The “Interaction Engine”: a Common Pragmatic Competence across Linguistic and Non-Linguistic Interactions

Giovanni Pezzulo

Abstract—Recent research in cognitive psychology, neuroscience and robotics has widely explored the tight relations between language and action systems in primates. However, the link between the pragmatics of linguistic and non-linguistic interactions has received less attention up to now. In this article we argue that cognitive agents exploit the same cognitive processes and neural substrate—a general pragmatic competence—across linguistic and non-linguistic interactive contexts. Elaborating on Levinson’s idea of an “interaction engine” that permits to convey and recognize communicative intentions in both linguistic and non-linguistic interactions, we offer a computationally-guided analysis of pragmatic competence, suggesting that the core abilities required for successful linguistic interactions could derive from more primitive architectures for action control, non-linguistic interactions, and joint actions. Furthermore, we make the case for a novel, embodied approach to human-robot interaction and communication, in which the ability to carry on face-to-face communication develops in coordination with the pragmatic competence required for joint action.

Keywords: pragmatics, joint action, embodied cognition, internal model, mindreading, face-to-face communication, language

I. INTRODUCTION

The links between language and action system have recently received a lot of attention in cognitive psychology, neuroscience and robotics. Most studies have focused on two main issues. The first issue is how the meaning of words and symbols is grounded in action and perception [29], [112], [124]–an idea that has gained attention in recent cognitive neuroscience research [16], [106], [125]. The second issue is the passage from skills to language [4], [5], [6] and in particular how the ability of the action system to build sequences of motor primitives could be reused for acquiring syntax, compositional semantics and linguistic constructions, or structured mappings between form and meaning of sentences [51]; see also [122].

In this article we focus on a third issue that has received less attention in cognitive and developmental robotics: the relations between the pragmatics of linguistic and non-linguistic interactions. According to Morris [87, p. 35], pragmatics “concerns the relation of signs to their interpreters”; this definition is neutral with respect to linguistic and non-linguistic issues. A linguistically-oriented definition is that pragmatics “is the study of language usage” and is concerned solely with performance principles of language use, with a particular emphasis on the contextual aspects that influence interpretation [79].

These definitions emphasize a role of pragmatics in driving interpretations that go beyond the mere encoding and decoding of messages according to grammatical rules. The “encoding and decoding” model of communication is in fact insufficient for many reasons. First, there are many elements that are not fully specified in the utterance (a typical example is reference resolution, required to understand what “I” or “you” mean in a sentence). Second, it is often the case that utterances convey (additional) messages that go beyond their literal meaning, but that the speaker wants the hearer to infer: the so-called implicatures [64]. Third, linguistic communication is an important part of social life that goes well beyond the conveying of messages (being them literal or implicatures), and is part of a continuous activity of coordination or competition with others, and to influence their beliefs, goals, and actions. This is acknowledged in speech acts theory [8], [118]: together with the locutionary level (the verbal, syntactic and semantic aspects of utterances), a speech act can be equally described at the illocutionary level as conveying for instance directives (e.g., requests or commands) or commissives (e.g., promises), and consequently as having non-linguistic effects (the perlocutionary level), which typically correspond to an agent’s prior intention.

Grice [64] argues that pragmatics studies how to bridge the gap between the literal meaning of a sentence (sentence meaning) and what the speaker intends to convey (speaker meaning), including its non-linguistic intentions. The ability to understand another speaker’s intended meaning—essentially, a mindreading ability—is called pragmatic competence. Thanks to this competence, a hearer uses (i) utterances and linguistic knowledge (e.g., grammar and lexicon), and (ii) non-linguistic elements (the environmental and social context) as pieces of evidence about the speaker’s meaning and intentions. Non-linguistic factors are important in that interpretation is highly context-sensitive, and any facet of the situation in which a sentence is uttered can offer hints for its disambiguation.

The view that the pragmatics of action and language systems are interrelated is not novel in psycholinguistics. Levinson [80] has convincingly argued that a pre-linguistic interaction engine constitutes the pragmatic foundation that supports language and produces its universals. In this view, the pragmatic functions include multiple abilities: (communicative and extra-communicative) intention attribution, multimodal signaling (using gestures, facial expressions, etc.), mastery of the “give
and take of interaction” [80, p. 40], turn-taking, the use of functions such as denying, requesting, greeting, and ultimately the ability to engage in cooperative activities and to pursue common objectives. These interactional abilities do not derive from language but provide a foundation for it. Many other scholars, such as Wittgenstein [135], Austin [8], Grice [64], Leech [78], Barwise and Perry [19], Clark [39], Sperber and Wilson [120], Tomasello [128], and Bara [13], while having distinct (philosophical, linguistic or cognitive) perspectives, all have acknowledged the relations between language and action-based pragmatic competence.

Despite so, in practice the study of the two domains proceeds separately. As remarked in the Introduction, the issue of how action and language are related is nowadays popular in cognitive robotics, but the focus is principally on the grounding of words and on grammatical structures, and rarely on the pragmatic dimension. At the same time, non-linguistic interactions have been rarely addressed in linguistic studies of pragmatics, with the noteworthy exception is the analysis of pointing gestures in Kaplan [72]. Furthermore, the standard methodology used in computational linguistics consists in encoding pragmatic and semiotic information using linguistic structures rather than by relating them to the interaction abilities of situated agents [1]. In other words, an agent’s pragmatic competence is modeled as a fully linguistic competence (i.e., revolving around the ability to use words) rather than as an ability that is used also in non-linguistic interactive tasks, as we propose.

In this article we propose a framework for studying action and language pragmatics in a coherent way, by pointing at possible cognitive and computational mechanisms that are shared across them (we are principally concerned with the issue of addressing speakers intentions rather than other important issues in pragmatics such as reference resolution).

To this aim, in sec. II we propose theoretical arguments in favor of a common pragmatic competence across linguistic and non-linguistic interactions, and of a reuse of a pre-linguistic interaction engine [80] in linguistic domains. In sec. III we offer a computationally-oriented analysis of the interaction engine, highlighting a continuity of mechanisms across individual motor control, social interaction and linguistic communication. In particular, we emphasize the importance of joint action dynamics in communicative domains, linguistic and non-linguistic. Finally, in sec. IX we advance suggestions for a novel, embodied computational approach for the modeling of face-to-face communication and other linguistic phenomena, in which situated and linguistic abilities, including pragmatic competence, co-develop and are both ultimately grounded in sensorimotor action and interaction.

II. A COMMON PRAGMATIC COMPETENCE ACROSS LINGUISTIC AND NON-LINGUISTIC INTERACTIONS

As highlighted in the Introduction, context-sensitive intention understanding is not unique of language understanding, but it is also required in both non-linguistic (cooperative and competitive) interactions. This fact suggests two hypotheses, which we discuss below: (1) the pragmatic competence required in face-to-face communication and verbal exchanges is the same as the pragmatic competence of non-linguistic interaction, and (2) cognitive agents reuse the same processing and neural substrate across non-verbal and verbal communicative contexts.

A. The pragmatics of language are the same as the pragmatics of interaction

Non-linguistic communication has a rich pragmatic dimension. Indeed, even in non-linguistic contexts, humans (and other animals) use their own body and actions in semiotically salient ways, as it is highly advantageous for successful interaction and communication. Early (non-linguistic) examples of communicative implicatures of gestures are signaling requests with eye movements; indicating objects for communicating something about them; pointing at something or directing eye movements to attract one’s attention (or distracting it); performing movements or assuming facial expressions for showing (or hiding) one’s intentions; performing abrupt movements for signaling important events in the external world, or for signaling approval or disapproval; trying to persuade or intimidate by displaying body strength or related facial expressions; executing sequences of actions in an odd order or with a certain rhythm to capture attention or to be fun; signaling availability to engage in a social interaction by performing salutation or approval gestures, etc.

Communicative intentions such as requests, commands or greetings (among others) can be expressed both via conventionalized gesture or facial expressions, such as pointing with a finger, nodding or smiling, whose meaning is standard and easily decoded by the observer, or via standard actions. For instance, someone can break a nut to signal that he is strong, enraged, or that he has the possession of that object; someone can shut a door to signal that it was mistakenly left open by someone other; someone can try to lift a box to signal that it is heavy and to ask for help\(^1\). In this sense, Speech acts [8] such as asking a question, requesting help, challenging somebody, or informing somebody can be considered as linguistically-mediated versions of non-linguistic behaviors that express the same communicative intentions in non-linguistic interactions. Supporting this view is the fact that language and gesture are often used together, for instance to stress an argument, or interchangeably (e.g., in a noisy pub I can indicate you the presence of a friend by pointing at her rather than verbally; but if you cannot see me pointing, then I can shout), and the choice of linguistic or non-linguistic medium does not change meaning and pragmatics. For instance, pointing at an object is the same if performed manually, verbally (“look at that glass”), or both, which is often the case. Similarly, requesting help for lifting a heavy box is the same if the request is made verbally, or by displaying an awkward attempt to lift the object (interpreting the latter as a help-to-lift request means treating it as an indirect speech act [117]). Overall, in both linguistic and non-linguistic cases, utterances and actions are pieces of evidence about the performer’s (or speaker’s) meanings and intentions, and require that the observer (or hearer) has

\(^1\)The encoding of communicative intentions in standard actions is called behavioral implicit communication in [33], [34].
B. Reuse of non-linguistic pragmatic abilities in linguistic domains

Because pragmatic competence is the same in the two domains, the cognitive mechanisms for conveying non-linguistic communicative intentions through actions can be also used for conveying linguistic communicative intentions through speech acts (and, conversely, for recognizing them). Cognitive mechanisms cover specialized neural machinery for intention recognition and joint action, as well as the behavioral repertoire that supports it, such as the mechanisms for pointing at objects, regulating joint attention, assuming characteristic body poses and facial expressions. This view is alternative to the idea that linguistic competence is a prerequisite for mentalizing and mindreading abilities [32], or that specialized mechanisms are required for processing the pragmatics of language [121].

There are at least four reasons for an extensive linguistic reuse of the pragmatics of non-linguistic interaction. First, communication does not start with language and linguistic signs. Body movements, facial expressions and gestures have a rich semiosis (i.e., they are a rich system of signs that can carry on meaning) which is commonly used to convey communicative intentions during non-linguistic interactions. The semiotic dimension of body and action include elements, such as facial expressions, whose meaning is stable across cultures, as well as culture- or community-specific gesture, and even novel and extemporaneous actions that make sense only in a certain context; as such, it can serve a limitless set of communicative intentions, such as requests, greetings or commands. Because communicative intentions overlap across linguistic and non-linguistic tasks, the semiosis of body actions can be extensively used across non-linguistic, linguistic and mixed interactions (e.g., performing movements during dialogue).

Second, it has been convincingly argued that the cognitive infrastructure of communication is pre-linguistic [4], [47]. Even non-linguistic animals have specialized neural mechanisms for conveying and recognizing intentions in facial expressions and body actions, which form a core pre-linguistic pragmatic competence. In the next Sections, we will show that as most problems arising in linguistic and non-linguistic interactions are related at the computational model, they could be solved by similar mechanisms in the brain, such as for instance those for intention recognition and joint action.

Third, linguistic exchanges and face-to-face communication are typically situated in a rich sensorimotor context, which is largely the same across linguistic and non-linguistic interactions. Therefore, the mechanisms that have evolved to interact successfully within a shared, non-linguistic sensorimotor context (e.g., mechanisms for joint attention, the signaling of a request by pointing at an object or by directing eye movements) could be reused during linguistic interactions.

Fourth, face-to-face communication is essentially a joint action [39], [126]. As real-time coordination requires significant cognitive resources, such as the ability to monitor and anticipate another’s actions to be able to coordinate with it at some point, evolution could have favored the arise of automatic mechanisms of understanding and alignment of goals and behavior, and for achieving common goals [15], [22], [55], which in turn could have provided the foundations of equivalent abilities in linguistic domains.

In sum, we hypothesize that during evolution, linguistic communication reused the neuro-cognitive mechanisms (such as intention recognition mechanisms, mechanisms for orienting social attention, pointing at objects) originally developed by non-linguistic primates for conveying and recognizing non-verbal communicative intentions, and for achieving joint goals (see [3] for a discussion of neural reuse in the brain). Our view is in line with, and extends, Arbib’s [4] proposal that during evolution linguistic competence developed on top of the sophisticated action and gesture system of non-linguistic primates.

It is worth noting that our arguments on reuse apply at the evolutionary level only. If we focus on cognitive development, we observe that children co-develop the pragmatics of action and language. Many empirical studies indicate that the development of actions, gestures, and spoken words is highly interrelated at the neural and behavioral levels, such that that gesture and environmental constraints scaffold language acquisition, making language itself situated and inextricably linked to behavior; in turn, language skills permit the acquisition of richer pragmatic abilities. For instance, Bates...
[20] has emphasized the continuity between prelinguistic and linguistic signaling of communicative intentions in children (see also [30], [31], [60] for studies on pre- and paralinguistic communication). This process is made possible by the cognitive abilities inherited by our earlier ancestors, and facilitated by the inherently linguistic nature of children’s learning environments.

III. THE INTERACTION ENGINE: FIVE SCENARIOS AND THEIR COMPUTATIONAL REQUIREMENTS

The arguments discussed so far point at common mechanisms for processing the pragmatics of linguistic and non-linguistic interaction. However, it is unclear what the key constituents of the pre-linguistic interaction engine are, and how they can be (re)used in linguistic domains.

In the next Sections, by adopting the formalism of Bayesian systems for illustrative purposes, we offer a conceptual framework for answering these two questions. In particular, (1) we discuss which are the key constituents of such an interaction engine, and (2) we show that there is a “fil rouge” that links computational problems faced in non-linguistic and linguistic interactions. In the following we summarize our arguments:

A. Key constituents of the interaction engine

Although the interaction engine includes many sets of mechanisms (see the Introduction), three of them are particularly important for our analysis:

i mindreading mechanisms, which permit to estimate the cognitive variables (beliefs and intentions) of other agents;
ii mechanisms for achieving communicative goals, and to influence the cognitive variables of other agents;
iii mechanisms for sharing representations and optimizing joint goals or goals and actions of others (in addition to one’s own).

All three are fundamentally based on predictive processes and forward models (see [28], [66], [70], [96], [98], [136] for further analyses on the role of prediction for bridging sensorimotor action and higher cognition).

B. A “fil rouge” that links the interaction engine in non-linguistic and linguistic domains

These three sets of mechanisms do not only operate in linguistic domains. Rather, the two former mechanisms bear similarities with state estimation and action selection processes, respectively, in the individual control of action. The latter is characteristic of (linguistic and even non-linguistic) joint action scenarios, and is particularly important to understand the pragmatic dimension of linguistic exchanges, as face-to-face communication is essentially a joint action.

As resumed in Tab. I and Tab. II, we trace a parallel (at the computational level) between perceptual and planning problems in the individual domain, and the problems that “observer” and “actor” agents solve in social situations. We discuss how these problems become increasingly difficult, but not essentially different (at the computational level), in five scenarios of increasing complexity: individual, interactive but non communicative, interactive and communicative, joint action, and linguistic. In all these scenarios, we discuss the role of the three aforementioned key constituents of the interaction engine (marked as (i), (ii), and (iii) in Tab. I and Tab. II), also pointing to associated neuro-cognitive processes.

The computationally-guided analysis that we undergo in the next Sections supports our theoretical claim that the latter (linguistic) scenario could require the same interaction engine as used in the other (social but not linguistic) scenarios, which in turn are “socially-oriented” sophistications of the mechanisms for individual goal-directed action.

IV. INDIVIDUAL ACTION SCENARIO

A control-theoretic view of the architecture of goal-directed action is schematized in fig. 1 (see also [66], [69], [70], [93], [99], [100], [136]). Essentially, once an intention is formulated, it is realized by means of a combination of one or more pairs of forward and inverse models [138]. Inverse models take goal and current state as inputs, and generate a motor command so as to minimize the discrepancy between them. For example, if a goalkeeper is saving a penalty kick, the inverse model calculates the necessary body and hand movements from the current position to the expected ball position. Forward models take current state and an efferent copy of the executed motor command as inputs, and produce expectations relative to the next body and hand position, and the impact with the ball. (In this example, additional forward models can be in play that predict the ball trajectory.) A more complex example, which emphasizes hierarchical action organization [23], [67], is pushing a button for opening a box. Here opening the box represents the distal intention; if the agent knows that to open the box it is necessary to push a button, it can form the goal to push the button and then trigger the necessary internal models for doing so.

Fig. 2 illustrates a model of intentional action in the Dynamical Bayesian Network (DBN) formalism, a kind of graphical model that is especially suited for the modeling dynamic systems and (PO)MDP problems [88]. In this formalism, nodes indicate random variables (gray nodes are observables), with associated probability distributions over possible states (e.g., environmental states S, or agent’s actions A), and arrows indicate probabilistic relations between variables.

Our model is composed of five components: beliefs (B),
Computational mechanisms

(i) achieving goals relative to the environment

(ii) using language to achieve goals relative to the environment

Neuro-cognitive processes

(iii) joint action control (takes joint goal into consideration, uses shared representations)

TABLE I
FORMAL SIMILARITY OF PROBLEMS ACROSS INDIVIDUAL, INTERACTIVE, JOINT ACTION AND LINGUISTIC SCENARIOS: OBSERVER

<table>
<thead>
<tr>
<th>Task of perceptual processes</th>
<th>Computational mechanisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual scenario</td>
<td>(i) estimating the state of the observed system (i.e., hidden environmental variables)</td>
</tr>
<tr>
<td></td>
<td>Kalman filtering, particle filtering, Luenberger observer</td>
</tr>
<tr>
<td>Interaction scenario, non communicative aspects</td>
<td>(i) mindreading (estimating cognitive variables of another agent)</td>
</tr>
<tr>
<td></td>
<td>motor resonance, action simulation, emulation, action and intention understanding, inverse planning</td>
</tr>
<tr>
<td>Interaction scenario, communicative aspects</td>
<td>(i) mindreading for recognizing communicative intentions</td>
</tr>
<tr>
<td></td>
<td>the same mechanisms as above</td>
</tr>
<tr>
<td>Joint action scenario</td>
<td>(iii) formation of shared representations (SRs)</td>
</tr>
<tr>
<td></td>
<td>behavioral entrainment, mutual emulation, joint attention; the explicit goal of forming SRs</td>
</tr>
<tr>
<td>Linguistic scenario</td>
<td>(i) mindreading for recognizing communicative intentions in speech acts, (iii) formation of shared communicative context</td>
</tr>
<tr>
<td></td>
<td>language understanding as mental simulation, interactive alignment, mechanisms for maintaining reference</td>
</tr>
</tbody>
</table>

Fig. 2. Goal-directed action model in the Dynamical Bayesian Network (DBN) formalism [88]. Nodes indicate random variables (gray nodes are observable), and arrows indicate probabilistic relations between variables. The figure shows an agent’s beliefs (B), intentions (I), actions (A), unrolled in three time steps t, t + 1 and t + 2. The model includes also variables representing observations (O) and the state of the environment (S). The curved edge between the dotted squares is a shorthand for the relations $B_t \rightarrow B_{t+1}$, $I_t \rightarrow I_{t+1}$, $A_t \rightarrow A_{t+1}$.

intentions (I), actions (A), states (S), and observations (O). Beliefs, intentions and actions form the (hidden) cognitive variables of an agent.

$B$ represents the agent’s beliefs, which can include both perceptually-tied information and knowledge that goes beyond observation, such as the operating context (e.g., “we are in a room”). Encoded in the beliefs, and in the causal relationships between beliefs, intentions and action, there is also the agent’s general world (and task) knowledge. The agent’s beliefs are not necessarily a complete representation of the “true” state of the world, but can be partial and biased—this reflects an assumption of bounded resources (this is why in this model $B$ is distinct from $S$).

$I$ represents the agent’s intentions. In the next sections, we discuss three kinds of intentions: standard (praxic) intentions (i.e., non-communicative intentions that are aimed at changing the environmental dynamics), communicative intentions (i.e., intentions that are aimed at changing the hidden, cognitive variables of another agent), and joint intentions (i.e., intentions that are shared

See [50], [97], [101] for more details. Furthermore, note that, for the sake of simplicity, here we do not consider the value of actions; however, the model can be easily extended so as to include rewards associated with states, see [24], [102].
between two or more agents). Intentions can have abstract and distal content (e.g., becoming famous, having a successful conversation) or more concrete content (e.g., grasping a cup, informing you about something), and can be realized by means of multiple courses of actions, possibly extended over time.

- A represents the agent’s repertoire of actions, which have performative effects and change the state S in various ways. In this paper we discuss two kinds of actions: standard (praxic) actions (e.g., opening a window) and communicative actions (limiting ourselves to the illocutionary level of speech acts, e.g., asking what time it is). We assume that actions are not simply movements, but have action goals; this also implies that actions can be multiply realized by different patterns of movements (conversely, each action can realize more than one action goal). This is in agreement with the view that the brain encodes a (hierarchically organized) “motor vocabulary” of basic actions, which are goal-directed (in the sense that they encode implicitly a transition \( A \rightarrow s_{goal} \)) and can be flexibly recombined.

- S represents the (world and agent) state, which can encode the external world, the robot state, as well as another robot’s state, etc. Note that the “true” S is hidden (or latent), in the sense that an agent cannot fully access it, but only estimate it through its observations.

- O is a function of the observable and accessible subset of S. It can include, for instance, features or objects in the environment, the observable agent movements, etc. (This definition entails that there are parts of the state S that are not directly observable, but must be inferred.)

This formulation highlights the goal-directed nature of choice and action control. First, in keeping with the philosophical idea of practical reasoning [26], intentions depend on beliefs and actions depend on beliefs and intentions. Second, the architecture includes inverse and forward models \((P(A|B, I) \text{ and } P(S_{t+1}|A, S_t))\), respectively, which are essential elements of goal-directed, model-based instrumental controllers (as opposed to model-free controllers), see [12, 90].

The agent shown in Fig. 2 faces two fundamental problems: updating its beliefs through perception, and realizing its intentions by planning and acting. These operations can be described at the computational level as Bayesian inferences, which essentially calculate probability distributions over the relevant variables based on estimation of other variables that are causally related to them in the DBN [21].

**A. State estimation: estimating hidden environmental variables**

An agent’s observations O can only partially and approximately correspond to the “true” state of the environment S, which is not directly observable. This is due to the limited resources of the agent, to the fact that certain characteristics of the environment are not directly observable (e.g., due to occlusions), and to noise and uncertainty that are intrinsic in any measurement. Acting on the basis of incorrect knowledge of the current state S can be deleterious for the agent, since its actions could produce results that are different from those intended. For this reason, it is valuable for the agent to try estimating S instead than simply observing O. In generative models hidden states can be estimated based on observations by computing the posterior probability \( P(S|O) \) using the standard Bayes rule. Since S changes over time with (partially unknown) dynamics \((s_t \rightarrow s_{t+1} \rightarrow s_{t+n})\), recursive estimation methods such as Kalman filtering [71] or particle filtering [52] can be used to estimate them by optimally combining measurements and predictions generated by the forward model. Note that in this perspective an agent’s perception and state estimation are not passive processes, but consist in the proactive modeling of the changes in the environment, both as an effect of one’s own actions, and of external events.

An important part of state estimation is affordances recognition, or the recognition of which action possibilities are currently more likely given the environment. This problem has recently attracted attention in cognitive science after the discovery that specific motor programs, such as for instance for grasping a cup or a dice, are automatically elicited in the mere presence of appropriate objects [130]; the putative brain substrate for this mechanism are canonical neurons in the human and monkey premotor cortex [109]. From a computational viewpoint, affordances can be considered as enhancing the prior probability \( P(A) \) of certain actions in the mere presence of objects [85]. In this way, affordance recognition biases the planning process, because they make the afforded actions more likely to be selected and more ready to be executed.

**B. Action planning and execution: changing environmental dynamics**

In this computational framework, planning can be described in terms of probabilistic inference of which action (or sequence of actions) could realize a desired goal [7], [24], [129] (often, this is cast as an optimization process so that the action or action sequence is optimal). This can be done by conditioning (i.e., fixing) the values of B and I or \( S_{goal} \), so that the value of \( [A_I, A_{I+n}] \) can be inferred. Intuitively, this process consists in recognizing the present and imagining the intended future so to derive the actions that achieve it. Formally, this can be done by computing the MAP (maximum posteriori probability) action.

---

4For the sake of simplicity, here we assume that intentions can be always mapped into a (future) goal state \( s_{goal} \) to be realized at time \( t_1 \) (or within a time interval). In other words, we do not address the problem of decomposition, and implicitly assume that any intention (e.g., becoming famous) is associated to one or more goal states that satisfy it (e.g., my picture in the Time magazine). Thus, goals are treated as end states that satisfy an intention.

5The fact that intentions can govern action sequences is implicitly represented by the link \( I_1 \rightarrow I_{t+1} \).

6Note that different from many AI architectures [107], [111] beliefs are grounded in sensorimotor experience, as they are produced by an active estimation of the environment (through the relations between S and B), and afford hypothesis testing (via the relations between B and A).

7A simpler alternative to Kalman filtering is using a Luenberger observer, or building an “observer” that assumes a similar form to the observed system dynamics so as to reconstruct all the state variables, being them observable or not [81]. Once an observer is built, the discrepancy between the observed system and the observer is feed back to the observer to adjust its parameters and output.
sequence conditioned on achieving the goal (in a time horizon $T$) [7], as well as in many other ways [21]. In control theoretic terms, this is equivalent to measuring the discrepancy between the estimated (current) and desired (goal) states of the world, and select actions that reduce it [84]. Successively, during execution, this procedure can be repeated, and so action plans can be refined so as to deal with uncertainty and motor errors until the goal state is reached.

V. Interaction Scenario

The machinery of individual goal-directed action can be extended to social scenarios with two (or more) interacting agents, in which the goals are for instance helping, hindering or imitating another agent. For the sake of simplicity, here we discuss the case of two actors: agent 1 and agent 2, playing the (interchangeable) role of “actor” and “observer”, respectively. (Where appropriate, we distinguish their variables with numbers, too; for example, agent 1’s actions are referred as $A^1$, and agent 2’s actions as $A^2$.) Agent 1 and agent 2 dwell in the same environment, and can perform actions that affect both the external world and the other agent, such as moving an object where the other agent can see or reach it, or communicating with the other agent.

Fig. 3 illustrates this actor-observer scenario. Contrary to the individual scenario, here an agent’s observations do not only include environmental dynamics and the effects of its own actions, but also the actions executed by another agent. Here we discuss how the aforementioned mechanisms that permit to estimate hidden environmental variables, predict and plan individual actions, are reused and extended to support the actions of social actions when they play the role of observers and actors (these roles are considered here for illustrative purposes, in that agents can act simultaneously). In particular, internal models (inverse and forward) used in individual action planning and control are reused for planning social actions and understanding actions performed by others [58], [136], providing key functionalities to the interaction engine.

A. Estimating another’s (hidden) cognitive variables: mindreading

Analogous to the necessity of estimating hidden environmental variables, a social agent benefits from estimating the (hidden) cognitive variables of other agents: the beliefs, intentions and actions that generate their observable behavior (i.e., $P(A^2|B, I)$, $P(I^2|B)$, and $P(B^2|S)$). We generically refer to the estimation of hidden variables of other agents as a mindreading ability. In our framework, mindreading abilities form the core of the interaction engine, and play a significant role in all the cognitively-mediated social actions, communicative and non-communicative.

Recent advances in neuroscience suggest that mindreading abilities are explained by subpersonal mechanisms, which recruit (automatically) the motor system, and realize a basic and almost effortless ‘motor understanding’ of behavior and intentions of others, which does not rely on explicit (belief and intention) representations. It has been proposed that primates are automatically “attuned” to the goals of their conspecifics via mirror neurons [57], [110]. This form of motor resonance provides understanding of observed actions in terms of one’s own motor repertoire. In addition, this mechanism could produce automatic alignment of behavior during interaction, by enhancing the prior probability $P(A)$ of one’s own actions after seeing the performance of others executing the same action. A related but distinct proposal is that action understanding is mediated by a predictive mechanism of motor simulation: inverse and forward models used in individual action control are fed with sensory inputs derived from the observation of another acting agent, and run “in simulation” so to derive what action would have more likely produced the same sensory effects [50], [70], [136]. Action simulation could provide action and intention understanding, by reactivating the motor codes necessary for producing the same results. Essentially, the observer takes the actor’s observed movements as input, and “simulates” which action in its own motor repertoire could have produced them; then, it attributes this action, and the corresponding goal, to the actor. In computational terms, it calculates $P(A^2)$ by running $P(S_{t+1}^1|A^1, S_t)$, and comparing predicted and actual state ($s^c_{t+1}$ and $S_{t+1}$, respectively). The action which produces less discrepancy between $s^c_{t+1}$ and $S_{t+1}$ is then interpreted as the observed action. A second possible role for action simulation is facilitating perceptual processing [66], [134]. This mechanism is analogous to Kalman filtering. Briefly, the forward model receives perceptual stimuli relative to the observed action, and runs so to produce the same perceptual predictions that it would have produced if the observer had actually run the same action. In other words, agent 1 can better estimate the observed movements of agent 2 (that are part of $P(S_{t+1})$) by running $P(S_{t+1}^1|A^1, S_t)$ (under the assumption that agent 2 is running the same action $A$ as its own).

Automatic mechanisms could not be sufficient for human-like theory of mind, but it might act in concert with more
sophisticated rule-based inferential processes, which also take into account explicit knowledge of other individuals, and how they behave in social situations; these mechanisms are often called “mentalizing” or employing a theory of mind [2], [25], [43], [54], [46], [116]. To perform these inferences, it has been proposed that humans adopt a form of teleological reasoning, which provides explanations and predictions of other’s actions based on a rationality principle (i.e., by assuming that agents tend to select the most rational actions given their constraints and goals [61]). Knowledge of other agents also includes social scripts such as “in this situation people behave so-and-so”. From a computational viewpoint, this mechanism has been explained as an inverse planning, or an action-to-intention inference, with the aim to derive which intention could have generated the choice of the observed action. Essentially, the inference $P(I^2|A^2)$ is done via one’s own explicit knowledge (in B), not via the motor system [11], [44].

Note that the presence of other individuals influences an agent’s state estimation process, by creating social affordances [74], [83], or novel possibilities for action that depend on the actions of other individuals (e.g., somebody can move an object where I can reach it, or faraway from me) or in certain cases on their mere presence (e.g., in the presence of another individual, a weighty object can be perceived as lighter). Therefore, the motor facilitation that we have discussed earlier as an enhancement of the prior probability of actions can be modulated by the presence of other individuals.

B. Social actions: helping, hindering, and imitating others

Mindreading abilities are useful for both cooperation and competition. For instance, in many cases intention recognition is necessary in order to coordinate, as well as for helping a friend or hindering an adversary. Once an agent is able to mindread another agent, it can more advantageously coordinate with it, calculate the effects of its possible actions on it, or know what are the complementary actions that it can take (specialized neurocognitive mechanisms for the latter abilities have been studied in [89]). In addition, mindreading abilities permit to reuse the aforementioned planning abilities for planning actions that help or hinder others. One way to plan helping actions is setting a goal $G^2$ of another agent as one’s own intention, and reusing the same machinery of individual action control for optimizing its achievement (with the possible use of constraints, such as for instance that $G^2$ does not collide with one’s own goals). Hindering actions can be planned in a similar way, by intending the contrary of another’s goal, or an inconsistent goal [42].

Another relevant kind of social action is imitation. In this framework, imitation is a straightforward consequence of action and intention recognition via action simulation. While in action simulation the motor outputs are inhibited, for imitation to take place it is necessary that once an observed action is recognized in terms of one’s own actions (and associated goals), then the necessary motor codes for performing them are activated [48].

VI. Communicative scenario

The third scenario extends the second one and addresses non-linguistic communication and communicative intentions. Many animals learn basic forms of communication by trial and error, or inherit them genetically. However, we are interested in more advanced forms of communication in which actions are intentionally used as signs to communicate with their conspecifics. In our perspective, communicative and non-communicative intentions and actions have a largely similar structure and can be modeled using the same architecture as in Fig. 2. The most important difference is that the objective of performing communicative actions and speech acts is changing the observer’s cognitive (hidden) variables rather than its behavior.

Communicative intentions can be realized without awareness of the receiver to be the addressee of a message. However, this only works well when the content of the communicative intention is easy to infer based on the practical action performed (for instance, if you notice that I take my glasses, you can easily infer that I use glasses for reading, and I can use this act for informing you without you noticing that I am doing so). In more complex, especially cooperative scenarios it is often useful to recognize that the other agent is intentionally communicating (rather than only acting), and what is its communicative intention.

A prerequisite for these more complex forms of communication is that both observer and actor assume an intentional stance toward each other [49], and recognize each other as capable of mindreading. In addition, both need the pragmatic competence to process the pragmatics of actions (body movements, facial expressions) so as to recognize or convey communicative intentions, respectively. An observer needs to use intention recognition mechanisms to decode the actor’s intentions, including its communication intentions. An actor, in turn, needs to infer what is the most likely interpretation that the observer will extract from his actions, what is the best action to make the observer believe something, etc.; this ability has been called recipient design [114] and is based on intention recognition as well.

A. Using mindreading for (communicative) intention recognition

Decoding the pragmatics of action is not trivial, since the same action can have opposite communicative intentions; for instance, giving an object can express gratitude or prostration. In addition, communicative intentions are often context-dependent. For instance, if two alpinists face a difficult crag, and one of them points at it, this can mean “let’s climb it” or “it is too difficult” depending on their common knowledge of their competence, or courage. Or, if they face two possible
paths, and one of them points at the left one, it can mean “let’s go there” if there is the peak summit, or “let’s avoid it” if there is a precipice.

Despite this complexity, still from a computational viewpoint inferring communicative intentions is equivalent to inferring standard intentions. The same mechanisms of action simulation and inverse planning as described in sec. V-A are good candidates for decoding communicative intentions, too. The decoding process is facilitated by stereotyped social signals, such as the direction of gaze of the actor or its facial expressions, which are associated with communicative intentions in a less ambiguous way.

B. Planning communicative intentions and recipient design

A hallmark of social action is that, because the actor knows that the observer has hidden cognitive variables, it can intentionally try to influence and change them. This is indeed one of the main purposes of communication (being it recognized by the observer, or not). Influencing the beliefs, intentions and actions of another agent is typically more efficient than only affecting its behavior; for instance, informing a child that red mushrooms are poisonous is (should be) more effective than stopping it each time it picks up a red mushroom. This influence is represented by the dotted edges in fig. 3. This means that agent 1 can intentionally plan an action aimed to change agent 2’s B, I, and A, and can predict the effects of its actions in terms of modifications of agent 2’s B, I, and A, as well as of S9.

Whether or not the actor wants the observer to believe that his action has communicative purposes, the problem of deciding what action can better fulfill the communicative intention is called recipient design. It can be hypothesized that the same mindreading mechanisms supporting intention recognition can be adopted by speakers to decide what communicative action could better achieve its communicative intentions. Support for this view comes from a recent neuroimaging study [91], which shows that similar brain areas are activated when subjects recognize intentions and when they predict the intention recognition process of a co-actor.

Besides the use of mindreading mechanisms, planning communicative actions is similar to planning performative actions in that a communicative intention is first selected, and then the actions that realize it with higher probability are selected and actuated with the aid of internal models. For example, the actor can plan actions that maximize the probability that the observer will believe a certain fact (i.e., ∈ A2), do a certain action (i.e., ∈ A2), or infers the communicative intention correctly (i.e., that he infers ∈ I1 with low uncertainty) using the computational methods discussed sec. IV-B. There is a wide range of strategies that can be used to influence beliefs, intentions and actions of others, which include for instance directly asking or informing them, selecting actions that strongly suggest complementary actions to be taken (or capitalizing on automatic processes of mutual alignment of behavior), and orienting another’s attention towards certain facts that afford desired reactions.

VII. Joint action scenario

A joint action scenario is a sophistication of the basic social scenario, in which two or more co-actors coordinate their actions in space and time to bring about a common goal [27], [119]. Guiding together a canoe and walking in a crowded street illustrate well the difference between joint actions and simpler forms of interaction, respectively. Although in both cases each agent adjusts its actions by taking into account the predicted consequences of the actions of the other agent, only in the former case there is a common goal that gives additional constraints to planning.

Some aspects of joint action could be supported by automatic processes of alignment of behavior; one example is the implicit coordination of hand clapping in a concert. These automatic mechanisms can be treated as enhancing the prior probability P(A) of selecting certain actions. However, more complex forms of joint action, such as for instance playing together in an orchestra, or assembling together a construction, require a sophistication of the computational model that we have introduced so far. Below we discuss the two novel features of the joint architecture: the presence of shared representations, and the achievement of joint intentions (or even of another’s intentions), in addition to one’s own.

A. Sharing representations: automatic and intentional processes

In complex joint action tasks, not only behavior, but also representations of two (or more) agents, such as for example goal and task representations, become aligned. The aligned subsets of beliefs and intentions of two (or more) agents has been called shared representations (SR) in the psychological literature [74], [119]. The idea of SRs stems from the concept of common ground, which explains how the success of communication depends on mutual knowledge, beliefs and assumptions, and permit to coordinate content (for instance, establishing common reference to an object) and behavior (for instance, deciding when to talk) [40], [39]. Common ground includes both background knowledge that is shared before the interaction (for instance, as an effect of a common culture) and knowledge that develops in the course of the joint action (for instance, by making reference or putting attention to the same object)\(^{10}\).

Shared representations can be formed unintentionally or intentionally, and through at least three mechanisms: entrainment, motor simulation and signaling strategies [74], [97]. There is ample evidence of automatic forms of alignments and mutual imitation of behavior and facial expressions during interaction (and possibly part of the cognitive structure that produces the aligned behavior), as well as emotion matching; an example is the so-called chameleon effect [35]. A related

\(^{10}\) Although the term “shared representations” suggests elsewhere, not all the SRs have to be represented explicitly; some elements, such as the table we are referring to, can be off-loaded in the external environment and accessed when needed, for instance by orienting attention.
mechanism is the (mutual) priming and percolation of linguistic representations in the interactive alignment model of dialogue [59], [103] (see also [55] for the relevance of bottom-up processes in social cognition). These mechanisms, which are akin to motor resonance, automatically entrain behavior and simplify prediction and mutual understanding. In addition, part of the automatic activation of SRs is explained by memories of past interactions and social conventions, whose discussion is beyond the scope of this paper. A second example of automatic formation of shared representations is through motor simulation: experimental studies reveal that people engaged in joint action can use their own internal models to predict actions of others, and their timing so as to better coordinate with them [73], [134].

Shared representations can be formed and maintained intentionally, too. Some examples are agents maintaining a coherent conversational context, orienting joint attention toward one object in the environment, agreeing on the common intention to move a table together, or on the more specific goal of moving it toward the left or the right. In keeping with the idea that intentional action selection (in individual and social scenarios) considers the costs of actions, pursuing the goal of forming SRs needs not to be altruistic but can have long-term advantages that surpass their short-term costs. Often, the intentional formation of SRs involves the use of communication and signaling procedures, such as stressing sentences or changes of context (e.g., in conversations), pointing at relevant objects or attracting another’s attention to the elements that should be part of the SR, repeating expressions, etc. This is not necessarily done with conventional (or linguistic) communication, but can consist in producing perceptual cues for the other agents, or even performing actions that constrain another’s actions. As revealed by a recent study [101], subjects use signaling actions as part of a wider strategy that is aimed at influencing another’s cognitive variables, but only when this is functional to the joint action.

Irrespective of the fact that representations are shared intentionally or unintentionally, the benefits in terms of interaction success are the same. Indeed, agents that share representations can more easily predict each other, and make their actions and the communicative intentions behind them, easier to detect; in sum, alignment of representations favors the unfolding of smooth interactions and diminishes the cognitive load required to perform predictions and mindreading. As shown in Fig. 4, in our computational framework SRs can be described as the overlap of a subset of agents and/or a subset of agents and/or a subset of actions. Thanks to this alignment, for instance, at time \( t \) agent 1 can predict the behavior of agent 2 by computing \( P(A^2_{t+1}|A^1_t, B^{sr}, I^{sr}) \) rather than firstly inferring \( B^2 \) and \( I^2 \) and then computing \( P(A^2_{t+1}|A^1_t, B^2, I^2) \). Because agent 1 already knows \( A^1_t, B^{sr}, \) and \( I^{sr} \), it does not need to estimate them, and thus its cognitive operations are facilitated. In turn, for this strategy to work, it is necessary that the two agents strive to remain predictable (based on SR) and explicitly signal violations of the SR; hence the necessity of implementing the intentional strategies for maintaining the SR, including signaling actions, which we have described so far.

\[ I^{sr} = \text{intentional} \]

\[ B^{sr} = \text{behavior} \]

\[ A^{sr} = \text{actions} \]

**B. Joint action control: achieving joint intentions**

Besides the formation of shared representations, joint action requires an accurate coordination of two mutually interacting planning processes (in the mind of two distinct agents). For this reason, in joint action, the objective of action planning is slightly different than in individual action control: the common intention \( I^{sr} \) has to be taken into consideration, and the goals of the other agent that are functional to the common goal have to be facilitated or at least not hindered. Then, in joint action the optimal choice for both agents is maximizing the success of the whole plan rather than one’s own part of it, as it would be natural in the individualistic perspective\(^{11}\).

In terms of our computational framework, this maps into a control strategy, which we can call joint action control, which aims at maximizing the success of the whole interaction rather than the success of one’s own part [97], [101]. A related perspective is that of teamwork, or the idea that agents see themselves as part of a team when they consider what to optimize during cooperative tasks [9]. The same computational mechanisms of individual planning as described in sec. IV-B can be used, such as the computation of a (MAP) action sequence conditioned on achieving the selected goal(s), except that the joint goal and the predicted behavior of other actors have to be considered as additional constraints in this process. It is worth noting that, because another’s sub-goals act as constraints, this method naturally leads to helping actions during the interaction.

A second aspect of this strategy is the intentional maintenance of SR, along the lines that we have discussed above. In particular, joint action control prescribes that, to obtain long-term benefits, agents perform signaling actions (only) when the co-actor cannot use the SR for predicting. In terms of our model, a signaling action follows the same rules as communicative actions (described in sec. VI-B), except that the objective is changing the content of SR, rather than of beliefs and intentions in general. Similar considerations about

\(^{11}\text{Similar considerations can be done for the optimization of other ingredients of the interaction, such as the minimization of joint rather than individual effort, timing and costs.}\)
what (signaling) actions could better change SR according to the performer’s intentions apply, too (see [101] for evidence that, when planning signaling actions, performer agents take the observer’s uncertainty into consideration).

In sum, the joint action control strategy is distinct from all the other forms of individual and social strategies, in that signaling actions, helping actions, and actions that fix the SR, which would be non-optimal under only individualistic constraints, can be all modeled within an optimal control framework whose aim is maximizing the success of the whole joint task.

VIII. LINGUISTIC SCENARIO: FACE-TO-FACE COMMUNICATION

Throughout the paper we illustrated the computational mechanisms that solve interaction problems and give agents a pragmatic competence to succeed in interactive scenarios. In this Section, we analyze the same interaction engine at work in a linguistic domain: face-to-face communication.

The first aim of this Section is showing that, in keeping with the idea that there is a single pragmatic competence, the functioning of the interaction engine is not significantly different in non-linguistic and linguistic domains, and the cognitive mechanisms for conveying and recognizing communicative intentions in speech acts, and forming shared linguistic representations, are essentially the same as those that pursue equivalent interactive goals in non-linguistic domains.

The second aim of this Section is analyzing the dynamics of interaction in terms of the computational framework that we have proposed. Our analysis stems from the idea that conversation is a form of joint action [39] that mainly uses speech and gesture, in which two (or more) agents pursue the common goal of having a successful communication in addition to (possibly non-overlapping) individualistic goals, such as informing somebody about something, convincing somebody, being informed, or make a good impression. In principle, interaction problems can be solved by a bounded form of recursive mindreading [139]. However, given that this mechanism is very demanding in terms of cognitive resources, it is still unclear why interactions are apparently so simple.

There have been many theoretical proposals concerning this point. Grice [64] proposes that a key role in the inferential process is played by conversational maxims and a general rationality principle. Sperber and Wilson [121] emphasize that communication capitalizes on the (automatic) ability of listeners to grab relevant messages, to orient attention and inference towards the relevant aspects of the dialogue, so as to narrow the set of possible interpretations (although they express skepticism that the standard mindreading equipment is sufficient to deal with the complexities of linguistic tasks). Not only this ability is useful from the listener’s viewpoint, but also from the speaker’s. They propose that the planning of communicative intentions is simplified by the ability of speakers to make reliable considerations of relevance (for the listeners): for instance, the ability to predict what will attract the attention of listeners, and what they will infer, can be leveraged to manipulating their mental states. In this framework, a “speaker who wants to make her utterance as easy as possible to understand should formulate it (within the limits of her abilities and preferences) in such a way that the first interpretation to satisfy the hearer’s expectations of relevance is the one she intended to convey” [121, p. 19]. Yet another facilitating aspects is highlighted by Pickering and Garrod [59], who propose that interaction dynamics produce an “interactive alignment” of linguistic representations and situation models, which in turn greatly facilitates mutual understanding and predictability without the need to maintain separate representations for one’s own and the other agent. In addition to automatic mechanisms, people adopt coordination smoothers, such as exaggeration of movements and selection of actions with low variability, for facilitating another’s action prediction and recognition [132].

Our analysis complements these proposals by stressing the crucial role of joint action control and the goal of remaining predictable, as elucidated in sec. VII-B. We propose that joint action dynamics, and in particular the formation of shared representations, facilitate the planning and recognition of communicative intentions by driving considerations of what is relevant, what should be expected, and what should be communicated (shared) during the interaction.

By discussing two sample dialogues, we make the point that (1) linguistic and non-linguistic elements with the same pragmatics are used, and (2) interactive dynamics can be described in terms of the interaction engine illustrated in the preceding Sections, giving rise to a joint action control strategy.

A. Joint action control in play during linguistic exchanges: a cooperative case

What follows is a short dialogue between Bruce Wayne, who is in reality Batman, and Lucius Fox, who is CEO of the Wayne Enterprises, a multinational owned by Bruce Wayne, and knows the secret activities of his employer (from The Dark Knight movie, 2008).

Bruce Wayne: I need a new suit.
Lucius Fox: Yeah, three buttons is a little ’90’s, Mr. Wayne.
Bruce Wayne: I’m not talking fashion, Mr. Fox, so much as function.
[hands him a diagram]
Lucius Fox: You want to be able to turn your head.
Bruce Wayne: Sure would make backing out of the driveway easier.
Lucius Fox: I will see what I can do.

1) The interactive engine uses speech and motor acts on purpose: Bruce wants Lucius to understand what he needs. Although his behavior can be described as a standard intention-al action selection process (see below), it is worth noting that in the first line Bruce selects a speech act (a request: asking Lucius a new suit), and in the third line he selects a non-linguistic act that achieves the same goal (handing a diagram). Not only this example shows that Bruce’s planning process is highly flexible and can use linguistic and non-linguistic means, but it also indicates that the selection is not causal but tied to the specific needs of the interaction. In fact, the initial speech act is more efficacious for setting the context of the interaction
(talking of a new suit), but at the same time it is ambiguous, because it does not describe exactly what is Bruce’s goal. The successive use of a diagram is more efficacious for conveying the goal of having a lighter and more efficient suit, whose specific features would have been harder to describe verbally.

A second indication that the pragmatics of linguistic and non-linguistic elements are treated equally is the fact that Lucius is able to extract the communicative and non-communicative intentions of Bruce from both the speech act and the diagram. After the first speech act, we can assume that Lucius correctly understands that Bruce is talking of a new Batman suit rather than the classy suit that Bruce is wearing (although his reply is ironic). After seeing the diagram, Lucius correctly infers that Bruce wants the new suite because he needs novel features, such as for instance turning the head. In sum, not only the planning process but also intention recognition mechanisms can process linguistic and non-linguistic elements flexibly.

Moreover, in this dialogue the interaction engine manifests itself also in the non-linguistic elements, such as gestures, facial expressions and eye movements. For instance, handing the diagram is an implicit request that is done not verbally. At the end of the dialogue, Lucius signals that he understands correctly by both giving a reassuring verbal feedback (“I will see what I can do”) and an equally expressive facial expression.

Finally, parallel to the explicit dialogue there is an implicit exchange of messages that is all non-verbal, which revolves around the fact that Bruce is ready to take more risks with the new, lighter suit. Lucius statement (“You want to be able to turn your head”) expresses acknowledgement of this fact, and Bruce’s ironic answer (“Sure would make backing out of the driveway easier”) is intended to play down and divert attention from the risky aspects of Batman’s activity. If these communicative intentions were expressed verbally, the whole situation would have been more embarrassing and demanding in terms of cognitive resources.

It emerges from our analysis that the use of linguistic and non-linguistic elements does not correspond to different aims or different strategies; rather, their selection (and recognition) is part of a planning process that flexibly selects the best means to achieve communicative and non-communicative intentions. Thus, processing linguistic and non-linguistic elements requires the same pragmatic competence. Hereby we describe this pragmatic competence in terms of the interaction engine elucidated earlier.

2) The interaction engine and joint action control in play: In this dialogue, Bruce has a non-communicative intention (having a novel, lighter and more efficient suit), which entails a communicative intention (communicating the goal) as subgoal. To achieve his goals, Bruce uses the action planning and selection methods described in sec. IV-B; in addition, in keeping with feedback control strategies, Bruce monitors the interaction and uses feedback from Lucius’ replies to modify his plan until his goals are fulfilled (here the goals are defined as termination states like “Lucius expresses understanding of Bruce’s intention”, or “Lucius shows his willingness to realize the suit”).

In turn, Lucius can use mindreading abilities described in sec. V-A to reverse this process and recognize the speech acts, communicative and non-communicative intentions of Bruce. (We can assume that Lucius has the goal to correctly understand and follow Bruce’s instructions, thus his mindreading process is functional to his individualistic goals, not just the reaction to Bruce’s actions.) Of course, his mindreading process can fail or be partial, and indeed the initial ironic reply of Lucius could mean that he is unsure of what Bruce exactly wants. Understanding what Lucius means requires going well beyond the literal meaning of his utterance, but Bruce’s pragmatic competence permits him to easily understand what is wrong. In this way, he can take corrective actions until his goals are achieved; again, this can be described within standard feedback control as the optimization of an individual plan with feedback from another’s actions.

However, there are many elements that suggest that the two actors are not only pursuing their individual goals, but also interactive goals. First, Lucius actively gives feedback about his understanding. For each turn, he selects speech acts that signal to Bruce that his goals are or are not satisfied. As discussed in [40], turns in dialogues achieve two objectives: first, they carry on the interactive process, and second, they give evidence about the speaker’s understanding of the content of the preceding interaction. As we have discussed, a particularly useful strategy for conveying communicative intentions making one’s own utterances surprising given the SR. After his first sentence Bruce expects an appropriate reply (for instance, a confirmation that the novel suit will be more solid or lighter). By selecting a surprising and off-topic answer, Lucius conveys the communicative intention of knowing more. The reply is not casual, but it is intended to highlight what Lucius has correctly understood (that they are talking of a suit) and what he does not yet knows (what the new suit should be like). In terms of our model, this means that in his individual plan Lucius is explicitly selecting actions that are less ambiguous for Bruce.

Second, Bruce actively guides Lucius’ predictions, inferences and considerations of relevance by building SRs, and at the same time he monitors the content of SR to ensure that Lucius will not perform incorrect inferences. Bruce’s first sentence sets “the suit” as the main dialogue topic. Bruce is parsimonious in his choice of words in that he (incorrectly) predicts that Lucius will easily infer that the dialogue revolves around a more effective suit (or maybe his short sentence is aimed at creating suspense). As the (ironic) response of Lucius revolves around ‘fashion’ instead, Bruce formulates a second sentence with the aim to correct the SR. Importantly, Bruce’s correction of SR is specific, in that “suit” is maintained and “fashion” is corrected in “function”. In terms of our model, this corrective actions of Bruce can be considered as part of a strategy for forming SRs that offer a solid ground for the predictions of the other agent. In other words, in addition to fulfilling his individual communicative and non-communicative intentions $I^1$, Bruce’s planning has the additional constrain that Lucius can correctly infer and predict his beliefs, intentions and actions based on SR (i.e., that Lucius can reliably infer $P(B_1^1, B_{t+1}^1|SR)$, $P(I_1^1, I_{t+1}^1|SR)$, and $P(A_1^1, A_{t+1}^1|SR)$, although in most cases Lucius does not
need to to this in practice).

Overall, our analysis indicates that both agents help the cognitive processes of the other agents when they plan their actions, and use SR for coordinating and guiding them. This idea is compatible with joint action control, which prescribes that agents optimize the joint intention of achieving communication success rather than only their own (communicative and non-communicative) intentions. The hallmark of this strategy is performing actions that do not only achieve individual goals but facilitate inferential processes of the other agent, and this is mainly done by maintaining reliable SRs so as to be predictable, and occasionally violating them for conveying communicative intentions; see [50], [97].

B. Competitive cases

A potential objection to our discussion of the first dialogue is that, because the two characters collaborate and both take profit from exchanging information, interactive goals and the formation of SRs are just side-effects of the fulfillment of Bruce and Lucius' individualistic goals. A more compelling case for the fulfillment of interactive goals can be made by analyzing the following dialogue, which takes place between two competitors: the brave Batman and the foolish Joker. Differently from the former case, the two are enemies, so it is not expected that Batman and Joker collaborate. Despite so, as we will show, similar joint action dynamics, including SR formation, are also in play in this competitive scenario. The situation is the following: after kidnapping Dent, The Joker is imprisoned; now Batman is rudely interrogating him.

**Batman:** Where is Dent?
**The Joker:** You have all these rules and you think they'll save you.

[Batman slams the Joker against a wall]

**Batman:** I have one rule.
**The Joker:** Then that's the rule you'll have to break to know the truth.
**Batman:** Which is?
**The Joker:** The only sensible way to live in this world is without rules. And tonight you're gonna break your one rule.
**Batman:** I'm considering it.
**The Joker:** No, there's only minutes left, so you're gonna have to play my little game if you want to save one of them.

1) Joint action control in play: Although Batman and The Joker are enemies, they collaborate for building SRs and pursue the common goal to have a successful dialogue (although this common goal is ultimately functional to different individualistic goals). The counterintuitive advantage of collaboration is that linguistic exchanges can be understood only in the light of the shared communicative context, in the absence of which the individualistic goals cannot be achieved.

In the first sentence, Batman makes a request. In terms of our model, this is the result of a planning process aimed to fulfill Batman's communicative and non-communicative intentions, but at the same time it sets a communicative context that entails precise expectations about The Joker's reply (which should be an answer). The reply of The Joker, however, is not an answer to Batman's question, and is unpredictable given the SR. It is important to appreciate the fact that this surprising event is not casual, but the Joker violates Batman's prediction with the explicit aim to convey his communicative intentions (he does not want to reveal where Dent is, and wants to convince Batman that pursuing good is hopeless). According to our model, the surprising event forces Batman to mindread The Joker so as to extract his communicative intentions. Although the literal meaning of The Joker's sentence is not sufficient to fully understand them, by slamming The Joker, Batman signals that he correctly understood the intentions of The Joker, and he does not appreciate them.

The use of surprise to convey communicative intentions is even more evident in the last sentence, when The Joker says: "save one of them" rather than "save him" as it would be natural for referring to Dent. The Joker uses this unpredictable sentence intentionally, so as to signal that he kidnapped another person in addition to Dent (in fact, he kidnapped Dent's girlfriend Rachel), and that the SRs have to be revised. In terms of our model, The Joker selects an action $A_1$ so that $P(A_1|SR)$ is low, and the surprising event guides Batman's mindreading (or, in other words, so that Batman puts his attention on what exactly using $A_1$ rather than a more plausible alternative entails). The intention behind this sentence is forcing Batman to make a dolorous decision by choosing who of them to save (indeed, Batman is tricked and saves Dent, which then becomes a dangerous enemy of Batman for revenge).

Overall, in this dialogue The Joker's surprising answers are not casual but always signal precise communicative intentions, which go well beyond the literal meaning of his utterances. In keeping with the idea that joint action control is in play even in competitive scenarios, The Joker can "play his game" with Batman only because he makes a subtle, still "lawful" use of SRs: he uses them as long as he wants to guide Batman's expectations, and then violates them to convey communicative intentions.

2) Use of linguistic and non-linguistic elements: A second important element that emerges from this dialogue is that linguistic and non-linguistic goals, actions and context are mixed. During the interaction Batman and The Joker use their body actions, the direction of their gaze, the tone of their voice and their facial expressions as extra instruments to encode these intentions and complement the speech acts, for instance for signaling what is relevant. Indeed, facts that happen in the sensorimotor, non-linguistic context (e.g., Batman grappling and slamming The Joker, the patterns of eye movements) are as relevant as utterances for understanding the pragmatics of this dialogue.

In some cases praxic actions and speech acts can be used interchangeably. For instance, lines 2-4 of the dialogue is a "trial of strength" between Batman and The Joker; they are off-topic and the utterances have little literal meaning, and this suggests that each one wants to scare the adversary and impose his view of life (and indeed the two are fighting at the same time). A similar, but non linguistic trial of strength happens in another part of the movie, when Batman threateningly rides a motorcycle towards The Joker, who refuses to move. Again, pragmatic ability spans across linguistic and non-linguistic events. Not only the boundaries between these two contexts
are often blurry, but they can be selected depending on the situation: in the former but not the latter situation, the two can speak to each other, so using language is more appropriate.

IX. CONCLUSIONS

In this article, we have offered a cognitive model of the general pragmatic competence that underlies both verbal and non-verbal interaction. Elaborating on Levinson’s [80] idea of an “interaction engine” that underlies the ability for meaningful interactions, linguistic and non-linguistic, we have described the cognitive ingredients of the interaction engine, and offered a computationally-guided analysis of the problems they face in increasingly complex scenarios, individualistic, social and then linguistic. This analysis has led to the identification of a common set of mechanisms that could have originated for the sake of non-linguistic interactions, and that were then reused for processing the pragmatics of linguistic interactions.

Although the paper focuses on the cognitive mechanisms behind intentional action planning and recognition, an equally important aspect of the pragmatic competence is the rich semiosis of body and actions. Even in this case, there is a lot of space for reuse from non-linguistic to linguistic domains. Indeed, communicative intentions can be conveyed by virtually any kind of action, body gestures or facial expressions, and the highly expressive behavioral repertoire of primates is an ideal substrate for conveying and recognizing communicative intentions. During evolution, specialized actions and gestures have developed to carry out communicative implicatures, such as facial expressions, mechanisms for capturing or directing attention, pointing at objects, and giving emphasis to actions (e.g., repeating or exaggerating them). The pre-existing mechanisms of goal-directed planning and understanding, including mindreading mechanisms, supported the achievement of communicative intentions in non-specialized actions and gestures. All these elements provide a flexible pragmatic competence, which, we argue, is used in linguistic communication, too, since non-linguistic and linguistic communication have largely the same structure and constraints.

A. Relations with other linguistic theories that emphasize a link between linguistic and non-linguistic pragmatics

As discussed in sec. I, many linguists have emphasized a relation between linguistic and non-linguistic pragmatics. Some of them have also described a mechanistic account of pragmatic competence. Here we discuss the similarities and differences between these theories and the ideas that we have presented here.

Grice [64], [65] has clearly recognized that pragmatics involve intention recognition as a key element. In this framework, the hearer’s pragmatic competence is described in terms of expectations that rely on conversational maxims (quantity, quality, etc.) under the assumption that conversation is a rational and cooperative activity. Although Grice acknowledges that talking is a special case of purposive behavior, the majority of his studies (and of those of other scholars that pursue a similar research agenda, such as Austin [8] and Searle [118]) have focused on linguistic (e.g., grammatical) rather than “generally cognitive” aspects of this process. We have taken a different perspective by stressing that non-linguistic and linguistic interactions require the same pragmatic competence and recruit the same (inferential, mindreading) mechanisms, and associated cognitive strategies. This makes our proposal more in line with cognitive theories (e.g., Relevance Theory, see later) than with the philosophical-linguistic approach of Grice. In addition, we have suggested joint action control as a (rational) strategy that guides interactions, and expectations within it. Although this idea is prima facie similar to the “principles of cooperative dialogue” of Grice, in that they are useful sources of expectations that help inferential processes, there are two important differences. First, our strategy gives a significant role to the formation of shared representations. Second, we propose joint action control as a cognitive mechanism rather than at the level of a maxim of conversation (this is a distinguishing factor between cognitive and linguistic/philosophical theories). To what extent conversational (or more broadly communicative) maxims could be re-conducted to joint action control and its dynamics remains an important direction for future research.

Relevance Theory [120] is a cognitive framework; it stresses that pragmatic competence consists in a cognitive mechanism that goes beyond language. In this framework, intention understanding is a resource-bounded cognitive process guided by expectations of relevance rather than expectations on maxims as described earlier. Similar to Relevance Theory, we have proposed a cognitive framework for understanding face-to-face communication in continuity with situated interaction. However, we have emphasized the continuity of strategies and cognitive mechanisms from individual control of action to joint action, and stressed the importance of joint action strategies for understanding the dynamics of communication (and what is relevant in it). In particular, we have argued that people actively guide the expectations of others by selecting actions so as to remain predictable, by purposively forming a common ground, and occasionally signaling communicative intentions by violating it. Our proposal is compatible with Relevance Theory in that we emphasize that the common ground provides important information to evaluate what is relevant (and indeed we have proposed that people engaged in joint action do, by default, interpret observed actions and communications as being relevant for the current interaction). At the same time, we propose that the same interaction engine of non-linguistic communication is also in play in linguistic domains, contrary to the idea of a dedicated linguistic comprehension module [121].

A third theoretical proposal that is related to ours is the idea of “interactive alignment” in dialogue [59], [103], which stresses automatic mechanisms of alignment on top of the common coding of perception and action [105] (also producing imitative behavior in non-linguistic interactive scenarios [35]). The thesis that we put forward can be seen as complementary to this theory, as it emphasizes deliberate strategies for SR formation and signaling, which act in concert with automatic mechanisms [97].
B. Toward embodied models of face-to-face communication

Traditional human-robot interaction and dialogue systems are limited in how they represent and process pragmatics. These systems (with some exceptions, see later) are typically endowed with linguistically-encoded pragmatic information, for instance in the form of scripts, which provide non-linguistic knowledge and context to the linguistic interactions. These scripts are not grounded in the sensorimotor context, do not include grounded models of objects to which they refer to (e.g., this chair), do not connect with an agent’s non-linguistic actions and goals, and do not use similar computational processes in linguistic and non-linguistic contexts (e.g., pointing at objects, shared attention). In brief, the pragmatic competence of these systems is a fully linguistic competence, and is based on the ability to manipulate linguistic information rather than on the ability to act and interact in the external environment, as argued for in grounded theories of cognition.

However, recently the common structure of action and dialogue has been recognized in the design of dialogue models [1], [41], [75], [86], [95]. For instance, while most speech processing models focused only on few inferences, and particularly in how to infer $S$ from $O$ (e.g., recognizing words and sentences from noisy speech signal), more recent plan-based dialogue agents are designed to perform and interpret speech acts, by using logics as well as probabilistic representations for estimating $P(A|S)$. Still, most of these models lack embodiment and do not relate these inferences to situated interaction. In addition, the joint nature of dialogue and its implications for computational processing has been less investigated. In particular, the (automatic and deliberate) mechanisms for forming shared representations, and the computational methods for using SRs so as to simplify communication have not yet been incorporated in existing dialogue systems.

A novel methodology to design face-to-face communication systems, human-robot and robot-robot interactions is required, in which communicative and linguistic abilities develop on top of, and in coordination with, the “interaction engine” that afford the same pragmatic abilities in non-linguistic domains. In this way, the pragmatics of linguistic communication are ultimately supported by the computational machinery of situated, sensorimotor interaction, along the lines that we have described in this article. In this novel methodology the ability to engage in face-to-face communication develops in coordination with pragmatic competence required for non-linguistic communication and (especially) joint action; this is alternative to the more common realization of linguistic processing as separated from action-perception systems.

We propose that this research objective should be pursued within the developmental robotics framework [133]. The increasingly complex scenarios analyzed in this paper provide a tentative roadmap for the progressive development of an interaction engine, and its usage in linguistic and non-linguistic domains. A first possibility is taking the staged methodology literally, and expose the robot to increasingly complex learning scenarios, such that pragmatic abilities required in linguistic exchanges develop on top of an existing interaction engine for situated interaction. A second, potentially more complex but also more plausible alternative, is co-developing linguistic and non-linguistic abilities, as observed in children, which are exposed to linguistic contexts from the beginning of their life (see sec. II-B).

A good starting point to develop (embodied) models of communication endowed with pragmatics are recent cognitive robotics models of imitation [48], [108], joint action [36], [45], [101], planning and action recognition [10], [11], [115], [129], [131], human-robot interaction [92], [113], experimental semiotics [56], [123], as well as models that embed language learning within a shared sensorimotor context between teacher and learner [53] (see [127] on this issue), and dialogue within a shared context [86]. These models, among others, include some aspects of the interaction engine and pragmatic competence that we have discussed.

An open issue for future research is scaling up the abilities already possessed by the aforementioned models to the linguistic domain, along the lines that we have suggested in this article, that is, by modeling a continuum of non-linguistic and linguistic pragmatic competences. To this aim, below we discuss three main challenges that this research initiative has to face.

1) Challenge 1. Understanding the neural bases of communication and pragmatics: Our proposal is linked to other cognitive theories that emphasize a link between language and action systems. Similar to Arbib [4], we have stressed the evolutionary aspects of the passage from motor to linguistic skills, except that we focused on pragmatic, semiotic and communicative functions of action and language rather than on grammatical and syntactical aspects. By showing that the computational problems behind individual action planning, non-linguistic and linguistic communication are related, we have offered a way to link pragmatic and communicative actions that complements Arbib’s view, and links with other evolutionary arguments that stress how linguistic communication has shared beliefs, shared attention and cooperation as precursors [13], [128]. There are at the moment few data that can confirm or disconfirm our theoretical and computational analysis. The reason is that, although mindreading has been studied extensively in social neuroscience, most studies focus on standard (praxic) actions, such as for instance grasping for eating, rather than on communicative functions behind actions, which are more studied by developmental psychology (but see [91]). A convergence of social neuroscience and developmental studies is therefore necessary to increase our knowledge of the cognitive machinery behind the processing of pragmatics.

2) Challenge 2. In search of common principles for reconciling language and action in cognitive theories: Our proposal is linked to other studies in cognitive science that deny a special, modular role for linguistic abilities. Although we have focused on pragmatics, similar proposals of “embodied” linguistic theories are becoming popular in cognitive science, such as those emphasizing modality-specific (rather than amodal) processing of language [16], [106], a continuity of action and language systems and their grammatical structure [4], the grounding of symbols in sensorimotor [29], [112] and social [122] domains, and the sensorimotor origins of abstract
As a result of these lines of research, a common view is emerging in the cognitive neuroscience literature that all linguistic processing is an embodied process, which is implemented by the same mechanisms of intentional action control that we have described. Our proposal fits nicely in this broad view, and captures an important yet unexplored aspect of linguistic processing: pragmatics. Note that, up to this moment the arguments for the embodiment of symbols, syntax, semantics and (now) pragmatics have proceeded quite independently; at this stage, we also need an integrated view of how all these levels coordinate. For instance, during dialogue the interactive process that we have proposed has to act in coordination with a more properly “linguistic” processing, which captures syntactic and semantic elements of language.

A promising direction of research is finding a common cognitive process that underlies all levels. It has recently been proposed to interpret language understanding as mental simulation, or a simulation process that involves linguistic entities but is essentially the same as in the sensorimotor domain in computational and neural terms, and spans from grammatical aspects to the simulation of meaning [14], [63], [104], [140]. For instance, Glenberg and Gallese [62] offer a theory of linguistic competence that builds on the components of sensorimotor control of action that we have discussed, and particularly on internal modeling. Note that the notion of internal models apply to other levels as well. For example, they can be used for the control of the phono-articulatory system (like any other body movement); thus, they offer an explanation of the functioning of (the phonetic aspects of) locutionary acts, which are one of the components of speech acts.

Although these lines of research bring the promise to unify the domains of action and language at the level of cognitive mechanisms, much remains to be done to demonstrate that the same (or at least very similar) processes of mental simulation could span all the levels of linguistic processing, syntactic, semantic, and pragmatic.

3) **Challenge 3. Providing evidence that the non-linguistic interaction engine is sufficient for processing the pragmatics of language:** A challenge for the idea of mental simulation as a unifying principle is the fact that, despite our emphasis on commonalities across linguistic and non-linguistic scenarios, still linguistic scenarios are more complex than non-linguistic ones, for at least two reasons.

First, the processing of communicative intentions obeys to a complex set of syntactic and semantic rules, which are characteristic of language and less prominent in other forms of joint action. Although these rules require a long period of learning to be mastered properly, in turn they give greater flexibility on what can be expressed linguistically.

Second, note that in our examples we have focused on face-to-face communication, where the sensorimotor context is (largely) shared. However, there are many examples of linguistic domains where this is not the case, because the two agents are not in the same spatial environment (e.g., in a phone call), or because they refer to objects or events that are not perceptually available (e.g., past or future events). For instance, in one of the dialogues that we have analyzed, Batman and The Joker refer to the fact that Dent is imprisoned, which would be difficult or even impossible without the use of language.

When the link to what is referred to cannot be offered by perception, language itself could play a role by affording the realization of shared communicative contexts, in which words, rather than percepts, serve as referents of objects and events. In other words, as recognized in studies of common ground [39], [38], language extends the context of interaction from the sensorimotor to more distal and abstract domains. For instance, certain illocutionary acts such as promising something, which refer to future situations, could require language to be expresses and/or developed. Levinson [79] convincingly argues that certain aspects of verbal communication, such as the possibility to address many persons at once, favor the formation of larger groups and of more complex interactions than non-verbal communication. (These are just examples of the general fact that language enhances cognition in multiple ways; see [37].)

Using a linguistically-created context rather than (or in addition to) a shared sensorimotor context does not hinder performance; indeed, people watching (at the TV) the dialogue between Batman and The Joker, or reading a book, understand the pragmatics correctly. At the same time, one could argue that as linguistic exchanges requires significantly more knowledge than other joint actions, language processing can be more costly in terms of memory and attention resources. In addition, the price of expressive power and of wider SRs is that intention recognition and the assignment of reference are computationally more difficult in linguistic domains. It is still unknown if this additional sources of complexity requires the adoption of dedicated (i.e., linguistic) sets of mechanisms for comprehension, as proposed for instance by Sperber and Wilson [121], or if language can support embodied simulations that replace direct experience, as suggested by embodied theories of language processing [18], [82], [140]. Some support from a reuse of sensorimotor strategies comes from the consideration that, even when people do not actually share a context, they can behave as if they did; for instance, during phone calls they can point at objects that the other cannot see. However, assessing the validity of embodied simulation theories in the context of linguistic exchanges, or the existence of dedicated mechanisms for processing the pragmatics of language, remain open issues for future research.

To this aim, here we advance a hypothesis that stems from our proposal of a joint action control strategy for solving interaction problems. We make the case that interaction strategies (and dynamics), rather than additional cognitive mechanisms, keep the computational problems of language processing within tractable boundaries. For instance, during dialogue cognitive agents formulate queries so that their intentions are easy to infer, give answers that are maximally informative (and possibly make one’s own intentions and beliefs available to the other agent), take turns to avoid processing noisy speech, check frequently if the other agent agrees or not, etc. In turn, these strategies could facilitate language processing, lowering the cognitive load and memory requirements. Part of the
working of these strategies is captured nicely by the Gricean maxims [64], and part is better modeled within our proposal of a joint action control; still, a complete description of how humans deal with the complexities of language processing is missing.

As a final remark, it is worth noting that part of the interaction problem could be solved by the adaptation of the linguistic code rather than (or, better, in addition to) the adaptation of the interacting minds. Languages co-evolved with cognitive abilities for the sake of efficient communication. Languages are flexible instruments, which include sophisticated tools such as linguistic cues (e.g., gender, type) that signal reference with low uncertainty, syntactic constructs that facilitate the parsing of intentions, and metaphorical constructs that map abstract situations into sensorimotor ones [77]. These linguistic tools can be used in combination with existing means such as gesture, emphasis and facial expressions to convey meaning with low ambiguity, and to facilitate pragmatic inference.

By recognizing the importance of understanding the linguistic code together with the interaction engine, we emphasize that the developmental robotics methodology that we have proposed cannot be divorced from linguistic studies, and a tight collaboration between linguists and cognitive scientists is required to advance our understanding of the pragmatics of language.

ACKNOWLEDGMENTS

Research funded by the EU’s FP7 under grant agreements n. FP7-231453 (HUMANOBGS) and n. FP7-270108 (GoalLeaders). I thank Dr. Fabian Chersi, Dr. Haris Dindo, and two anonymous reviewers for their precious suggestions.

REFERENCES
