institutional environments consists in acknowledging institutional diversity. The need to respond to so many different contingencies in so many different action situations lead agents to multiply and diversify institutional arrangements. We need to understand this diversity and recognize that different contexts may require different combinations of institutional forms. For example, shall we abandon in all cases the much-criticized paradigm of substantive rationality? No; the point is not that this kind of behaviour does not exist. It exists in some economic situations, like competitive posted-price markets [21], where a price is announced indicating what a firm will pay for a commodity or the price at which firm will sell it, this being done without a link to an actual particular exchange of that commodity. In some situations, posted prices of the major companies are aggregated to form postings, serving as benchmarks for a market. For instance, in the Western Canadian oil market, the major companies post prices as a differential to the West Texas Intermediate posted price. In such a context, the environment makes the situation relatively simple to the agents: the price is viewed as a parameter; only the quantity needs to be chosen. That is the kind of situation that favours the behaviour corresponding to the assumptions of the substantive rationality paradigm. What we must recognize is that the domain of application of the substantive rationality paradigm is not universal.

The point is not to exclude certain types of social agents, but to understand that different institutional environments (combinations of different institutions) are likely to produce different results in different contexts. Durfee [10] already remarked the same about MAS: "It does not seem possible to devise a coordination strategy that always works well under all circumstances; if such a strategy existed, our human societies could adopt it and replace the myriad coordination constructs we employ, like corporations, governments, markets, teams, committees, professional societies, mailing groups, etc." We cannot, therefore, give a recipe to build institutional environments. Instead, we devote the rest of this section giving examples of how the institutional inspiration can be useful in modeling and implementing complex scenarios of many-to-many interaction. These examples advocate the importance of an institution-based approach for multi-robot systems in real-world human populated environments. The experiment in Sect. 6 below explores the use of these concepts.

4.1 Mediated Interaction with Representations

One crucial point of the institutional approach is that institutions allow direct and immediate interaction being replaced by indirect and mediated interaction of a much more sophisticated kind. Money is a classical example of the power of institutions in providing the means for mediated interaction [7]: "Adam Smith pointed out the hindrances to commerce that would arise in an economic system in which there was a division of labor but in which all exchange had to take the form of barter. (...) A person wishing to buy something in a barter system has to find someone who has this product for sale but who also wants some of the goods possessed by the potential buyer. Similarly, a person wishing to sell something has to find someone who both wants what he has to offer and also possesses something that the potential seller wants." The use of money overcomes this difficulty allowing mediated or indirect interaction.

Now, the most powerful features of institutional mediated interaction depend on the use of representations. Robotics capabilities directed only to immediate material features of the environment are not enough to make robots able to recognize deontology (permissions, obligations, and prohibitions). For instance, while a robot might detect bills as green rectangular shapes, without the proper representation it could not recognize the permissions (e.g., the possibility to trade the bill for other necessities), obligations, and prohibitions associated. Moreover, real agents in the real world at large frequently act based, not necessarily on accurate knowledge about reality, but on representations, internal models of the external world. Fortunately, common ground exists between economics and computer sciences concepts of representations and models of the world, easing the task of conceptualizing human-machine mediated interaction. Denzau and North [8] use the concepts of "mental model", "ideology" and "institution" to classify representations. James Albus, proponent of a classical global architecture of an intelligent machine, clearly states the basic link between mediated interaction and internal (mental) models of the (external) world [1]: "The world model contains knowledge of things that are not directly and immediately observable. It enables the system to integrate noisy and intermittent sensory input from many different sources into a single reliable representation of spatiotemporal reality."

Internal models of the external world can help to address stability issues within systems of multiple robots. Let us state,

within this framework, that stability concerns the response of a collective to a perturbation on the coupling between the agents/robots (eliminating communication links among collective members). Now, during periods where no communication at all is possible among a population, models of the world can replace, at least for a while, actual data from natural and social world. Where some communication is still available, persisting links can be used to update specific aspects of the working models of the world. This concept can be applied to a network of static cameras and mobile robots that interact with human groups, where the mobile robots require a prediction of the group behaviour in order to act so as to satisfies its needs (e.g., robots playing with children in a hospital). The cameras can provide data about the humans' behaviour, process it to estimate the actual behaviour, and communicate that information to the robots. However, in the occasion of a camera failure or in blind spots of the camera network, human observation may not be possible. In that case, the robots will have to rely on the latest available estimate of the humans' motion/gestures and use their social behaviour models (the "working model of the world") to predict their next moves as its best bet.

The experiment in Sect. 6 below shows how complex social settings may require agents endowed with the internal representations needed for mediated interaction.

4.2 Heterogeneity

To deal with individual action in a social context, it is crucial to understand that not all individuals have the same interests. We have to take into account heterogeneity within a population. Ostrom [22] suggests two key aspects in understanding this heterogeneity. First, not all individuals value the same way respecting or disrespecting social norms. Second, how different individuals in the same action arena value the decision to be taken depends on the time horizon and the set of related opportunities each enjoys outside that particular action arena. Discount rates (how opportunities are perceived) and norms (how norms are perceived) are sources of heterogeneity within a population. Given that discount rates will be used in the experiment introduced in Sect. 6 below, let us go a bit deeper on this concept now.

Discount rates are used to compute the current value of a future payment. In financial terms: how much I need to invest now, at a given discount rate, to have, in N years, the amount X. Discount rates measure the opportunity cost of capital: at the current discount rate, shall I invest my money or rather spend it now. The application of this concept to social dynamics is apt to represent the range of different opportunities enjoyed by different sets of participants in an action arena [22]. This use of discount rates parallels the use of discount factors in Reinforcement Learning to model short-sighted and more far-sighted learners [33].

Suppose we want to understand how a group of individuals act as interdependent appropriators of a natural resource system that is sufficiently large as to make it costly (and in some cases infeasible) to exclude one potential appropriator from accessing it. Take as an example an inshore fishery with a traditional small operation of a reduced number of local fishers in small boats. Renewable resource systems have problems of sustainability: the average rate of withdrawal must not exceed the average rate of replenishment. In some cases, investments made in maintenance and repair can improve sustainability. This kind of resources has subtractive attributes: the fish harvested by one boat are not there for someone else. Crowding effects and overuse, as well as the risk of opportunistic behaviour, are chronic in this kind of resource systems.

Now, how different individuals discount future benefits in different ways is a critical element of the dynamics of this kind of situation. Different discount rates depend on several factors, all related to different time horizons and different opportunities enjoyed by the agents. "In a fishery, for example, the discount rates of local fishers who live in nearby villages will differ from the discount rates of those who operate the larger trawlers, who may fish anywhere along a coastline. The time horizons of the local fishers, in relation to the yield of the inshore, extend far into the future. They hope that their children and their children's children can make a living in the same location. More mobile fishers, on the other hand, can go on to other fishing grounds when local fish are no longer available." [22]

The concept of discount rate allows the modeling of complex social environments where different actors ascribe different weights to the same opportunities or actions that are available to everyone. In collective robotics, especially in swarm robotics, homogeneous teams are often assumed. However, this is far from reality in human societies. Thus, if one wants robot teams interacting naturally with humans, and/or methodologies that ascribe human-like behavior to robot collectives, they should be heterogeneous, and different discount rates are a significant way of introducing heterogeneity. The experiment described in Sect. 6 below deals with a realistic situation of heterogeneity (represented by means of discount rates) heavily impacting the sustainability of a system.

5 Towards Implementing Shared Institutional Environments

The concepts from Institutional Economics reviewed in the previous section open a series of challenges we must face if we are to move towards an institutional approach to Social Robotics. To mention just some of these challenges: there is a dynamical interplay between the inner world and the outer world (which comprises other individuals) of any individual,

and it is not easy to determine where that boundary exactly lays; given the heterogeneity of individuals, some phenomena can be better understood at population level; it is not possible to find a mix of institutional forms that constitutes a unique response to a given social situation (at least the inspiration in human societies does not help here). Some of these challenges are closely related to the dynamic and historical character of institutional environments.

In [31] we gave a definition that tries to capture the substance of this institutional problematic: "Institutions are cumulative sets of persistent artificial modifications made to the environment or to the internal mechanisms of a subset of agents, thought to be functional to the collective order". Several qualifications in this definition reflect the historical dimension of the institutions (both for humans and robots): "persistent" excludes occasional, fortuitous modifications, while "cumulative" is meant to exclude ad hoc interventions (e.g. breaking the legs of people to implement a prohibition to run away) as institutional devices (broken legs can persist for a while, but cannot accumulate from a collective and historical point of view); "artificial" requires the modifications to be feasible by the agents, their ancestors, their teammates or their designers, excluding "natural" (e.g. evolutionary) modifications usually not manageable directly by the agents; the requirement of being thought functional to the collective order also evokes persistence, excluding the invention or strong modification of an institution just to face particular circumstances.

When we try to implement computationally a concept inspired in human sciences or in philosophy, often a level of vagueness surfaces. This requires a constant dialogue between different approaches, an effort we have been pursuing. This dialogue resulted recently in another step in this journey.

As part of the long term goal of implementing Institutional Robotics, we have recently introduced and implemented a formal model of institutional controllers for the coordination of robot collectives [26], using Petri nets (PN) [5] as a tool. Institutions, taken as coordination artefacts, are part of the robot controller, working as norms or procedures the agent has to follow in some circumstances. Each institution is formalized as an Executable Petri Net (EPN) that takes into account robot actions and sensor readings and encapsulates a behaviour that can be executed by the robot. This formalization is extended with conditions that dictate the start and end of execution of the institution, and with deontic operators stipulating how it relates to other institutions. Start and end conditions allow the institution to be *active* or *idle* depending on the situation the robot is in (since institutions are not usually relevant for all situations). On the other hand, deontic operators allow the regulation of the state of activity of related institutions to prevent inadequate concurrent execution of sets of institutions (since not all institutions can be executed simultaneously). Individual behaviours are also part of the robot controller. While social interaction (coordination among robots) is controlled by institutions, interaction with (other aspects of) the environment is controlled by individual behaviours. The robot controller, designated as Institutional Agent Controller (IAC) and shown abstractly in Fig. 3a, is generated by the composition of individual behaviour and a set of institutions, with the composition procedure being guided by the deontic operators of institutions.

In Fig. 3a we graphically represent an abstracted IAC. The lower layer of the IAC contains the EPN representations of institutions and individual behaviours, for instance, institution *Inst*. These are constructed using actions (associated with PN places) and events (associated with PN transitions) and define a formal controller for the robot, translating sensor readings and internal memory into actuator outputs. Institution *Inst* is constructed with two events and two actions that are sensed and executed by the robot. Note that this structure is not a fixed template for institutions but rather an example. Each institution and individual behaviour can be seen as a module (areas surrounded by the dashed line) that can be composed in order to generate the IAC. This composition is performed algorithmically in the higher layer of the IAC. Each module (institutions and individual behaviours) is represented by a macro place, shown in Fig. 3a in bold (m_{Inst} and m_{Ind}). Through their connections to the lower layer EPNs these places specify if a given module is active. The decision of activating or idling an institution is performed using start and end conditions that are part of the definition of institution. In Fig. 3, these are specified as transitions connecting the macro and the idle places of *Inst* (m_{Inst} is marked when *Inst* is active, $idle_{Inst}$ is marked when *Inst* is idle). Deontic relations among institutions, and between institutions and individual behaviours, are specified in the definition of institution (in the form of deontic operators) and implemented in the higher layer composition. In Fig. 3a, concurrent execution of institution *Inst* and individual behaviour *Ind* is forbidden (*Inst* prohibits *Ind*), thus the necessary structure for this regulation is added to the IAC. Further details on the workings of deontic operators and the composition and execution algorithms for IAC are given in [26]. The IAC methodology has been validated by designing and implementing a controller for the wireless connectivity maintenance case study [25], using up to 40 robots, and comparing the results obtained by probabilistic models with real results and results obtained using other control methodologies [24,26].

EPN formalization of institutions and individual behaviours is generic, in the sense that the same kind of action can be implemented by different robots (e.g. heterogeneous in terms of hardware) in different ways without any disturbance to the social interaction (like, in human societies, the same institution can be implemented by alternative means: e.g. money is implemented by different currencies in different

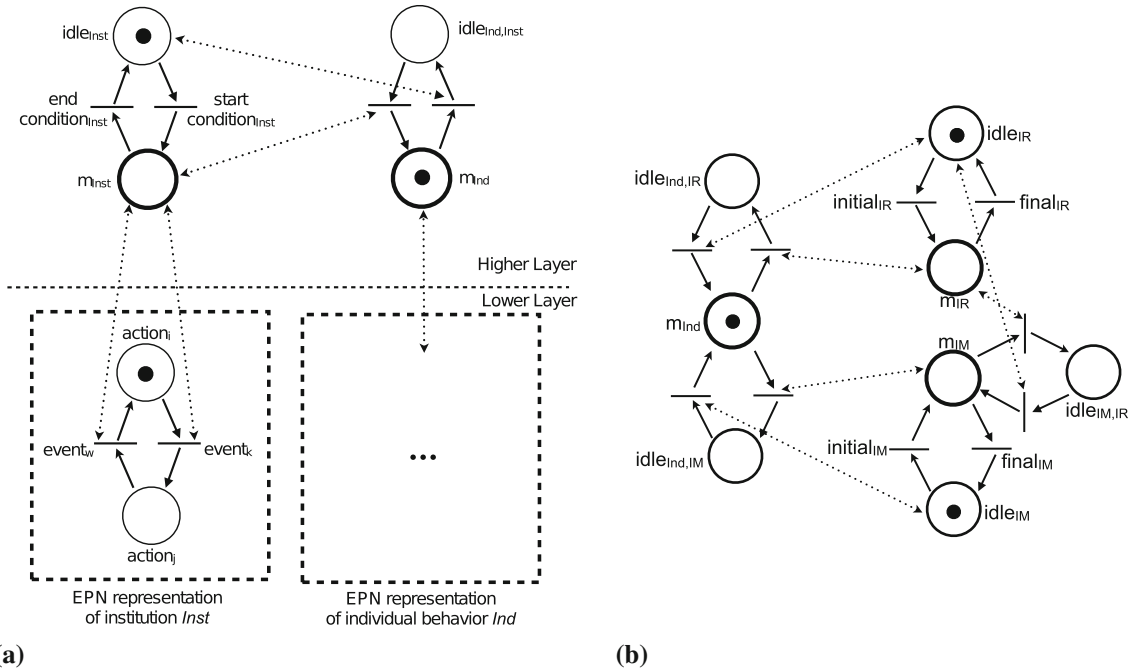

Fig. 3 **a** Abstracted institutional agent controller (IAC) graphical representation. *Lower layer* modular EPN representation of institution *Inst* and individual behavior *Ind* constructed using robot actions and events. *Higher layer* regulatory places and transitions added during IAC composition. Macro places m_{Inst} and m_{Ind} specify if the modules are active, idles places $idle_{Inst}$ and $idle_{Ind,Inst}$ specify if the modules are idle. Initial and final conditions for *Inst* regulate this decision. **b** Higher layer composition net of IAC for corridor case study. Place m_{Ind} represents the individual behavior Ind. Place m_{IR} represents institution I_R. Place m_{IM} represents institution I_M

monetary zones). Each robot in a collective setting mediated by institutions runs its IAC: institutions become part of the inner life of the agent. At the same time, each institution is distributed in the multi-robot system since its representation is replicated in each agent. This way, we can start the powerful engine of the interactive workings of inner life and outer life mechanisms of the agent and his social and natural environment. Within the IAC methodology, abstract representation and modularity replicate in robot collectives some crucial features of human institutions, namely the faculty of changing the institutional environment (and so influencing the interaction among agents) without interfering intrusively with the internal world beneath individual behaviours.

We can apply the IAC methodology to produce a robot controller for the case study described in Sect. 3. To do so we must distinguish between individual behaviours and institutions. The individual behavior of the robots specifies how the task at hand is accomplished. Picking up virtual items and deploying them is an individual behavior, since the robot interacts only with the environment. A single robot could accomplish the deployment task, although performance would be critically reduced. Thus, we specify the individual behavior with an EPN Ind. The main social behavior is the traffic regulator institutional role. We specify this behavior as an institution I_R that manages the role of traffic regulator. Its start condition $initial_{IR}$ is the detection of a conflict in the corridor and its end condition $final_{IR}$ is the end of regulation (time limit). However, the institutional role is not the only social behavior present. Institutional roles depend on other robots' behaviors, in the sense that they must recognize and/or permit such role playing by particular robots. A second social behavior present in this task is the recognition and compliance with the traffic regulator. The behavior corresponds to an institution I_M that manages the reception of messages from the traffic regulators and their relay. Its start condition $initial_{IM}$ is the reception of a stop message and its end condition $final_{IM}$ is the reception of a go message. To prevent inadequate concurrent execution of behaviours, institution I_M forbids the execution of the individual behavior Ind and institution I_R forbids the execution of all other behaviors. The composition of these behaviours in the higher layer of an IAC is shown in Fig. 3b. The full IAC would be composed of this layer plus the lower layer EPNs for each behaviour, which we choose to not show herein since they are out of the scope of this paper. As a feasibility test, we implemented the scenario described in Sect. 3 using this IAC, showing no observable distinction between how the task is executed with respect to a more traditional Finite State Automata (FSA) methodology (used to produce the results described in Sect. 3). However, and since the purpose of this

experiment was only to show that our proposed formalization is supported by an adequate adhesion to real results, we currently have no analytical or numerical comparison of the two different implementations. A more detailed comparison between the IAC and FSA methodologies is tackled in [26] for a different case study.

Three points in the IAC methodology deserve special attention as contribution to advancing the institutional approach.

Firstly, there is no absolute, natural distinction between individual behaviours and institutional behaviours of one agent. This is true even in the case of natural species creatures (human beings, for example): a behaviour that was initially induced by social pressure can be internalized by learning and even become automated to the point of ceasing to be easily inspected. Or, conversely, an automatic behaviour may become the target of social scrutiny, be inspected and eventually be consciously modified or even abandoned. In the case of our artificial creatures (robots), since they do not belong to a species, the distinction is largely one of design choice. In the IAC methodology, the distinction is based on the differentiation between interactions with the environment (behaviours taken as closely related to the robots own goals and their individual "struggle for survival") and social interactions (among robots). The crucial difference is that each robot cannot escape the initial design of individual behaviours by the makers and programmers of the system (hardware and software), while behaviours resulting from institutions can either be subject to a decision (of the individual robot) to conform or not to conform to the norms or even be engineered by deliberate modifications of some institutions (made by the robotic collective or by the humans sharing the environment with the robots).

We use this distinction between interactions with the environment and social interactions as a guideline to construct the IAC. However, some choice is still left to the designer of the system. Consider the example of a robot driver, moving on a road from point A to point B, while obeying the traffic code norm to drive on the right side. Such norm could be implemented as an individual behaviour (meaning that it has been internalized by the robot and is now related to its own goals) or as an institution (meaning that it is a social norm that helps coordination with other drivers and is executed only under certain conditions). Both options are valid depending on the particularities of the system. For instance, if the robot operates only in circumstances where the right side norm is valid and must always be executed, then it can be implemented as an individual behaviour. Otherwise it should be implemented as an institution.

Secondly, the use of deontic operators to specify how one particular institution relates to others (e.g. blocking the concurrent execution of some of the institutions), while preserving modularity (each institution can be executed independently or together with other institutions, and modifications to one institution will automatically translate into the robot controller without any further ado), effectively models the fact that institutions do not work in isolation. The crucial concept is rather "institutional environment", a network of coordination artefacts that come in many forms and articulate in many ways, their impact on collective order resulting both from their particular features and from their combination.

Thirdly, at the current state of development of the IAC methodology behaviours (individual or institutions), initial and final conditions for institutions, and deontic operators, must all be implemented or chosen by the designer of the system. This represents a decrease in the autonomy of the robots being controlled. However, what institutions should be executed is still a choice the robots must take, based on their own sensor information and internal memory. The choice of PNs as the tool to specify IACs was made taken into account the possible use of several extensions in our work. For instance, making use of stochastic PNs will allow us to specify probabilities for the execution of each institution, and thus represent non-conformity with institutions in our IAC.

In the future, it is our goal to provide the robots with tools that will enable them to autonomously create institutions out of need (e.g., the traffic controllers after observing repeated traffic jams in narrow corridors).

From the point of view we are exploring in this paper - institutional environments shared by humans and social robots - this contribution is a significant step forward. Given the uncertainty, heterogeneity and divergent interests among agents that would characterize a many-humans-to-many-robots relationship within a complex social environment, the possibility of adapting the coordination of robots to the context may be of interest to humans sharing with them the same social space. In the IAC methodology this is done without modifying the basic (individual) behaviours of robots. The social control results from modifying institutions the robots recognize. Institutions, being generic, can be implemented in robots with different hardware and different architectures. This scenario approximates the kind of social control that we consider acceptable in human societies and may allow a more informal relationship with robots sharing our social environment.

6 A More Socially Complex Application of Institutional Robotics

Within the Institutional Robotics approach we are interested in considering more complex social environments shared by robots and humans. In the case study presented in this section we increase the social complexity in three fronts by considering: (i) a large number of individuals (experiments performed with 50 simulated robots); (ii) heterogeneity in the popula-

tion; and (iii) a social dilemma in which the robots must take an individual decision.

We envision that sustainability will be an important property in truly social human-robot mixed systems, where robots act as independent entities with their own goals. Herein, we study the possibility of an institutional approach turning an unsustainable system into a sustainable one. We consider that sustainability describes the ability of a robot team to keep its members operational during run time. In this section we introduce a transport and assembly task, designated as the *piece assembly case study*, and describe only a part of the experiments performed. A broader view, including more details and results (experiments with up to 1000 simulated robots), can be found in [24]. Robots in a heterogeneous team need to collect *components* of different types, which are the basic building blocks needed for *pieces* to be assembled. There are different types of components to be assembled in specific ways. These components are present at specific locations of the environment, and must be transported to a particular area called *assembly site*, where the assembly process takes place. The team goal is to maximize the number of pieces assembled.

Robots spend energy while moving through the environment and are rewarded with energy, both when they deliver components to the assembly site (immediate rewards, received in the moment of delivery) and when pieces are correctly assembled (collective rewards, received upon piece completion and divided by all agents that contributed with components to the piece). In some situations, either by virtue of the physical environment or by virtue of choices of individual robots, the energy level of some robots will drop below zero and they will become non-operational, possibly leading the system to an unsustainable state.

There is a social dilemma of how to explore the resources present in the environment. Depending on how the components are delivered to the assembly site, robots can get either high or low immediate rewards. However, deliveries that give a high reward are not always useful for the piece assembly process, possibly leading to some delay, while deliveries that give a low reward can help speeding up the process. Robots can give priority to their individual goal of remaining operational (preferring high immediate rewards but possibly contributing less to piece completion) or to the collective goal of maximizing team performance (accepting low immediate rewards and depending more on collective rewards). This decision reflects a conflict of interests between individual robots and team.

The heterogeneity in our robotic team comes from considering two types of robots that take different decisions regarding this dilemma. Let us explain. Immediate rewards depend only on the individual robot delivering the component to the assembly site, and are received in the moment of delivery; at the time of delivery, the robot knows exactly the value of the immediate reward corresponding to each of the available ways it has to deliver the component it carries at that time. In contrast, collective rewards depend on other robots delivering the appropriate components at the right time to allow the correct assembly of the whole piece, and are received at a later time, when the piece is completed (or not received at all, if piece assembly fails). Robots of "*type 1*", whose decision process includes a high discount rate, have a short time horizon, and therefore take into account only immediate rewards, being blind to collective rewards, so giving priority to their individual goals. The decision process of robots of "*type 2*" includes a low discount rate, giving them a longer time horizon, and so making them able to weigh both immediate and collective rewards. In this way, robots of *type 2* (but not robots of *type 1*) tend to make decisions that include knowledge about the benefits resulting from the success of the assembly process, which is a collective process.

Here, some possible states of the world are not "facts" that exist independently of agents. On the contrary, the concrete reachability of some future states of the world depends on other individuals' dispositions to behave in such and such manners. This is important from an institutional point of view, because such dispositions depend on the representational capabilities of the agents.

We propose two approaches to this task that differ only in the manner in which the piece assembly process occurs. A fully *decentralized* approach decomposes the process into individual decisions taken by the robots delivering the components, these being dependent on the robots' priorities (individual vs collective). On the other hand, the *institutional* approach puts the burden of piece assembly on one robot from the team, using an institutional role designed for this purpose, which we will designate as *institutional assembler*. The institutional assembler role is general, in the sense that any robot in the team can perform the role and no specialized robot is needed. Other robots transport the components to the assembly site, where the robot executing the institutional assembler role conducts the assembly process and attributes the rewards to the corresponding robots. Since the institutional approach requires an extra effort in terms of coordination, we consider that using the specified institution has an associated cost (implemented as a percentage of the collective rewards attributed and designated as *assembler fee*). Both approaches were implemented using the EPN/IAC methodology.

We performed simulations with both approaches (100 runs each). For the decentralized approach we varied the proportion of *type 1* robots in the population from 0 to 100 %. This allows us to observe the impact of having a heterogenous population and an increasing number of robots more inclined to fulfill their individual goal, rather then the collective goal. This parameter has no impact on the institutional approach, since the assembly process is conducted by the institutional assembler and not dependent on the decisions of individual

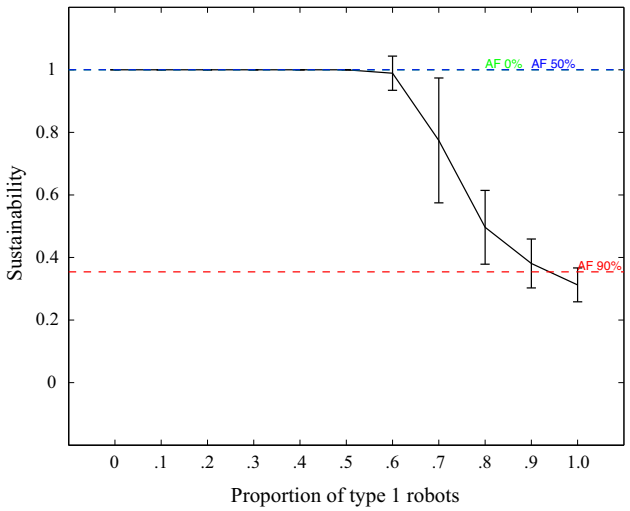

Fig. 4 Average sustainability (*bars* represent variance) for decentralized approach with different proportions of *type 1* robots in the population (*black line*). *Horizontal dashed lines* represent values for institutional approach (not dependent on proportion of robots): 0 % AF (*green*), 50 % AF (*blue*), 90 % AF (*red*). (Color figure online)

robots. For this approach we tested different values for the assembler fee (AF): 0, 50 and 90 %. This parameter allows us to study what costs for institutions are acceptable to the system before it becomes unsustainable.

In Fig. 4 we display the sustainability values obtained (mean proportion of experimental time robots remain operational). In the decentralized approach, sustainability is at maximum value for populations with less than 60 % proportion of *type 1* robots and drops as this parameter is increased. In the institutional approach, for some values of AF, the system is sustainable irregardless of the proportion of *type 1* robots in the population. This shows that, in some cases, an institution is able to turn an unsustainable system into a sustainable one. Nevertheless, the cost associated with the institutional assembler must be taken into account. At 90 % AF the system becomes unsustainable.

We relate heterogeneity in the system not only with different types of robots in the population, but also with different types of goals being pursued. This relation between heterogeneity and different goals is not present in all instances of multi-robot systems and different case studies. Moreover, it is possible that other types of heterogeneity in the population (for instance in systems considering division of labor) can actually improve performance and sustainability. However, we believe that this case study can represent a class of distributed robotic systems where a social (or moral) dilemma is considered by the robots, and where different types of robots take different decisions regarding social dilemmas. An institutional use of discount rates was crucial to model heterogeneity in the population and its impact on the sustainability of the system.

7 Conclusion and Future Work

A confluence of several scientific disciplines, technologies, and philosophical approaches are making possible a scenario of "metamorphosis of objects", in which human and non-human, natural and artificial creatures will build unprecedented relations of sociability. We have proposed in this paper to think this process as a horizon for Social Robotics, going beyond proximity sociability and considering institutional environments shared by humans and robots in many-to-many (partly) anonymous social relationships. Given the fact that institutions are the tool human societies use to deal with this kind of complex social interaction, we looked at leading economists and philosophers having dedicated sustained efforts to understand the workings and functions of social institutions in search of inspiration to our approach.

The reported experiments show what we have achieved in advancing the goal of designing robots that embody our institutional inspiration. Demonstrating experimentally the impact of Institutional Robotics in the naturalness and performance assessment of human-robot interaction, namely for the many-to-many case, is a challenging endeavor, given the complexity of the required experimental tests, involving several autonomous social robots and humans. Also, the development of concepts at the intersection of disciplines, on the one hand, and conducting experiments in robotics that use those concepts, on the other hand, can only be done by successive approximations.

Despite the difficulties, we plan to carry out real scale experiments with networked robot systems, composed of static sensors and mobile robots, interacting with people, within the frame of an ongoing EC FP7 project (MOnarCH). The robots will interact with children, staff, and visitors in the pediatric ward of an oncological hospital, playing several roles in edutainment activities (acting as school teaching assistants, playing interactively with children, and helping staff assistants to maintain the children in a socially interesting dynamics that improve their quality of life as inpatients). The hospital is a challenging environment, not only because it is a realistic scenario, but also because of the strict ethical regulations in force in such a place [29]. Integrating robot systems in such a mixed society with human beings will require having models (even if coarse) of human groups activity dynamics and composing them with robot task plans and institutions described as IACs, so as to ensure that robots will take human-aware decisions, using socially acceptable rules.

Our approach can help to understand how natural and artificial creatures (e.g. human beings and robots) can meaningfully share the same world, and the modalities of that sharing. We see Institutional Social Robotics as part of this future: use human-inspired institutions to control systems

of multiple robots within sophisticated social spaces shared by humans and machines, shaping many-to-many human-machines (partly anonymous) social interactions over long periods of time, where humans can interact with robots within large populations without having to abandon neither their informal day to day behaviour nor their habitual expectations in dealing with other human individuals.

Acknowledgments This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through grants SFRH/BPD/35862/2007 (first author) and SFRH/BD/33671/2009 as part of the Joint Doctoral Program IST-EPFL (second author), as well as by FCT ISR/IST Pluriannual funding through the PIDDAC program funds.

References

1. Albus JS (1991) Outline for a theory of intelligence. IEEE Trans Syst Man Cybern 21(3):473–509
2. Bayindir L, Sahin E (2007) A review of studies in swarm robotics. Turk J Electr Eng 15(2):115–147
3. Bonabeau E, Dorigo M, Theraulaz G (1999) Swarm intelligence: from natural to artificial systems. Oxford University Press, New York
4. Bosch A, Sunder S (2000) Tracking the invisible hand: convergence of double auctions to competitive equilibrium. Comput Econ 16(3):257–284
5. Cassandras C, Lafortune S (2008) Introduction to discrete event systems. Springer, Berlin
6. Chlistalla M (2011) High-frequency trading: better than its reputation? Research briefing, Deutsche Bank Research. http://bit.ly/oPggJ1. Accessed 20 Oct 2012
7. Coase RH (2007) The institutional structure of production (prize lecture–1991 nobel prize in economic sciences). In: Ménard C, Shirley MM (eds) Handbook of new institutional economics. Springer, Dordrecht, pp 31–39
8. Denzau A, North DC (1994) Shared mental models: ideologies and institutions. Kyklos 47(1):3–31
9. Duffy BR (2006) Fundamental issues in social robotics. Int Rev Inf Ethics 6:31–36
10. Durfee EH (2004) Challenges to scaling up agent coordination strategies. In: Wagner TA (ed) An application science for multi-agent systems. Kluwer Academic Publishers, Dordrecht, pp 113–132
11. Gode DK, Spear S, Sunder S (2004) Convergence of double auctions to pareto optimal allocations in the edgeworth box. Working paper no. 04–30, Yale International Center for Finance. http://ssrn.com/abstract=1280707. Accessed 15 Dec 2012
12. Gode DK, Sunder S (1993) Allocative efficiency of markets with zero-intelligence traders: market as a partial substitute for individual rationality. J Polit Econ 101(1):119–137
13. Gode DK, Sunder S (1997) What makes markets allocationally efficient? Q J Econ 112(2):603–630
14. Hodgson GM (2006) What are institutions? J Econ Issues XL(1):1–25
15. Ménard C, Shirley MM (2005) In: Ménard C, Shirley MM (eds) Handbook of new institutional economics. Springer, Dordrecht, pp 1–18
16. Millikan RG (1984) Language, thought, and other biological categories. The MIT Press, Cambridge
17. Millikan RG (1993) White queen psychology and other essays for Alice. The MIT Press, Cambridge, MA
18. Mondada F, Bonani M, Raemy X, Pugh J, Cianci C, Klaptocz A, Magnenat S, Zufferey JC, Floreano D, Martinoli A (2009) The e-puck, a robot designed for education in engineering. In: Proceedings of the 9th conference on autonomous robot systems and competitions, Castelo Branco, 2009, vol 1, pp 59–65
19. Mutlu B, Forlizzi J (2008) Robots in organizations: the role of workflow, social, and environmental factors in human–robot interaction. In: Proceedings of the 3rd ACM/IEEE international conference on human–robot interaction, Amsterdam, 12–11 March 2008. HRI '08. ACM, pp 287–294
20. North DC (1991) Institutions. J Econ Perspect 5(1):97–112
21. North DC (2005) Understanding the process of economic change. Princeton University Press, Princeton
22. Ostrom E (1990) Governing the commons. The evolution of institutions for collective action. Cambridge University Press, Cambridge
23. Ostrom E (2005) Understanding institutional diversity. Princeton University Press, Princeton
24. Pereira JN (2013) Advancing social interactions among robots: an institutional economics-based approach to distributed robotics systems. Phd thesis, IST-EPFL Joint Doctoral Initiative - Instituto Superior Técnico (IST), École Polytechnique Fédérale de Lausanne (EPFL)
25. Pereira JN, Silva P, Lima PU, Martinoli A (2013) An experimental study in wireless connectivity maintenance using up to 40 robots coordinated by an institutional robotics approach. In: Proceedings of the IEEE/RSJ international conference on intelligent robots and systems (IROS), pp 5073–5079
26. Pereira JN, Silva P, Lima PU, Martinoli A (2014) Formalization, implementation, and modeling of institutional controllers for distributed robotic systems. Artif Life 20(1):127–141
27. Searle JR (1995) The construction of social reality, 1996th edn. The Penguin Press, New York
28. Searle JR (2006) Social ontology: some basic principles. Anthropol Theory 6(1):12–29
29. Sequeira J, Lima P, Saffiotti A, Gonzalez-Pacheco V, Salichs MA (2013) Monarch: Multi-robot cognitive systems operating in hospitals. In: ICRA workshop "Crossing the reality gap from single to multi- to many robot systems", Karlsruhe
30. Silva P, Lima PU (2007) Institutional robotics. In: F.A. Costa, L.M. Rocha, E. Costa, I. Harvey, A. Coutinho (eds.) Advances in artificial life. Proceedings of the 9th European conference, ECAL 2007, Lecture notes in computer science, vol 4648. Springer, Lecture notes in computer science, vol 4648. Springer, Berlin, pp 595–604
31. Silva P, Ventura R, Lima PU (2008) Institutional environments. In: From agent theory to agent implementation, proceedings of workshop AT2AI-6, AAMAS 2008–7th international conference on autonomous agents and multiagent systems, pp 157–164
32. Sundmaeker H, Guillemin P, Friess P, Woelfflé S (eds.) (2010) Vision and Challenges for Realising the Internet of Things. Luxembourg: Publications Office of the European Union, Cluster of European Research Projects on the Internet of Things
33. Sutton RS, Barto AG (1998) Introduction to reinforcement learning. MIT Press, Cambridge
34. Uckelmann D, Harrison M, Michahelles F (eds) (2011) Architecting the internet of things. Springer, Berlin and Heidelberg
35. Vertesi J (2012) Seeing like a rover: visualization, embodiment, and interaction on the mars exploration rover mission. Soc Stud Sci 42(3):393–414
36. Visscher PK (2003) Dance language. Encyclopedia of insects. Academic Press, San Diego, pp 284–288

Porfírio Silva (Ph.D) received the Licenciatura (4 years) and M.A. degrees in Philosophy in 1990 and 1996, respectively, and the Ph.D. (2007) in Epistemology and Philosophy of Science at Faculdade de Letras, Universidade de Lisboa. Under the label "artificial societies", he researches the impact of recent robotic developments on human sociability, taking as a vantage point the significance of institutions in human societies. As a post-doctoral researcher at the Institute for Systems and Robotics (Lisboa) he contributed to a robotics' team developing a long term program based on a concept he introduced during his PhD: "Institutional Robotics", a novel framework for the coordination of robot teams sharing social spaces with humans, inspired by philosophical approaches to institutions and by Institutional Economics. Currently he is a collaborator to the Center for Philosophy of Science of the Universidade de Lisboa (CFCUL) and a participating researcher in the FP7 MOnarCH project about the deployment of a networked robotic system in the pediatric ward of an oncological hospital.

José N. Pereira (Ph.D.) received B.Sc. and M.Sc. degrees in Applied Mathematics and Computation and in Mathematics and Applications, respectively, at Instituto Superior Tcnico (IST), Universidade de Lisboa, in 2007. In 2014, he completed his Ph.D. in the scope of the IST-École Polytechnique Fédérale de Lausanne (EPFL) Joint Doctoral Initiative, being granted a joint degree issued by both universities (Electrical and Computer Engineering at IST, Manufacturing Systems and Robotics at EPFL). Currently, he is a post-doctoral researcher at the Distributed Intelligent Systems and Algorithms Laboratory at EPFL, leading the group's effort in the FP7 MOnarCH project about the deployment of a networked robotic system in the pediatric wing of a hospital. His research interests include distributed and networked robotic systems, in particular their coordination and cooperation with humans, swarm robotics and multi-agent systems.

Pedro U. Lima (Ph.D., Associate Professor) received the Licenciatura (5 years) and M.Sc degrees in Electrical and Computer Engineering at IST in 1984 and 1989, respectively, and the Ph.D. (1994) in Electrical Engineering at the Rensselaer Polytechnic Institute, NY, USA. Currently, he is a Professor at IST, Universidade de Lisboa, and a researcher of the Institute for Systems and Robotics, where he is the coordinator of the Intelligent Robots and Systems group. He is the co-author of two books, and member of the Editorial Board of the Elsevier's Journal of Robotics and Autonomous Systems. His research interests lie in the areas of discrete event models of robot tasks and planning under uncertainty, with applications to networked robot systems. Pedro Lima was a Trustee of the RoboCup Federation (2003-2012), and is currently the Coordinator of the FP7 Coordination Action RoCKIn, about robot competitions.