SCIENTIFIC REPORTS

Towards a mathematical theory of meaningful communication

Bernat Corominas-Murtra[1,2,3], Jordi Fortuny[4] & Ricard V. Solé[2,3,5]

[1]Section for Science of Complex Systems, Medical University of Vienna, Spitalgasse, 23 1090 Vienna, Austria, [2]ICREA-Complex Systems Lab, Universitat Pompeu Fabra, Dr Aiguadé, 88, 08003 Barcelona, Spain, [3]Institut de Biologia Evolutiva. CSIC-UPF. CMIMA-Pssg. de la Barceloneta 37-49. 08003 Barcelona, Spain, [4]Departament de Filologia Catalana. Universitat de Barcelona. Gran Via de les Corts Catalanes, 585, 08007 Barcelona, Spain, [5]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501 USA.

Meaning has been left outside most theoretical approaches to information in biology. Functional responses based on an appropriate interpretation of signals have been replaced by a probabilistic description of correlations between emitted and received symbols. This assumption leads to potential paradoxes, such as the presence of a maximum information associated to a channel that creates completely wrong interpretations of the signals. Game-theoretic models of language evolution and other studies considering embodied communicating agents show that the correct (meaningful) match resulting from agent-agent exchanges is always achieved and natural systems obviously solve the problem correctly. Inspired by the concept of *duality of the communicative sign* stated by the swiss linguist Ferdinand de Saussure, here we present a complete description of the minimal system necessary to measure the amount of information that is consistently decoded. Several consequences of our developments are investigated, such as the uselessness of a certain amount of information properly transmitted for communication among autonomous agents.

Major innovations in evolution have been associated with novelties in the ways information is coded, modified and stored by biological structures on multiple scales[1]. Some of the major transitions involved the emergence of complex forms of communication, being human language the most prominent and difficult to explain[2]. The importance of information in biology has been implicitly recognized since the early developments of molecular biology, which took place simultaneously with the rise of computer science and information theory. Not surprisingly, many key concepts such as coding, decoding, transcription or translation were soon incorporated as part of the lexicon of molecular biology[3].

Communication among individual cells promoted multicellularity, which required the invention and diversification of molecular signals and their potential interpretations. Beyond genetics, novel forms of non-genetic information propagation emerged. At a later stage, the rise of neural systems opened a novel scenario to interact and communicate with full richness[2]. Human language stands as the most complex communication system and, since communication deals with the creation, reception and processing of information, understanding communication in information theoretic terms has become a major thread in our approach to the evolution of language.

In its classical form, information theory (IT) was formulated as a way of defining how signals are sent and received through a given channel with no attention to their meaning. However, in all kinds of living systems, from cells sharing information about their external medium, individuals of a given species surviving in a world full of predators or when two humans or apes exchange signals, a crucial component beyond information is its meaningful content[4]. The distinction is very important, since information has been treated by theoreticians since Shannon's seminal work[5] as a class of statistical object that measures correlations among sets of symbols, whereas meaning is inevitably tied to some sort of functional response with consequences for the fitness of the communicating agents. This standard scheme describing information transmission through a noisy channel[5] is summarized in figure (1)a. The most familiar scenario would be described by a speaker (S) and a listener or receiver (R) having a conversation in a living room. The air carries the voice of the first and is the channel, which would be reliable (low or zero noise) if nothing except R and S were present. Instead, the channel will become more and more unreliable (noisy) as different sources of perturbation interfere. These can be very diverse, from air turbulence and children laughing to another conversation among different people. Consistently with any standard engineering design, Shannon's picture allows us to define efficient communication in terms somewhat similar to those used -for example- within electric transmission networks. In this case, a goal of the system design is
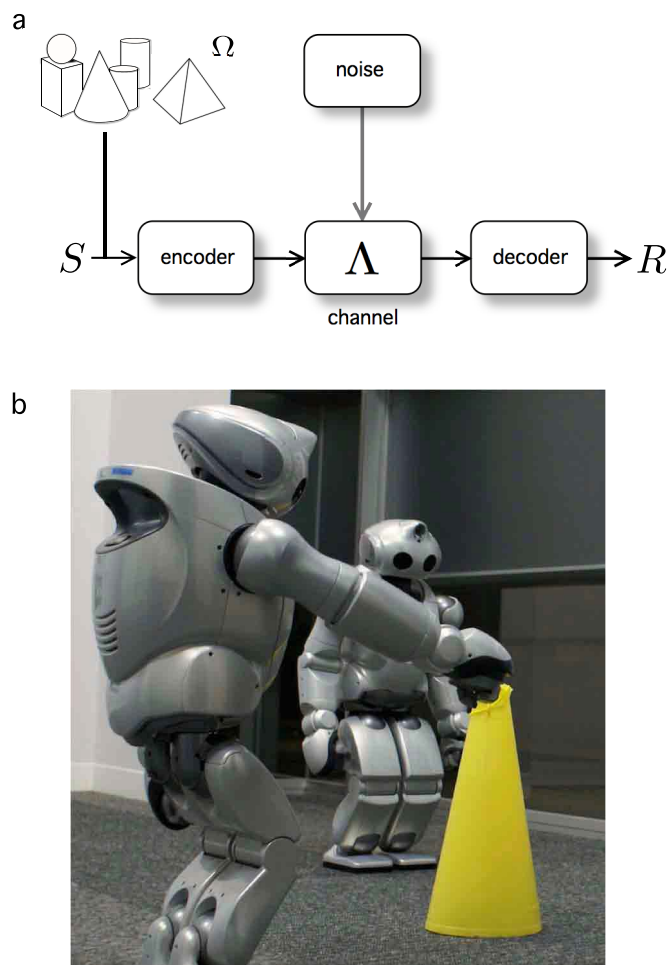
Figure 1 | In standard theory of information, as defined in Shannon's theory, a communication system (a) is described in terms of a sequential chain of steps connecting a source of messages (S) and a final receiver (R). The source can be considered linked to some external repertoire of objects (Ω). An encoder and a decoder participate in the process and are tied through a channel Λ, subject to noise. The acquisition and evolution of a language, as it happens in artificial systems of interacting agents, like robots (b), involves some additional aspects that are usually ignored in the original formulation of Shannon's approach. Those include the embodiment of agents and the necessary consistency in their communicative exchanges emerging from the their perceptions of the shared, external world. Picture courtesy of Luc Steels.

minimizing the heat loss during the transmission process. Information is a (physically) less obvious quantity, but the approach taken by standard IT is quite the same.

As a consequence of its statistical formulation, IT does not take into account "meaning" or "purpose" which, as noted by Peter Schuster[1], are also difficult notions for evolutionary biology. Despite this limitation, it has been shown to successfully work in the analysis of correlations in biology[6]. However, one undesirable consequence of this approach is that some paradoxical situations can emerge that contradict our practical intuition. An example is that a given pair of signals s_1, s_2 associated to two given objects or events from the external world could be "interpreted" by the receiver of the messages in a completely wrong way –"fire" and "water", for example, could be understood, as "water" and "fire", respectively. Measured from standard IT -see below- the information exchanged is optimal -even perfect- if "fire" ("water") is always interpreted as "water" ("fire"). In other words, full miscommunication can also score high, as perfectly "efficient", within Shannon's framework.

Therefore, one should approach the communicative sign as a *dual* entity that must be preserved as a whole in the communicative exchange. This crucial *duality sign* in communicative exchanges was already pointed out -with some conceptual differences to the version we will develop below-before the birth of information theory by the Swiss linguist Ferdinand de Saussure in his acclaimed *Cours de linguistique générale*[7].

It seems obvious that meaning -and its connection to some signal, in order to create the dual entity- plays an essential role and has been shaped through evolution: "the message, the machinery processing the message and the context in which the message is evaluated are generated simultaneously in a process of coevolution"[1]. In our bodies, proper recognition of invaders is essential to survival, and failures to recognizing the self and the non-self are at the core of many immune diseases[8,9]. Similarly, learning processes associated to proper identification of predators and how to differentiate them from inmates are tied to meaningful information. Beyond the specific details associated to each system, correct information storing and sharing, and the relevance of meaning is well illustrated by its impact on evolutionary dynamics. As pointed out in[3] we can say that, in biology, the coder is natural selection. In this way, the use of evolutionary game theoretic arguments has played a very important role in shaping evolutionary approaches to language and communication[10–15], but require some extension in order to properly account for meaningful information. Moreover, evolutionary robotics and the artificial evolution of protolanguages and proto-grammars is a unique scenario where such a framework naturally fits[16–22]. Evolving robots capable of developing simple communication skills are able of acquiring a repertoire of appropriate signals, share them and interpret correctly the signals sent by other agents. The coherent development of a shared set of symbols that is correctly used -and thus where "meaning" is preserved- becomes central. Such coherence results from the combination of a shared repertoire of signals together with a shared perception of the external world, as detected and perceived by the same class of sensing devices.

In this paper we develop and describe an information-theoretic minimal system in which the signal is linked to a referential value. This relation is assumed to be simple and direct, so that no other process than the mapping is assumed. Other forms of more complex meaning associations would deviate from the spirit of the paper, which is to introduce the minimum framework accounting for the conservation the simplest form of meaning. In a nutshell, we are going to derive an information-theoretic measure able to grasp the consistency of the shared information between agents, when *meaning* is introduced as a primitive referential value attributed to one or more signals.

Results

We start this section describing the minimal system incorporating referential values for the sent signals. Within this system, we show what is meant when we say that information theory is blind to any meaning of the message. We then derive the amount of consistently decoded information between two given agents exchanging information of their shared world, thereby fixing the problem pointed out above, and analyze some of its most salient properties, including the complete description of the binary symmetric channel within this new framework.

The minimal system encompassing referentiality. Our minimal system to study the referential or semantic consistency of a given information exchange will involve two *autonomous communicative agents*, A, B, a *channel*, Λ, and a *shared world*, Ω. Agents exchange information about their shared world through the channel -see figure (2). Now we proceed to describe it in detail.
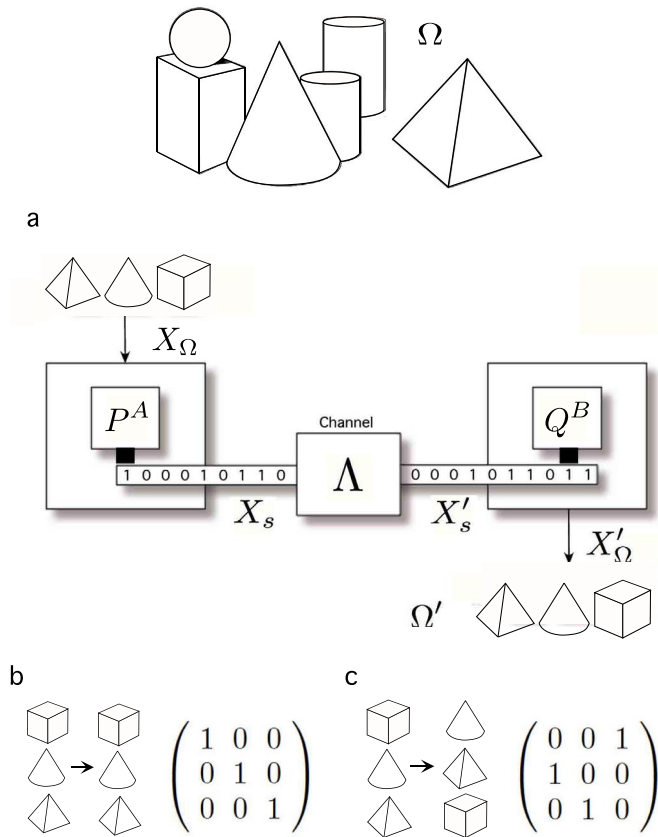
Figure 2 | Minimal communicative system to study the conservation of referentiality (a): A shared world, whose events are the members of the set Ω and whose behavior is governed by the random variable X_Ω. A coding engine, $\mathbf{P^A}$, which performs a mapping between Ω and the set of signals $\mathcal{S}$, being X_s the random variable describing the behavior of the set of signals obtained after coding. The channel, Λ, may be noisy and, thus, the input of the decoding device, $\mathbf{Q^B}$, depicted by X_s', might be different from X_s. $\mathbf{Q^B}$ performs a mapping between $\mathcal{S}$ and Ω, whose output is described by X_Ω'. Whereas mutual information provides a measure of the relevance of the correlations between X_Ω and X_Ω', *consistent information* evaluates the relevance of the information provided by consistent pairs with regard to the overall amount of information. In this context, from a classical information-theoretical point of view, situations like b) and c) could be indistinguishable. By defining the so-called consistent information we can properly differentiate b) and c) by evaluating the degree of consistency of input/output pairs -see text.

Description. An *agent*, $\mathbf{A}$, is defined as a pair of computing devices,

$$\mathbf{A} \equiv \{\mathbf{P^A},\mathbf{Q^A}\}, \tag{1}$$

where $\mathbf{P^A}$ is the coder module and $\mathbf{Q^A}$ is the decoder module. The shared world is defined by a random variable X_Ω, which takes values from the set of events, Ω, $\Omega = \{m_1, \ldots, m_n\}$, denoting the (always non-zero) probability associated to any event $m_k \in \Omega$ as $p(m_k)$. The coder module, $\mathbf{P^A}$, is described by a mapping from Ω to the set of signals: $\mathcal{S} = \{s_1, \ldots, s_n\}$. We will here assume $|\Omega| = |\mathcal{S}| = n$, unless the contrary is indicated. The mapping that represents the coder module is defined by means of a matrix of conditional probabilities $\mathbf{P^A}$, whose elements $\mathrm{P}_{ij}^A = \mathbb{P}^A(s_j|m_i)$ satisfy the normalization conditions (namely, for all $m_i \in \Omega$, $\sum_{j \leq n} \mathrm{P}_{ij}^A = 1$). The outcome of the coding process is depicted by the random variable X_s, taking values from $\mathcal{S}$ according to a probability distribution

$$q(s_i) = \sum_{j \leq n} p(m_j)\mathrm{P}_{ij}^A. \tag{2}$$

The channel Λ is characterized by the $n \times n$ matrix of conditional probabilities Λ, with matrix elements $\Lambda_{ij} = \mathbb{R}_\Lambda(s_j|s_i)$. The random variable X_s' describes the output of the composite system world + coder + channel, thereby taking values on the set $\mathcal{S}$, and follows the probability distribution q', defined as

$$q'(s_i) = \sum_k p(m_k) \sum_{j \leq n} \mathbf{P}_{kj}^\mathbf{A}\Lambda_{ji}. \tag{3}$$

Finally, the decoder module is a computational device described by a mapping from $\mathcal{S}$ to Ω; i.e. it receives $\mathcal{S}$ as the input set, emitted by another agent through the channel, and yields as output elements of the set Ω. $\mathbf{Q^A}$ is completely defined by its transition probabilities, namely, $\mathbf{Q}_{ik}^\mathbf{A} = \mathbb{P}^\mathbf{A}(m_k|s_i)$, which satisfy the normalization conditions (i.e., for all $s_i \in \mathcal{S}$, $\sum_{k \leq n} \mathbf{Q}_{ik}^\mathbf{A} = 1$). We emphasize the assumption that, in a given agent $\mathbf{A}$, following [14, 15] (but not [10, 11]) there is a priori no correlation between $\mathbf{P^A}$ and $\mathbf{Q^A}$.

Now suppose that we want to study the information transfer between two agents sharing the world. Let us consider $\mathbf{A}$ the encoder agent and $\mathbf{B}$ the decoder one, although we emphasize that both agents can perform both tasks. Agent $\mathbf{B}$ tries to reconstruct X_Ω from the information received from $\mathbf{A}$. The description of Ω made by agent $\mathbf{B}$ is depicted by the random variable X_Ω', taking values on the set Ω and following the probability distribution p', which takes the form:

$$p'(m_i) \equiv \sum_{l \leq n} p(m_l)\mathbb{P}_\mathbf{AB}(m_i|m_l), \tag{4}$$

where

$$\mathbb{P}_\mathbf{AB}(m_i|m_l) = \sum_{j,r \leq n} \mathrm{P}_{lj}^A \Lambda_{jr} \mathrm{Q}_{ri}^\mathbf{B}. \tag{5}$$

From which we can naturally derive the joint probabilities, $\mathbb{P}_\mathbf{AB}(m_i,m_j)$ as follows:

$$\mathbb{P}_\mathbf{AB}(m_i,m_j) = \sum_{l,r} p(m_j)\mathbf{P}_{jl}^\mathbf{A}\Lambda_{lr}\mathbf{Q}_{ri}^\mathbf{B}. \tag{6}$$

We say that X_Ω' is the *reconstruction* of the shared world, X_Ω, made by agent $\mathbf{B}$ from the collection of messages sent by $\mathbf{A}$. Summarizing, we thus have a composite system where the behavior at every step is described by a random variable, from the description of the world, X_Ω to its reconstruction, X_Ω' -see figure (2a):

$$\Omega \overset{X_\Omega \sim p}{\rightarrow} \mathbf{A} \overset{X_s \sim q}{\rightarrow} \Lambda \overset{X_s' \sim q'}{\rightarrow} \mathbf{B} \overset{X_\Omega' \sim p'}{\rightarrow} \Omega. \tag{7}$$

At this point, it is convenient to introduce, for the sake of clarity, some new notation. We will define two matrices, namely $\mathbf{J(AB)}$ and $\Lambda(\mathbf{AB})$ in such a way that $J_{ij}(\mathbf{AB}) \equiv \mathbb{P}_\mathbf{AB}(m_i,m_j)$ and $\Lambda_{ij}(\mathbf{AB}) \equiv \mathbb{P}_\mathbf{AB}(m_j|m_i)$. Finally, we will define the probability distribution $\Lambda_i(\mathbf{AB}) \equiv \{\Lambda_{i1}(\mathbf{AB}), \ldots, \Lambda_{in}(\mathbf{AB})\}$. This new notation will enable us to manage formulas in a more compact way.

Information-theorethic aspects of this minimal system. First we shall explore the behaviour of mutual information in this system. Detailed definitions of information-theory functionals used in this subsection are provided in the Methods section. Under the above described framework, we have two relevant random variables: *the world* X_Ω and the *reconstruction* of the world X_Ω'. Its mutual information $I(X_\Omega : X_\Omega')$ is defined as[5,23,24]:

$$I(X_\Omega : X_\Omega') = H(X_\Omega) - H(X_\Omega|X_\Omega'). \tag{8}$$

The above expression has an equivalent formulation, namely

$$I\left(X_\Omega : X'_\Omega\right) = \sum_{i,j \leq n} J_{ij}(\mathbf{AB}) \log \frac{J_{ij}(\mathbf{AB})}{p(m_i)q(m_j)}, \quad (9)$$

where the right side of the above equation can be identified as the *Kullback-Leibler* divergence between distributions $J(\mathbf{AB})$ and $p \cdot q$:

$$I\left(X_\Omega : X'_\Omega\right) = D(J(\mathbf{AB})\|p \cdot q). \quad (10)$$

Within this formulation, the mutual information is the amount of accessory bits needed to describe the composite system X_Ω, X'_Ω taking as the reference the distribution $p \cdot q$, which supposes no correlation between X_Ω and X'_Ω.

Let us underline a feature of mutual information which is relevant for our purposes. As is well-known, $\max I(X_\Omega, X'_\Omega) \leq H(X_\Omega)$, and equality holds if there is no ambiguity in the information processing process, meaning that the process is *reversible*, in logical terms. Thus, every event $m_i \in \Omega$ has to be decoded with probability 1 to some event $m_j \in \Omega$ which, in turn, must not be the result of the coding/decoding process of any other event. In mathematical terms, this means that $\mathbf{P^A}, \mathbf{Q^B}, \Lambda \in \Pi_{n \times n}$, being $\Pi_{n \times n}$ the set of $n \times n$ permutation matrices, which are the matrices in which every file and column contains $n - 1$ elements equal to 0 and one element equal to 1 -see Methods section. It is worth emphasizing that $\delta_{n \times n}$, the $n \times n$ identity matrix is itself a permutation matrix. Notice that if $\Lambda(\mathbf{AB}) \neq \delta$ some symbol m_i sent by the source is decoded as a different element m_j. This shift has no impact on the information measure $I\left(X_\Omega : X'_\Omega\right)$ if $\Lambda(\mathbf{AB}) \in \Pi_{n \times n}$, and this is one of the reasons by which it is claimed that *the content of the message* is not taken into account in the standard information measure. Actually, it is straightforward to show -see Appendix B- that only $n!$ out of the $(n!)^3$ configurations leading to the maximum mutual information also lead to a fully consistent reconstruction -i.e., a reconstruction where referential value is conserved. This mathematically shows that, for autonomous agents exchanging messages, mutual information is a weak indicator of communicative success.

Derivation of consistent information. Now we have a complete description of the minimal system able to encompass referential values for the sent signals. It is the objective of this section to derive an information-theoretic measure, different from mutual information, that will allow us to evaluate the amount of consistently decoded information.

Preliminaries. The rawest evaluation of the amount of consistently decoded pairs is found by averaging the probability of having a consistent coding/decoding process during an information exchange between agent $\mathbf{A}$ and agent $\mathbf{B}$. This corresponds to the view of an external observer simply counting events and taking into account only whether they are consistently decoded or not. This probability, denoted as $\theta_{\mathbf{AB}}$, is obtained by summing the probability of having consistent input output pair, i.e.:

$$\theta_{\mathbf{AB}} = \operatorname{tr} J(\mathbf{AB}) = \sum_{i \leq n} J_{ii}(\mathbf{AB}). \quad (11)$$

This formula has been widely used as a communicative payoff for an evolutionary dynamics in which consistent communication has a selective advantage[11,14,15]. We observe that the probability of error $p_e(\mathbf{AB})$ in this scenario is given by $p_e(\mathbf{AB}) = 1 - \theta_{\mathbf{AB}}$. Therefore, thanks to Fano's inequality -see Methods section-, we can relate this parameter to the information-theoretic functionals involved in the description of this problem, namely:

$$\boldsymbol{\theta_{\mathbf{AB}}} \leq 1 - \frac{H\left(X_\Omega | X'_\Omega\right)}{\log(n-1)}. \quad (12)$$

From this parameter, we can build another, a bit more elaborated functional. We are still under the viewpoint of the external observer

who is now interested in the fraction of information needed to describe the composite system X_Ω, X'_Ω that comes from consistent input/output pairs when information is sent from $\mathbf{A}$ to $\mathbf{B}$. This fraction, to be named $\sigma_{\mathbf{AB}}$, is:

$$\sigma_{\mathbf{AB}} = \frac{\operatorname{tr}(J(\mathbf{AB})\log J(\mathbf{AB}))}{H\left(X_\Omega, X'_\Omega\right)}. \quad (13)$$

We observe that the above quantity is symmetrical in relation to X_Ω and X'_Ω. These two estimators provide global indicators of consistency of the information exchange.

Consistent information. However, we can go further and ask us *how much of the information from the environment is consistently decoded by agent* $\mathbf{B}$ *when receiving data from* $\mathbf{A}$. As a first step, we observe that, since $J_{ij}(\mathbf{AB}) = p(m_i)\Lambda_{ij}(\mathbf{AB})$, we can rewrite equation (9) as:

$$I\left(X_\Omega : X'_\Omega\right) = \sum_{i \leq n} p(m_i) \sum_{j \leq n} \Lambda_{ij}(\mathbf{AB})\log\frac{\Lambda_{ij}(\mathbf{AB})}{p'(m_j)}$$
$$= \sum_{i \leq n} p(m_i)D(\Lambda_i(\mathbf{AB})\|p'). \quad (14)$$

Knowing that $D(\Lambda_i(\mathbf{AB})\|q)$ is the *information gain* associated to element m_i, $p(m_i)D(\Lambda_i(\mathbf{AB})\|q)$ is its weighted contribution to the overall information measure. If we are interested in the amount of this information that is consistently referentiated, we have to add an "extra" weight to $p(m_i)$, namely $\Lambda_{ii}(\mathbf{AB})$, which is the probability of having m_i both at the input of the coding process and at the output. Thus, since

$$\Lambda_{ii}(\mathbf{AB})p(m_i)D(\Lambda_i(\mathbf{AB})\|q) = J_{ii}(\mathbf{AB})D(\Lambda_i(\mathbf{AB})\|p'), \quad (15)$$

the amount of *consistent information* conveyed from agent $\mathbf{A}$ to agent $\mathbf{B}$, $\mathcal{I}(\mathbf{AB})$, will be:

$$\mathcal{I}(\mathbf{AB}) = \sum_{i \leq n} J_{ii}(\mathbf{AB})D(\Lambda_i(\mathbf{AB})\|p'). \quad (16)$$

Since this is the most important equation of the text, we rewrite it using standard probability notation:

$$\mathcal{I}(\mathbf{AB}) = \sum_{i \leq n} \mathbb{P}_{\mathbf{AB}}(m_i, m_i) \sum_{j \leq n} \mathbb{P}_{\mathbf{AB}}(m_j|m_i)\log\left(\frac{\mathbb{P}_{\mathbf{AB}}(m_j|m_i)}{p'(j)}\right). \quad (17)$$

We observe that the dissipation of consistent information is due to both standard noise $H\left(X_\Omega|X'_\Omega\right)$, and another term, which is subtracted to $I\left(X_\Omega : X'_\Omega\right)$, accounting for the loss of referentiality. Using equations (8), (9) and (16) we can isolate this new source of information dissipation, the *referential noise*, $v(\mathbf{AB})$, leading to:

$$v(\mathbf{AB}) = \sum_{i \leq n} D(\Lambda_i(\mathbf{AB})\|q)\left[\sum_{k \neq i} J_{ik}(\mathbf{AB})\right]. \quad (18)$$

Therefore, the total loss of referential information or *total noise* will be described as

$$\eta(\mathbf{AB}) \equiv H\left(X_\Omega|X'_\Omega\right) + v(\mathbf{AB}). \quad (19)$$

The above expression enables us to rewrite equation (16) as:

$$\mathcal{I}(\mathbf{AB}) = H(X_\Omega) - \eta(\mathbf{AB}), \quad (20)$$

which mimics the classical Shannon Information, now with a more restrictive noise term. Interestingly, the above expression is not symmetrical: the presented formalism distinguishes the world, X_Ω, from its reconstruction, X'_Ω. If we take into account that, according to the definition we provided for an autonomous communicating agent, the information can flow in both senses ($\mathbf{A} \rightarrow \mathbf{B}$ and $\mathbf{B} \rightarrow \mathbf{A}$), we can

compute the average success of the communicative exchange between **A** and **B**, $\mathcal{I}(\mathbf{A} : \mathbf{B})$, as:

$$\mathcal{I}(\mathbf{A} : \mathbf{B}) = H(X_\Omega) - \frac{1}{2}(\eta(\mathbf{AB}) + \eta(\mathbf{BA})). \qquad (21)$$

$\mathcal{I}(\mathbf{A} : \mathbf{B})$ is the *consistent information about the world* Ω *shared by agents* **A** *and* **B**. In contrast to the previous one, the above expression is now symmetrical, $\mathcal{I}(\mathbf{A} : \mathbf{B}) = \mathcal{I}(\mathbf{B} : \mathbf{A})$, because both agents share the same world, represented by X_Ω. We remark that this is an information-theoretic functional *between two communicating agents*, it is not an information-measure between two random variables, like mutual information is. This equation quantifies the communication success between two minimal communicating agents **A**, **B** transmitting messages about a shared world.

Properties. In this section we draw several important consequences from the treatment just presented, based on the consistent information concept. The rigorous and complete proofs behind them can be found in the Methods section, together with a brief discussion about the actual consistency of this measure when applied to single agents in a population (i.e., the 'self-consistency' or coherence that an individual agent should also keep about the world).

The binary symmetric channel. We first consider the simplest case, from which we can easily extract analytical conclusions that help us gain intuition: the *Binary Symmetric Channel* with uniform input probabilities. We are concerned with a world Ω having two events such that $p(1) = p(2) = 1/2$, two agents **A** and **B** sharing information about this world, and a binary channel, Λ. The agents' and channel configuration are assumed to be of the following form:

$$\Lambda(\mathbf{AB}) = \begin{pmatrix} 1 - \epsilon & \epsilon \\ \epsilon & 1 - \epsilon \end{pmatrix}, \qquad (22)$$

being $\Lambda(\mathbf{AB}) = \mathbf{P}^A \Lambda \mathbf{Q}^B$, as defined at the beginning of the results section. We will refer to ϵ as the *referential shift*, which is the probability that a given event is wrongly decoded in the reconstruction of Ω. In this minimal system all functionals can be easily evaluated. First, we have that $I(X_\Omega, X'_\Omega) = 1 - H(\epsilon)$, and that $\theta_{\mathbf{AB}} = 1 - \epsilon$, being $H(\epsilon)$ the entropy of a *Bernouilli process* having parameter ϵ -see Methods section. This leads to the following expression of the consistent information:

$$\mathcal{I}(\mathbf{AB}) = \theta_{\mathbf{AB}}(1 - H(\epsilon)) = \theta_{\mathbf{AB}} I(X_\Omega, X'_\Omega). \qquad (23)$$

We can also easily compute $\sigma_{\mathbf{AB}}$:

$$\sigma_{\mathbf{AB}} = \theta_{\mathbf{AB}} \frac{1 - \log\theta_{\mathbf{AB}}}{2 - H(\epsilon)}. \qquad (24)$$

The behavior of consistently decoded information is shown in figure (3). In these plots we confronted the behavior of $I(X_\Omega, X'_\Omega)$, $H(X_\Omega, X'_\Omega)$ and $H(X_\Omega|X'_\Omega)$ with their analogous counterparts when referentiality is taken into account, nalemy $\mathcal{I}(\mathbf{AB})$ and $\sigma_{\mathbf{AB}}$ and $v(\mathbf{AB})$ (and $\eta(\mathbf{AB})$) respectively. We can observe the symmetric behavior of the first ones against ϵ, which highlights the total insensibility to referentiality conservation of these classical measures. Instead, we observe that $\mathcal{I}(\mathbf{AB})$, $\sigma_{\mathbf{AB}}$, $\eta(\mathbf{AB})$ and $v(\mathbf{AB})$ do reflect the loss of referentiality conservation, showing a non-symmetric behavior with a generally decreasing trend as referentiality is progressively lost.

Decrease of information due to referential looses. One interesting consequence of equation (23) is that, except for very restricted situations, the presence of noise has a negative impact on the value of the consistent information, leading to the general conclusion that:
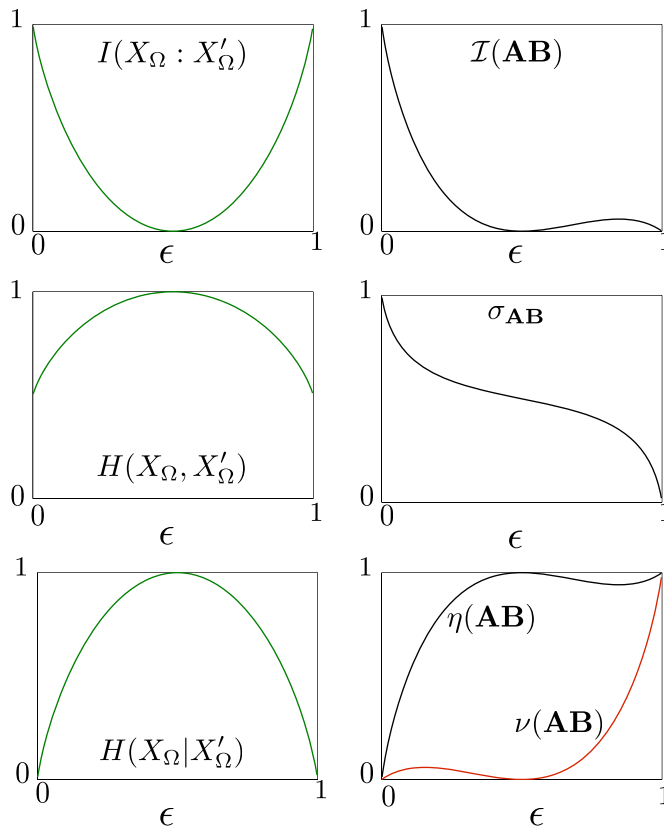

Figure 3 | The binary symmetric channel when we enrich the communication system with a referential set shared by coder and decoder agent. Plots correspond to the different values of the binary symmetric channel along ϵ, the referential shift parameter, from $\epsilon = 0$ (total information with no loss of referentiality) to $\epsilon = 1$ (total information with total loss of referentiality). On the left, from top to bottom, we have the classical, well known plots of $I(X_\Omega : X'_\Omega)$, $H(X_\Omega, X'_\Omega)$ (normalized to 1) and $H(X_\Omega|X'_\Omega)$. On the right, we have the equivalent ones accounting for the referentiality conservation, namely, on top, $\mathcal{I}(\mathbf{AB})$, next, $\sigma_{\mathbf{AB}}$ and in the last plot, we have $\eta(\mathbf{AB})$ (black line) and $v(\mathbf{AB})$ (red line). Units are given in bits. We observe that both $I(X_\Omega : X'_\Omega)$ (and $H(X_\Omega|X'_\Omega)$) have a symmetric behavior, with a minimum (maximum) at $\epsilon = \frac{1}{2}$ (total uncertainty). On the contrary, $\mathcal{I}(\mathbf{AB})$ does not show a symmetric behavior, showing two minima, at $\epsilon = \frac{1}{2}$ and at $\epsilon = 1$. There is a local maxima at about $\epsilon \approx 0.85$, which is a by-product of the combination of the loss of uncertainty of the system and a small but non-vanishable degree of referentiality conservation.

$$\mathcal{I}(\mathbf{AB}) < I(X_\Omega : X'_\Omega). \qquad (25)$$

This latter inequality shows that, in most cases, in the absence of a designer, part of the information properly transmitted is actually useless for communication in a framework of autonomous agents. As demonstrated in the Methods section, the strict inequality holds in general. Indeed, the above relation becomes equality only in the very special case where there is perfect a matching between the two agents (i.e.: $\Lambda(\mathbf{AB}) = \delta_{n \times n}$, being $\delta_{n \times n}$ the $n \times n$ identity matrix.) or trivially, in the case where $I(X_\Omega : X'_\Omega) = 0$.

But we can go further. Let us consider that we know that the system displays a given value of $I(X_\Omega : X'_\Omega)$ and, by assumption, we also know $H(X_\Omega)$. In these conditions, one can easily derive $H(X_\Omega|X'_\Omega)$ by simply computing $H(X_\Omega) - I(X_\Omega : X'_\Omega)$. But it is possible to set a bound to the value of $\mathcal{I}(\mathbf{AB})$ as well. As in many problems of information theory, the general case is hard, even impossible to deal with. However, several approaches become viable

in special but illustrative cases. Let us assume the paradigmatic configuration in which $(\forall m_i \in \Omega) p(m_i) = 1/n$ and where $\Lambda(\mathbf{AB})$ acts as a symmetric channel. In this case, we have that $\mathcal{I}(\mathbf{AB}) \leq \theta_{\mathbf{AB}} I(X_\Omega : X'_\Omega)$, where

$$\theta_{\mathbf{AB}} \leq \frac{I(X_\Omega : X'_\Omega)}{H(X_\Omega)}, \tag{26}$$

and, therefore:

$$\mathcal{I}(\mathbf{AB}) < \frac{I^2(X_\Omega : X'_\Omega)}{H(X_\Omega)} + 1. \tag{27}$$

(See the Methods section for the details of the above derivations). This tells us, after some algebra, that in this framework,

$$\eta(\mathbf{AB}) \geq 2H(X_\Omega|X'_\Omega) - \frac{H^2(X_\Omega|X'_\Omega)}{H(X_\Omega)} + 1. \tag{28}$$

Therefore, for $H(X_\Omega) \gg H(X_\Omega|X'_\Omega)$, we have that $\eta(\mathbf{AB}) \gtrsim 2H(X_\Omega|X'_\Omega)$, leading to

$$\mathcal{I}(\mathbf{AB}) \lesssim H(X_\Omega) - 2H(X_\Omega|X'_\Omega) \tag{29}$$

and, for example, for the case in which $H(X_\Omega) \approx 2H(X_\Omega|X'_\Omega)$ we have that:

$$\mathcal{I}(\mathbf{AB}) \lesssim H(X_\Omega) - \frac{3}{2}H(X_\Omega|X'_\Omega) \tag{30}$$
$$= \frac{1}{2} I(X_\Omega, X'_\Omega),$$

The above examples enable us to illustrate the strong impact of noise on the conservation of the referential value within a communication exchange -stronger than the one predicted by standard noise.

Discussion

Shannon's information theory had a great, almost immediate impact in all sorts of areas, from engineering and genetics to psychology or language studies[25]. It also influenced the work of physicists, particularly those exploring the foundations of thermodynamics, who found that the entropy defined by Shannon provided powerful connections with statistical mechanics, particularly in terms of correlations. It is mainly at that level -i. e. the existence of correlations among different subsystems of a given system- that the use of information theory has proved useful. But correlations do not ensure a crucial type of coherence that seems necessary when dealing with meaningful communication: the preservation of referentiality.

In this paper we have addressed a especially relevant problem, namely the development of an information-theoretic framework able to preserve meaning. This is a first step towards a more general goal, which would involve establishing the basis for an evolutionary theory of language change including referentiality as an explicit component. We have shown that, if consistent information is considered, its value is significantly lower than mutual information in noisy scenarios. We have derived an analytical form of consistent information, which includes referential noise along with the standard noise term. Our information measure defines a non-symmetrical function and properly weights the -more strict- requirement of consistency. We have illustrated our general results by means of the analysis of a classical, minimal scenario defined by the binary symmetric channel. The approach taken here should be considered as the formally appropriate framework to study the evolution of communication among embodied agents, where the presence of consistency is inevitable due to shared perception constraints. Moreover, it might also be useful as a consistent mathematical framework to deal with cognitive-based models of brain-language evolution[26–28]. At this point, we should point out an important issue: Consistency of the

communicative exchange is here evaluated between agents, not internally to a given agent *talking to itself*. Actually, there is no a priori any correlation between the coding and the decoding modules of a given agent. In doing so, we take the viewpoint proposed by [14] and [15]. Other approaches assumed an explicit link between the coding and decoding modules of the agent, thereby avoiding from the beginning the paradoxical situation in which two agents perfectly understand each other but, at the same time, they are not able to understand themselves [10, 11]. However, as shown in[29], this situation is unlikely to occur under selective pressures, for the frameworks depicted by these earlier works. In the Methods section is shown that the proposed framework has also the same property, i.e., that the maximisation of consistent communication in a given community of agents leads to the self-consistency of each of them, without the need of imposing it externally, thereby simplifying the mathematical apparatus.

The framework we have developed is somehow inspired by Saussure's duality of sign: a (linguistic) sign is a twofold entity compounded of a signifier and a signified. However, it must be mentioned that there is a substantial difference between the theory we have developed and a Saussurean approach. According to Saussure, the relation between a signifier and a signified is fixed with respect to the linguistic community that uses the sign. "The masses have no voice in the matter, and the signifier chosen by language could be replaced by no other". Saussure adopts therefore a 'static' approach to the study of signs, whereas we adopt a dynamic perspective that allows us to address the possibility that different agents assign different meanings to the same symbol, in which case referentiality is not preserved. In this way we extend evolutionary game-theoretic arguments in order to derive a measure of consistency of the shared information between agents by incorporating the (non-)preservation of referentiality.

In the presented work we took the simplest possible form of meaning, namely, its referential object. However, we said nothing about the object itself. Further works might explore the inclusion in the above proposed framework an explicit quantification of meaning beyond its referential value, to rank events of the world and to refine the role of the information functional to evaluate proper communication exchanges in selective scenarios. In addition, new hallmarks beyond the agent-channel-agent should be explored, leading to new forms of information which play a role in biological organisation and which are poorly reflected in such a schema.

Methods

Definitions. *Information theoretic functionals.* The following definitions are intended to be minimal. We refer the interested reader to any standard textbook on information theory, such as [23] or [24].

.-Given a random variable X_Ω taking values over the set Ω following a probability distribution p,

$$H(X_\Omega) = -\sum_{i \leq n} p(m_i) \log p(m_i) \tag{31}$$

is the standard *Shannon* or *statistical* entropy.

.-Given two random variables, X_Ω and X'_Ω,

$$H(X_\Omega|X'_\Omega) = -\sum_{1 \leq n} q(m_i) \sum_{j \leq n} \mathbb{P}(m_j|m_i) \log \mathbb{P}(m_j|m_i) \tag{32}$$

is the *conditional* entropy of X_Ω with respect X'_Ω, being, in that case, $\mathbb{P}(m_j|m_i) \equiv \mathbb{P}(X_\Omega = m_j|X'_\Omega = m_i)$. Additionally,

$$H(X_\Omega), X'_\Omega = -\sum_{1 \leq n} q(m_j) \sum_{j \leq n} \mathbb{P}(m_j, m_i) \log \mathbb{P}(m_j, m_i) \tag{33}$$

where $\mathbb{P}(m_j, m_i) \equiv \mathbb{P}(X_\Omega = m_i, X'_\Omega = m_j)$ is the *joint entropy* of the two random variables X_Ω, X'_Ω.

.-Given two probability distributions π_1, π_2 defined over the set Ω, the *Kullback-Leibler* divergence of *relative entropy* of π_1 with respect π_2 is:

$$D(\pi_1\|\pi_2) = \sum_{i \leq n} \pi_1(x_i) \log \frac{\pi_1(x_i)}{\pi_2(x_i)}, \tag{34}$$

which is the amount of extra information we need to describe π_1 taking as the reference distribution π_2.

.-*Fano's inequality.* The probability of error in decoding is bounded satisfies the following inequality:

$$p_e \geq \frac{H\left(X_\Omega | X'_\Omega\right) - 1}{\log(n-1)}. \tag{35}$$

.-A *Bernoulli process* is a stochastic process described by a random variable X taking value in the set $A = \{0, 1\}$, being $p(0) = 1 - \epsilon$ and $p(1) = \epsilon$. ϵ is the *parameter* of the Bernoulli process. Its entropy $H(X)$ is commonly referred as $H(\epsilon)$, since it only depends on this parameter:

$$H(\epsilon) = -(1-\epsilon)\log(1-\epsilon) - \epsilon \log \epsilon. \tag{36}$$

Permutation matrices. A *permutation matrix* is a square matrix which has exactly one entry equal to 1 in each row and each column and 0's elsewhere. For example, if $n = 3$, we have 6 permutation matrices, namely:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$
$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}. \tag{37}$$

The set of $n \times n$ permutation matrices is indicated as $\Pi_{n \times n}$ and it can be shown that, if $\mathbf{A} \in \Pi_{n \times n}$, $\mathbf{A}^{-1} = \mathbf{A}^T \in \Pi_{n \times n}$ and, if $\mathbf{A}, \mathbf{B} \in \Pi_{n \times n}$, the product $\mathbf{AB} \in \Pi_{n \times n}$. Furthermore, it is clear that $\delta_{n \times n} \in \Pi_{n \times n}$, being δ the identity matrix or *Kronecker symbol*, defined as $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$, otherwise.

Inequalities. We present the inequalities described in the main text in terms of three lemmas on the upper bounds of $\mathcal{I}(\mathbf{AB})$. The first one concerns inequality (25). The second one is general and supports the third, which proves inequality (27):

Lemma 1.- Let $\mathbf{AB}$ be two agents sharing the world Ω. The Amount of *consistent* information transmitted from $\mathbf{A}$ to $\mathbf{B}$ -when $\mathbf{A}$ acts as the coder agent and $\mathbf{B}$ as the decoder one- satisfies that

$$\mathcal{I}(\mathbf{AB}) = I\left(X_\Omega : X'_\Omega\right) \tag{38}$$

only in the following two extreme cases:

1. $I\left(X_\Omega : X'_\Omega\right) = 0$, or
2. $\Lambda(\mathbf{AB}) = \delta_{n \times n}$.

Otherwise, $\mathcal{I}(\mathbf{AB}) < I\left(X_\Omega : X'_\Omega\right)$.

Proof.- The first case is the trivial one in which there is no information available due to total uncertainty -corresponding to $\epsilon = \frac{1}{2}$ in the case of the symmetric binary channel studied above, see also figure (3). The second one is more interesting. Indeed, having $\Lambda(\mathbf{AB}) = \delta$ means that

$$\left(\mathrm{P}^\mathbf{A}, \Lambda, \mathrm{Q}^\mathbf{B} \in \Pi_{n \times n}\right) \text{ and } \mathrm{P}^\mathbf{A} = \left(\Lambda \mathrm{Q}^\mathbf{B}\right)^T, \tag{39}$$

where we use that, if $\mathbf{C} \in \Pi_{n \times n}$, $\mathbf{C}^{-1} = \mathbf{C}^T$, also having that $\mathbf{C}^T \in \Pi_{n \times n}$. Out of these two situations, $\exists J_{ik}(\mathbf{AB}) > 0$, in which $i \neq k$, since there are more than n non-zero entries in the matrix $\Lambda(\mathbf{AB})$, leading to

$$\mathcal{I}(\mathbf{AB}) < I\left(X_\Omega : X'_\Omega\right). \tag{40}$$

Lemma 2.- Let $\mathbf{AB}$ be two agents sharing the world Ω. The Amount of *consistent* information transmitted from $\mathbf{A}$ to $\mathbf{B}$ -when $\mathbf{A}$ acts as the coder agent and $\mathbf{B}$ as the decoder one- is bounded as follows:

$$\mathcal{I}(\mathbf{AB}) \leq \left(1 - \frac{H\left(X_\Omega | X'_\Omega\right) - 1}{\log(n-1)}\right) \left(\max_i \{D(\Lambda_i(\mathbf{AB}) \| p')\}\right). \tag{41}$$

Proof.- Let $\vec{v}$ and $\vec{u}$ be two vectors of $\mathbb{R}^n$. Its scalar product, $\langle \vec{v}, \vec{u} \rangle \sum_{i \leq n} v_i u_i$, is bounded, thanks to the so-called *Hölder's inequality*, in the following way:

$$|\langle \vec{v}, \vec{u} \rangle| \leq \left(\sum_{i \leq n} v_i^\alpha\right)^{\frac{1}{\alpha}} \left(\sum_{i \leq n} u_i^\beta\right)^{\frac{1}{\beta}}, \tag{42}$$

as long as α and β are *Hölder conjugates*, i.e., $1/\alpha + 1/\beta = 1$. The above expression can be rewritten, using the notation of norms as $|\langle \vec{v}, \vec{u} \rangle| \leq \|\vec{v}\|_\alpha \cdot \|\vec{u}\|_\beta$ -recall that, for $\alpha = \beta$

$= 1/2$ we recover the well-known *Schwartz inequality* for the euclidean distance. If we put $\alpha \to 1$ and $\beta \to \infty$ we obtain

$$|\langle \vec{v}, \vec{u} \rangle| \leq \|\vec{v}\|_1 \cdot \|\vec{u}\|_\infty, \tag{43}$$

where

$$\|\vec{v}\|_1 = \sum_{i \leq n} v_i; \text{ and } \|\vec{u}\|_\infty = \max_i \{u_i\}, \tag{44}$$

being the last one the so-called *Chebyshev's norm*. Now we want to apply this machinery to our problem. The key point is to realize that $\mathcal{I}(\mathbf{AB})$ can be expressed as a scalar product between two vectors, having the first one coordinates $J_{11}(\mathbf{AB})$, …, $J_{nn}(\mathbf{AB})$ and the second one $D(\Lambda_1(\mathbf{AB}) \| q)$, …, $D(\Lambda_n(\mathbf{AB}) \| p')$. We remark that this step is legitimated because all the terms involved in the computation are positive. Therefore, by applying the Hölder's inequality over the definition of $\mathcal{I}(\mathbf{AB})$, we have that

$$\mathcal{I}(\mathbf{AB}) = \sum_{i \leq n} J_{ii}(\mathbf{AB}) D(\Lambda_i(\mathbf{AB}) \| p')$$
$$\leq \left(\sum_{i \leq n} J_{ii}\right) \left(\max_i \{D(\Lambda_i(\mathbf{AB}) \| p')\}\right) \tag{45}$$
$$= \theta_{\mathbf{AB}} \left(\max_i \{D(\Lambda_i(\mathbf{AB}) \| p')\}\right),$$

being $\theta_{\mathbf{AB}}$ defined in equation (11). Now we observe that the probability of error in referentiating a given event of Ω is $p_e = 1 - \theta_{\mathbf{AB}}$. This enables us to use Fano's inequality to bound $\theta_{\mathbf{AB}}$:

$$\theta_{\mathbf{AB}} \leq \left(1 - \frac{H\left(X_\Omega | X'_\Omega\right) - 1}{\log(n-1)}\right), \tag{46}$$

thereby obtaining the desired result.

Lemma 3.- (Derivation of inequality (27)). Let $\mathbf{AB}$ be two agents sharing the world Ω and such that $(\forall m_i \in \Omega) p(m_i) = 1/n$ and that the channel defined by $\Lambda(\mathbf{AB})$ is symmetric. Then, the following inequality holds:

$$\mathcal{I}(\mathbf{AB}) < \frac{I^2\left(X_\Omega : X'_\Omega\right)}{H(X_\Omega)} + 1. \tag{47}$$

Proof.- The first issue is to show that, if $(\forall m_i \in \Omega) p(m_i) = 1/n$ and the channel defined by $\Lambda(\mathbf{AB})$ is symmetric, then $(\forall m_i \in \Omega) D(\Lambda_i(\mathbf{AB}) \| q) = I\left(X_\Omega : X'_\Omega\right)$. Indeed, since the channel is symmetric $p = p'$ and thus $H(X_\Omega) = H(X'_\Omega) = \log n$. Then take any $m_i \in \Omega$ and compute $D(\Lambda_i(\mathbf{AB}) \| p')$:

$$D(\Lambda_i(\mathbf{AB}) \| q) = \sum_{j \leq n} \Lambda_{ij}(\mathbf{AB}) \log \Lambda_{ij}(\mathbf{AB}) + \log n$$
$$= \log n - H\left(X'_\Omega | X_\Omega = m_i\right)$$
$$= \log n - \sum_{i \leq n} \frac{1}{n} \sum_{j \leq n} H\left(X'_\Omega | X_\Omega = m_i\right) \tag{48}$$
$$= I\left(X_\Omega : X'_\Omega\right),$$

where in the third step we used the property that, in a symmetric channel, $(\forall m_i, m_j \in \Omega) H\left(X'_\Omega | X_\Omega = m_i\right) = H\left(X'_\Omega | X_\Omega = m_j\right)$. Thus, if we average a constant value, we obtain such a value as the outcome (last step). Then, we apply inequality (41):

$$\mathcal{I}(\mathbf{AB}) \leq \left(1 - \frac{H\left(X_\Omega | X'_\Omega\right) - 1}{\log(n-1)}\right) I\left(X_\Omega : X'_\Omega\right)$$
$$< \left(1 - \frac{H\left(X_\Omega | X'_\Omega\right) - 1}{H(X_\Omega)}\right) I\left(X_\Omega : X'_\Omega\right) \tag{49}$$
$$\leq \frac{I^2\left(X_\Omega : X'_\Omega\right)}{H(X_\Omega)} + 1,$$

where, in the second step we used the fact that $H(X_\Omega) = \log n > \log(n-1)$ and in the third step we bound the remaining term

$$\frac{I\left(X_\Omega : X'_\Omega\right)}{H(X_\Omega)} \leq 1, \tag{50}$$

since $I\left(X_\Omega : X'_\Omega\right) \leq H(X_\Omega)$, thus completing the proof.

Achieving self-consistency maximizing consistent information. The structure of the functional accounting for the amount of consistent information shared by two agents -equation (21)- can lead to the paradoxical situation in which high scores on $\mathcal{I}(\mathbf{A} : \mathbf{B})$ do not imply high values of $\mathcal{I}(\mathbf{A} : \mathbf{A})$ or $\mathcal{I}(\mathbf{B} : \mathbf{B})$. In brief, the degeneracy of possible optimal configurations seems to jeopardize self-understanding even in the case in which communication is optimal. Interestingly, this apparent paradox can be

ruled out if we consider a population of agents, for several representative cases, as demonstrated in[29] using a version of θ_{AB}. For the particular case where $\eta_{AB} = 0$, we have seen at the beginning of this section that $\mathcal{I}(\mathbf{AB}) \le I(X_\Omega : X'_\Omega)$, having the equality only in the special case in which $\Lambda(\mathbf{AB}) = \delta_{n \times n}$, which, in turn, implies that $\mathcal{I}(\mathbf{AB}) = H(X_\Omega)$. The interesting issue is that in the presence of three or more agents $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$:

$$\left. \begin{array}{l} \mathcal{I}(\mathbf{A} : \mathbf{B}) = H(X_\Omega) \\ \mathcal{I}(\mathbf{A} : \mathbf{C}) = H(X_\Omega) \\ \mathcal{I}(\mathbf{B} : \mathbf{C}) = H(X_\Omega) \end{array} \right\} \Rightarrow \begin{array}{l} \mathcal{I}(\mathbf{A} : \mathbf{A}) = H(X_\Omega) \\ \mathcal{I}(\mathbf{B} : \mathbf{B}) = H(X_\Omega) \\ \mathcal{I}(\mathbf{C} : \mathbf{C}) = H(X_\Omega) \end{array} \quad (51)$$

i.e., maximizing the communicative success over a population of agents results automatically in a population of self-consistent agents, although there is no a-priori correlation between the coder and the decoder module of a given agent. Now we rigorously demonstrate this statement.

Lemma 3.- Let us have three $\mathbf{A}_i$, $\mathbf{A}_j$, $\mathbf{A}_k$ agents communicatively interacting and sharing the world Ω. Then, if $(\forall i < k)\mathcal{I}(\mathbf{A}_i : \mathbf{A}_k) = H(X_\Omega)$, then $(\forall i)\mathcal{I}(\mathbf{A}_i : \mathbf{A}_i) = H(X_\Omega)$.

Proof.- We observe, as discussed above, that the premise only holds if $(\forall i < k)$

$$\mathbf{P}^{\mathbf{A}_i}, \mathbf{Q}^{\mathbf{A}_i}, \Lambda, \mathbf{P}^{\mathbf{A}_k}, \mathbf{Q}^{\mathbf{A}_k} \in \Pi_{n \times n}, \quad (52)$$

and

$$\mathbf{P}^{\mathbf{A}_i} = \left(\Lambda \mathbf{Q}^{\mathbf{A}_k}\right)^T \wedge \mathbf{P}^{\mathbf{A}_k} = \left(\Lambda \mathbf{Q}^{\mathbf{A}_i}\right)^T. \quad (53)$$

Now we observe that, if $\mathcal{I}(\mathbf{A}_i : \mathbf{A}_k) = H(X_\Omega)$, $\mathcal{I}(\mathbf{A}_i : \mathbf{A}_j) = H(X_\Omega)$, we conclude that:

$$\mathbf{Q}^{\mathbf{A}_k} = \mathbf{Q}^{\mathbf{A}_j}; \wedge \mathbf{P}^{\mathbf{A}_k} = \mathbf{P}^{\mathbf{A}_j} \quad (54)$$

i.e., $\mathbf{A}_k = \mathbf{A}_j$. Now, knowing that $\mathcal{I}\left(\mathbf{A}_k : \mathbf{A}_j\right) = H(X_\Omega)$, then:

$$\mathcal{I}\left(\mathbf{A}_k : \mathbf{A}_k\right) = H(X_\Omega). \quad (55)$$

We can easily generalize this reasoning to an arbitrarily large number of communicating agents.

1. Schuster, P. The Origin and Processing of Biological Information. *Front. Life* **1**, 329 (2001).
2. Maynard Smith, J. & Száthmáry, E. *The major transitions in evolution.* (Oxford U. Press, Oxford, UK, 1997).
3. Maynard Smith, J. The Concept of Information in Biology. *Phil. of Sci.* **67**, 177 (2000).
4. Hopfield, J. Physics, computation, and why biology looks so different. *J. Theor. l Biol.* **171**, 53 (1994).
5. Shannon, C. E. A Mathematical theory of Communication (I). *Bell Syst. Tech. J.* **27**, 379 (1948).
6. Bialek, W. *Biophysics: searching for principles.* (Princeton University Press, Princeton, NJ, 2012).
7. Saussure, F. *Cours de Linguistique Générale.* (Bibliothèque scientifique Payot. Paris, France, 1916).
8. Atlan, H. Self creation of meaning. *Physica Scripta* **36**, 563 (1987).
9. Atlan, H. & Cohen, I. R. Immune information, self-organization and meaning. *Intl. Immunology* **6**, 711 (1998).
10. Hurford, J. Biological evolution of the Saussurean sign as a component of the language acquisition device. *Lingua* **77**, 187 (1989).
11. Komarova, N. L. & Niyogi, P. Optimizing the mutual intelligibility of linguistic agents in a shared world. *Art. Int.* **154**, 1 (2004).
12. Niyogi, P. *The Computational Nature of Language Learning and Evolution.* (MIT Press, Cambridge, Mass, 2006).
13. Nowak, M. A. Evolutionary Biology of Language. *Phil. Trans.: Biol. Sci.* **355**, 1615 (2000).
14. Nowak, M. A. & Krakauer, D. The Evolution of Language. *Proc. Nat. Acad. Sci. USA* **96**, 8028 (1999).
15. Plotkin, J. B. & Nowak, M. A. Language Evolution and Information Theory. *J. Theor. Biol.* **205**, 147 (2000).
16. Cangelosi, A. & Parisi, D. *Simulating the Evolution of Language.* (Springer, London, UK, 2000).
17. Floreano, D., Mitri, S., Magnenat, S. & Keller, L. Evolutionary Conditions for the Emergence of Communication in Robots. *Curr. Biol.* **17**, 514 (2007).
18. Steels, L. Language games for Autonomous Robots. *IEEE Int. Syst.* **16**, 16 (2001).
19. Steels, L. & Baillie, J.-C. Shared grounding of event descriptions by autonomous robots. *Robotics and Autonom. Syst.* **43**, 163 (2003).
20. Steels, L. Evolving grounded communication for robots. *Trends. Cogn. Sci.* **7**, 308 (2003).
21. Steels, L. The emergence and evolution of linguistic structure: from lexical to grammatical communication systems. *Connect. Sci.* **17**, 213 (2003).
22. Nolfi, S. & Mirolli, M. *Evolution of Communication and Language in Embodied Agents.* (Springer Verlag, Berlin, Germany, 2010).
23. Ash, R. B. *Information Theory.* (New York, Dover, NY, 1990).
24. Cover, T. M. & Thomas, J.-A. *Elements of Information Theory* (John Wiley and Sons, New York, NY, 2001).
25. Gleick, J. *The information: A history, a theory, a flood.* (Vintage, New York, NY, 2012).
26. Bickerton, D. *Language and Species.* (Chicago University Press, Chicago, IL, 1990).
27. Christiansen, M.-H. & Kirby, S. Language Evolution: Consensus and Controversies. *Trends. Cogn. Sci.* **7**, 300 (2003).
28. Deacon, T. W. *The Symbolic Species: The Co-Evolution of Language and the Brain* (Norton, New York, NY, 1997).
29. Corominas-Murtra, B. & Solé, R. V. Nework Topology and self-consistency in Language Games. *J. Theor. Biol.* **241**, 438 (2006).

Acknowledgments

Author contributions

B.C.-M., J.F and R.S. designed research, made analytic derivations and wrote the paper.

Additional information

Competing financial interests: The authors declare no competing financial interests.