PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Research

Cite this article: Andrade-Lotero EJ, Ortiz-Duque JM, Velasco-García JA, Goldstone RL. 2022 The division of linguistic labour for offloading conceptual understanding. *Phil. Trans. R. Soc. B* **378**: 20210360. https://doi.org/10.1098/rstb.2021.0360

Received: 29 November 2021
Accepted: 8 September 2022

One contribution of 23 to a theme issue 'Concepts in interaction: social engagement and inner experiences'.

Subject Areas:
cognition, behaviour

Keywords:
group cognition, cognitive artefacts, community of knowledge, semantic externalism, extended mind

Author for correspondence:
Robert L. Goldstone
e-mail: rgoldsto@indiana.edu

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.6251589.

The division of linguistic labour for offloading conceptual understanding

Edgar J. Andrade-Lotero[1], Julián M. Ortiz-Duque[2], Javier A. Velasco-García[2] and Robert L. Goldstone[3]

[1]School of Engineering, Science and Technology, and [2]School of Humanities, Universidad del Rosario, Bogotá, 111711, Colombia
[3]Department of Psychological and Brain Sciences and Program in Cognitive Science, Indiana University, Bloomington, IN, 47401, USA

RLG, 0000-0001-8357-8358

The division of linguistic labour (DLL), initially theorized by philosophers, has gained the attention of cognitive scientists in the last decade. Contrary to some controversial philosophical accounts of DLL, we propose that it is an extended mind strategy of offloading conceptual understanding onto other people. In this article, we empirically explore this proposal by providing an exploratory experimental paradigm to search for the mechanisms underwriting DLL and how they may work in practice. We developed a between-subjects experiment in which participants had to categorize two pairs of highly confusable dog breeds after receiving categorization training on just one pair of breeds. In the treatment group, participants were grouped in dyads and were allowed to interact with each other by means of the labels of these four dog breeds. In their queries to trained 'experts', novices frequently used labels to refer to breeds that they could not identify themselves. Experts were highly responsive to their paired novices' queries, and the rates of querying for the two members within a dyad were positively correlated. Independent categorization failure and offloading categorization success lead to subsequent increases in querying by novices, indicating adaptive use of offloading. Self-reports of breed knowledge were higher for experts within a dyad compared to isolated experts.

This article is part of the theme issue 'Concepts in interaction: social engagement and inner experiences'.

1. Introduction

In his famous paper *The Meaning of 'Meaning'* [1], Hilary Putnam presented the thesis that the meanings of some of a speaker's words cannot be in their head, because they might not have the necessary knowledge to recognize the objects referred to by these words. However, these words mean what they mean even when uttered by an uninformed speaker, and Putnam suggests that this is because such a speaker may delegate the recognition task to other speakers in their community who do possess the appropriate knowledge: the experts. This dependency of novices on experts for the identification of the referents of some words is known as the division of linguistic labour (DLL).

We intend our DLL framework to be sufficiently broad to capture situations in which the members of a community do not necessarily differ in their overall level of knowledge, but rather in their perspective or kind of knowledge. For example, chemists and firefighters each have their own expertise relative to the other when it comes to fire. Our framework does critically invoke a relative 'novice' offloading cognitive tasks such as recognizing, understanding and combatting fire onto a relative 'expert.' However, these designations of expertise must be understood relative to a particular context involving fire, such as understanding it at a molecular level versus eliminating it. Thus, when we use the term 'novice' it should be seen as a shorthand gloss for the more nuanced notion of a person whose current knowledge about something is insufficient to allow them to solve a currently required task involving that thing.

Although Putnam's arguments in the aforementioned paper centred mainly on natural kind words such as *water* or *gold*, it has been duly noted that DLL is a phenomenon that extends to all sorts of technical words (see [1, p. 160f.]; see also [2,3]). However, despite its ubiquity, DLL often goes unnoticed, since speakers do not usually think of themselves as novices. Indeed, a number of experiments have shown that there is a gap between what people feel they know about something and what they are actually able to explain about it [4–11]. Perhaps we feel that we know more about something because our layman concepts are quite sufficient to get by. However, more often than not, these layman concepts are sufficient *because* our everyday tasks are supported by experts who allow us to succeed at them.

Besides the aforementioned extension of the reach of the DLL phenomenon to all sorts of technical words, some philosophers have stretched DLL in a more striking and controversial manner. From its initial formulation about the meaning of words, DLL has been taken to show something about the content of our thoughts [3,12,13]. For example, in Tyler Burge's terminology [3], the novice speaker *possesses* concepts that they incompletely understand. In contrast with this controversial thesis, we maintain that, at its core, DLL is better viewed as a variety of cognitive offloading, that is, the widely documented strategies that people resort to for enhancing their internal cognitive limitations (e.g. writing down a shopping list, setting alarms on one's mobile phone, etc.) or for easing their information processing demands when engaged in a task (e.g. counting with one's fingers, rotating one's head to better read rotated text, etc.) [14–17]. There is, however, an important qualification. Many strategies discussed in this literature refer to actions involving the physical environment around the individual (a crucial exception being the phenomenon of transactive memory systems (TMS), to which we will come back later). Robinson Crusoe could use these sorts of strategies even when he was stranded on his forsaken island. However, Mr Crusoe would need to be around the relevant experts in order to use DLL (at least in the way in which we will be suggesting presently). The novice's utterances of a word can be seen as a social technology that requires structured interactions with other individuals [18]. In this way, a foundational part of DLL is a type of cognitive offloading in which the specific component of the environment that is used to extend one's mind [15] is other people. Given that a major part of a person's environment is the other people surrounding them, offloading onto others is a major resource that people have in achieving intelligent behaviour [6,19].

The DLL has been studied mainly in philosophical circles, but, to our knowledge, not from an empirical perspective. (Some empirical works [8,20] get close to it but do not directly study DLL.) We have developed an exploratory experiment in which the conditions for the emergence of division in a cognitive task, mediated by language, are instantiated in an image classification task. In the experiment, participants received points for correctly labelling dogs. Accurate labelling required participants to draw distinctions between two highly confusable pairs of breeds. Half of the participants were in a solitary condition in which they had to label the breeds on their own. The other half was in a paired condition where one participant was trained to distinguish between the two breeds in one pair while the other participant was trained to distinguish between the two breeds in the other pair. After training, paired participants could ask their partner whether a label was correctly assigned to a dog image that they were tasked with labelling. We are interested in exploring three things. Is the 'extended strategy' of asking one's paired partner a reliable way of easing one's own cognitive demand of learning a dog's correct classification? What are some of the factors that trigger the use of the extended strategy as it competes with the 'internal strategy' of labelling dogs by oneself? And what are the consequences of engaging in the extended strategy for the metacognitive process of perceiving one's own understanding?

The paper is structured as follows. In §2, we categorize different scenarios of DLL focusing on those in which it is natural to see the novice speaker's use of words as founded in a form of cognitive offloading. We also show how our experimental paradigm fits in with Risko's and Gilbert's metacognitive model for cognitive offloading [16]. In §3, we introduce the details of our experimental paradigm. The results of this experiment, presented in §4 and discussed in §5, clearly suggest that people's metacognitive processes when engaging with DLL accord with those of other forms of cognitive offloading.

2. A novice's utterances as stemming from cognitive offloading

(a) Novice utterances as embedded in social environments

There are situations in which we can communicate with words that we do not understand at the expert level. We can do this because our use of these words, considered within the framework of concrete social relations, takes advantage of the experts' understanding. The following example is adapted from [21]. Consider a man, say John, who is using his personal computer and realizes that it is too slow. He then turns to the technician at his office and asks them why this is happening. They tell John that one of the random-access memory (RAM) cards must be malfunctioning and that he needs a new one. Then, John goes to a service centre and requests a new RAM for his computer. The next day, John picks up his computer, tries it out and notices a certain increase in speed, but since he doesn't trust the service centre, he asks the technician at his office to check whether or not they installed a new RAM. Since John cannot identify RAM, he turns to his trusted expert to help him make the required identifications. And when he says 'RAM,' he says it knowing that his trusted technician will verify it for him. Note that this word in John's mouth is not linked to its referent by an internal characterization that John has in his own head. Rather the link exists because someone else makes it for him and so informs him.

Observe that when John went to the service centre to ask for his computer's RAM to be replaced, he had the linguistic capacity to use the word 'RAM' to mean exactly RAM. If instead of changing the computer's RAM, the service centre had overclocked the central processing unit in the computer (with the consequence that its speed improved) and had the technician at his office informed John of this, John would be entitled to demand that the service centre replace the computer's RAM as requested. John is using 'RAM' in an effectual manner despite not possessing the ability to define or even identify RAM.

Situations like this occur in our daily lives not only regarding our computers, but also our cars, home appliances, mobile phones, medical checkups, legal issues and so on. In all these situations, we use words even though their meanings are vague and fragmentary to us. We trust (or take for granted) that an expert will, immediately or at a later point, intervene to fill in the gap in our understanding to complete the task at hand.

To be sure, these are by no means all of the kinds of situations in which we use words we do not completely understand. We might use words, such as MRI or PCR, to tell other fellow novices the kind of tests we were asked to undergo, without any of us knowing exactly how to define such tests. More importantly, we often do not know anyone around us who actually can. This is a different kind of situation from our example with RAM. The difference lies in the proximity of the expert to which one can defer. So there are situations with immediate or direct interaction with an expert, and situations where no expert is directly available. We call the latter kind of situation *idle chatter*. The distinction is important because idle chatter could be explained by appealing to concepts such as 'layman meaning' and 'common ground,' since there is no need to talk about interactive cognitive offloading onto others. However, when an expert is immediately at hand, and they are required for their expertise, we need an alternative set of concepts to account for the situation, as we explain in the following section. This is why we focus on situations in which an expert is available to help complete a task.

(b) Common ground and pooled ground

Now, we would like to contrast our own perspective on communication with a more standard approach, which is usually conceived as a way to equalize the participants' level of understanding, i.e. the *common ground* perspective. Indeed, works such as Clark's [22] aim at characterizing communication as a process that requires a shared background of knowledge for the activity to be successful. From this perspective, what a speaker seeks is to align their knowledge with their interlocutor's knowledge.

This conceptualization is in contrast with our approach to DLL. In this case, speakers not only do not share a common background of beliefs, but neither do they seek to align their differences through communication. Rather, their goal is to make use of cognitive specialization to increase the efficiency of their own cognitive processes. For example, in the aforementioned case of John and his computer's RAM, John is not looking for any object that fits with his insufficient comprehension of RAM, but for the actual RAM. In this case, an equalization of participants' level of understanding would suppose either for the expert to understand RAM in some vague sense of 'something that makes the computer go faster' or for John to increase his knowledge about RAM, a very costly action for a mere computer repair. However, John is asking for a specific thing, new RAM (not, e.g. a faster configuration), without being able to identify a RAM or knowing anything further about its functioning.

This situation does not fit with the common ground perspective, but it does fit with a more recent perspective on communication, i.e. the *pooled ground* perspective. Here, communication occurs against the background of a functional joint task, in which participants do not possess homogeneous information nor do they seek it. As Raczaszek-Leonardi *et al.* point out:

> [The pooled ground is the] aggregate of the common ground and the relevant privileged ground that may never enter common ground (become mutual) yet is a basis for individual behavior influencing the [conversing] dyad. To pool knowledge in coordinative situations, language is thus used not only to confirm a shared vision of a situation, but also to 'scout' for and signalize mutually unavailable resources (information or skills), which would enable efficient functioning of the global system. [23, p. 7]

Pooled ground is a useful concept with which to study DLL since it shows that some speakers rely on the understanding of experts to perform tasks that they could not perform without this support and where it would be inefficient to retrieve the information to perform them. It also shows that this division of labour is regulated by a principle of efficiency, namely, achieving the goal of the joint task with the least possible effort.

(c) Words as cognitive artefacts

The resourcefulness with which people find ways to enhance their memory limitations or ease their information processing demands when engaged in a task has been widely documented [14–16], for example, by setting alarms on their mobile phones or by using calculators to make calculations at the supermarket. Furthermore, there is a long literature stemming from Wegner's pioneering work [24] on TMS [10,11], which refers to structured collectives in which not everyone knows all of the information that the group collectively possesses, but its members know who knows what so that anyone can access the information when necessary. These are obvious examples of cognitive artefacts, but language itself can also be seen in this light. As suggested by Clark [25], words in a public language might serve as tools that complement individual cognitive resources.

Even if we emphasize the individually grounded experience of using words [26], we should also recognize that this is a socially regulated activity, as Borghi & Cimatti note [27] (see also [28]). This approach might prove fruitful not only for abstract concepts, Borghi and Cimatti's target phenomenon, but also for situations where the speaker can use a word even though they lack complete knowledge of its meaning. This is precisely the DLL phenomenon, which is, essentially, a mechanism that allows a speaker to reduce the amount of information they need to apprehend by delegating the recognition task to other members of their community.

In order to better appreciate our theoretical contribution to the explanation of DLL, we can contrast our approach with a traditional perspective. According to the latter, a novice indirectly connects words with their referents in the world through a mental act called 'semantic deference' [13]. This allows the novice to connect with the experts in their community and, through them, with the referent of the word. However, what does this mental act that connects an individual cognition with that of other agents consist in? In our 'words as cognitive artefacts' approach, this mental activity is explained (or better, replaced) by the first-person experience embedded in a social niche that allows the novice to: (i) recognize a material symbol, namely, a signifier (i.e. the physical form of a word); (ii) realize that this signifier is 'entrenched' in their society [8]; and (iii) use it to refer to something in their world (i.e. their physical and cultural reality).

In the traditional perspective, semantic deference has been attempted to be explained by the possession of a mental item called a deferential concept [29]. By contrast, our approach suggests that such mental activity must be replaced by a practical ability and a social background that allows the novice to use a signifier to communicate in the pooled ground. The niche that the novice inhabits contains people interacting with each other as well as physical artefacts with which cognitive tasks are performed. Some of these artefacts are the signifiers, which allow novices to perform cognitive tasks that they could not do on their own, by involving experts who contribute to the pooled ground with their information or skills, 'which would enable efficient functioning of the global system' ([23, p. 7]). In the example of John and his computer, we are considering the entire activity in which John fixes his computer. One part of it is the linguistic action of requesting a new RAM, and the other part is the technician's task of finding and replacing the RAM card, but our interest is in John's request. This is a type of cognitive activity that John could not do by himself because he does not know how to recognize RAM and without a social context his *requesting* activity would not be possible.

A speaker uses a word as a cognitive artefact when they can offload the need for personally learning and storing the information associated with the word (e.g. the recognition criteria for identifying the word's reference). Such information could be included in the pooled ground through the use of words and the involvement of the relevant experts. Here, 'offloading' means that the novice uses the external resources of the expert, without including them in any way in the novice's individual cognition, i.e. without acquiring or improving any cognitive skill; but it also implies a deeper philosophical sense, namely, that a term can be used without a complete understanding of its meaning (in this case, without being able to recognize the term's reference).

However, DLL allows for more than just externally storing information in another person and retrieving it later, as is the case with the phenomenon of TMS [24]. In order to pinpoint where our proposal goes beyond TMS (as it was originally conceived by Wegner), we can use the characterization of a task-dependent pooled ground [23, p. 8], which consists of A not knowing x, B knowing x and A knowing that B knows x. In a TMS, A makes a query about x to extract from B the information to know x. In contrast with this, in DLL a speaker can use this word to achieve the task by taking advantage of the expert's performance without ever knowing x. In the literature, there are some broader uses of the term TMS (e.g. [30,31]) to which our account of DLL may contribute.

(d) Some aspects of cognitive offloading strategies

We want to explore in an empirical way the claim that the use of words by novices can be seen as a form of cognitive offloading. To this effect, we will use a framework based on a proposal of Risko's and Gilbert's, which integrates the processes that trigger a strategy of cognitive offloading and the consequences of its use [16], in an attempt to characterize cognitive offloading strategies. There are three important aspects to consider. First, when agents are engaged in a task, resorting to such strategies typically helps them overcome gaps in understanding and/or minimize their computational effort. Second, the selection of an internal versus extended strategy is based on beliefs about, and/or past experiences with, these two options. Third, the agent's use of extended strategies tends to alter their thinking about their internal capacities such as, for example, self-assessment of understanding (figure 1).

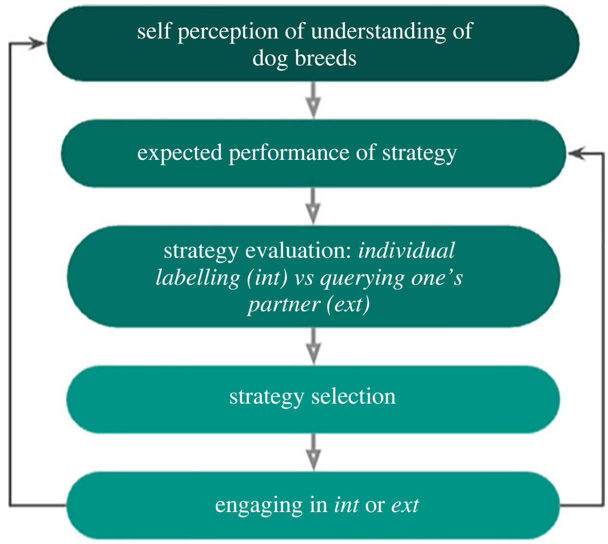

Figure 1. Framework for a description of DLL as a cognitive offloading strategy. Adapted from Risko's and Gilbert's framework for how cognitive offloading strategies depend on metacognitive aspects of the processes that trigger such strategies and the consequences of using them. (Online version in colour.)

Our paired treatment (see details below) matches this framework. A player has two choices for how to accomplish the task of classifying a dog. First, they could assign the most-likely dog breed based on their own ability. Here, individual expertise is privileged, either because it is trusted, or because one hopes to improve one's classification ability and practice is viewed as a means to that end. We will call this the *internal strategy*. Second, the player could ask a newly trained expert whether a classification is correct. Here the use of words is privileged since taking advantage of the expert's ability is seen as a greater benefit. We will call this option the *extended strategy*.

We believe that this experimental paradigm contributes to the emerging field of research on partial understanding. Keil [32], for example, points out the problem that normally our understanding of the world is very limited and yet it is still useful for us to get by. Here we propose that one mechanism by which the problem of incomplete understanding is solved in everyday life is the DLL. However, DLL as described above is an abstract philosophical idea without reference to the concrete mechanisms by which it is created or maintained. Our experimental paradigm seeks to recreate a situation of incomplete understanding in which there are incentives to form a DLL, to the effect of collecting exploratory data that allow insights about the cognitive and social elements supporting DLL. This internal versus external strategies framework that constrains our participants' behaviour does not embody a set of empirical predictions to be evaluated; rather, it provides a working test-bed to search for the mechanisms underwriting DLL and how they may work in practice.

3. Method

(a) Participants and procedure

Participants were 84 students from the Universidad del Rosario (40% male, 60% female; average age approximately equal to 20 years), who were invited by email to participate in the subject pool of the Rosario Experimental and Behavioural Economics Laboratory of the Universidad del Rosario. Participants were tested in six sessions; the first three corresponded to the paired condition and the last three to the solitary condition. For the sessions in pairs for which an odd number of participants were received, one participant was chosen to be included in the solitary condition. The numbers of participants per session were 3, 15, 24, 21, 8 and 13. The total number of participants in the paired condition was 40 and in the solitary condition was 44 (we had planned to collect at least 40 observations for each treatment). Before starting the session, each participant signed an informed consent document, which was previously approved by the Ethics Committee (social science room) of the Universidad del Rosario. Each participant carried out the test inside a sound- and sight-isolated workstation running a version of the experiment implemented in the *nodeGame* platform [33]. The test consisted of the presentation of instructions, followed by a brief quiz to confirm that the participants had understood the instructions in general. These provided the participants with knowledge about the game and the payment structure. After reading the instructions and passing the quiz, participants completed a short tutorial on the operation of the task, after which they performed 25 training rounds and 25 game rounds. Players were allotted 60 s to finish their guesses during each training round, and 90 s during each game round. At the end of the game rounds, participants were asked to rate their understanding of each category on a scale of 1 to 7 (1 = little or no understanding, 4 = moderate understanding, 7 = deep and detailed understanding), using a questionnaire adapted from [5].

For their participation in the experiment, participants received 10 experimental monetary units (EMUs; each unit equivalent to approximately 0.3 USD). Additionally, participants could earn bonus money based on their performance during the experiment. In both treatments, the payment received by each participant was determined by randomly choosing two game rounds and observing the score obtained in each of them. A score of 1 was rewarded with 2 EMUs; a score of 2 with 4 EMUs; a score of 3 with 6 EMUs; a score of 4 with 9 EMUs; and a score of 5 with 13 EMUs, so each participant could earn up to 26 EMUs additional to the show-up fee. Participants received their reward in cash immediately after the session.

(b) Materials

In §2(a), we described the following situation: John wants his computer to work again, goes down to a service centre and says 'Is my computer's RAM malfunctioning?', upon which a technician examines the computer and replaces the bad RAM card; so finally John goes back home to use his computer as he usually does. The intended scenario in our paradigm is similar: a player wants to earn a point by classifying a dog, so they ask their partner 'Is this dog a Cairn terrier?', upon which the partner provides a yes/no answer, which the player uses to classify the dog. Thus, we capture a kind of situation where the use of a word allows a speaker to achieve a goal even if they are not able to identify the referent of that word.

Regardless of the treatment, each participant was randomly assigned with equal probability to either the group trained to classify two hound breeds or the group trained to classify two terrier breeds (see the electronic supplementary material, figure S1). During the training rounds (see the electronic supplementary material, figure S2 top), each participant had to learn by trial and error how to correctly classify five dogs that appeared on their screen. However, participants being trained to discriminate between the two hound breeds were provided with a number of explicit characteristics that only Irish wolfhounds have (e.g. their hair colour and build), and participants being trained to discriminate between the two terrier breeds were given some explicit clues about Norwich terriers. At the end of the round, participants were told whether they had correctly classified each dog, but they were not provided with the correct label (see the electronic supplementary material, figure S2 bottom panel). In the paired condition, members of each pair were trained on complementary breeds (i.e. one participant was trained on hounds and the other on terriers). This information was explicitly given to the participants.

After the training rounds, players went on to the game rounds (figure 3). There the task changed slightly, as the number of categories available to classify each of five dogs increased from two to four. The five dogs, in turn, were also mixed (i.e. there were both hounds and terriers). The task at this stage was the same—to correctly classify each of the five dogs into one of the available categories. Both paired players had exactly the same dogs on their screens during each round. In addition, for paired players, communication was enabled through a button with which each participant could ask their partner if a certain dog belonged to a specific category. The player could ask 'Is this dog…?' and fill in the blank with the label of one of the four breeds. To that question, the partner could only answer 'correct,' 'incorrect' or 'I don't know.'

At the end of each round, both in the training and game rounds, each participant received feedback on their performance during the round.

(c) Measures

We can measure a player's understanding of a word by their accuracy of classifications made on their own, that is, when players did not query their partner or when they were not paired with a partner. Now, the novice's use of a word (the extended strategy) can be measured by the proportion of dogs about which the player queried their partner. Moreover, the effectiveness of such a strategy can be measured by the accuracy of classifications performed after the player queried their partner. We expect paired players to classify dogs better when they query their expert and use their response compared to when they decide to use their own best guess. Moreover, since players can 'opt out' from the extended strategy, we expect players to be sensitive to its benefit for dog classification and adhere more to it when they achieve low accuracy on their own.

4. Results

(a) Classification accuracy per treatment

To begin with, we verified that participants reached an adequate level of expertise in their respective *expert dog* breeds. (We will use the term 'a participant's expert dogs' to refer to the breeds of dogs on which the participant received training. Similarly, the term 'a participant's novice dogs' refers to breeds on which the participant did not receive training.) In figure 2*a*, we can see that after 25 rounds of training, participants achieved more than 95% accuracy ($M_{\text{paired}}^{\text{rounds 24–25}} = 0.97$, $\text{s.d.}_{\text{paired}}^{\text{rounds 24–25}} = 0.16$, $M_{\text{single}}^{\text{rounds 24–25}} = 0.95$, $\text{s.d.}_{\text{single}}^{\text{rounds 24–25}} = 0.21$) at their expert dogs, and this result is comparable in both treatments. To check the significance of this as well as all comparisons in this paragraph, a logistic mixed-effects regression was calculated to predict the participant's accuracy based on treatment and round (fixed effects), player, and item (random effects). The details and justification of this model are presented in the electronic supplementary material, S2.1. The contribution of treatment (baseline = paired condition) is not statistically significant ($\beta = -0.69$, $p = 0.107$; see the electronic supplementary material, table 1

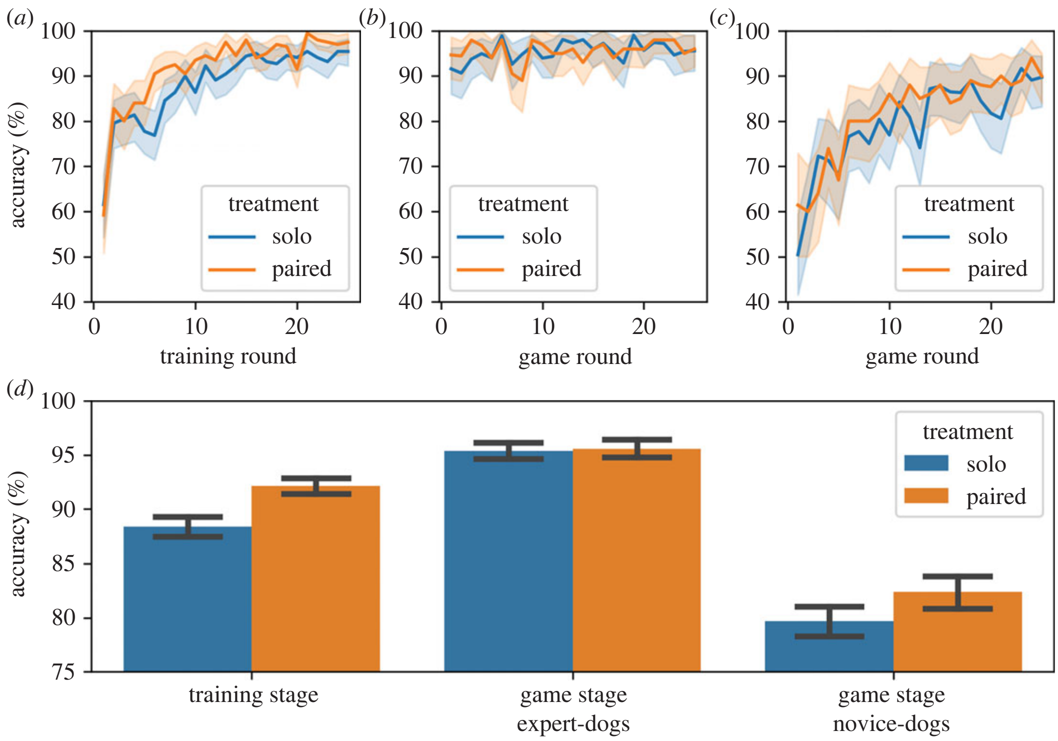

Figure 2. Dog classification accuracy. The vertical axis shows the average classification accuracy (1 = correct, 0 = incorrect; displayed as a percentage) of the five dogs presented to each player per round, averaged over all players. (*a*) Training rounds, indicating that training was successful in teaching participants how to classify trained breeds. (*b*,*c*) Classification accuracy for expert (centre) and novice dogs (right) during game rounds. The shaded regions correspond to a 95% confidence interval of the percentage estimator. (*d*) Accuracy for paired and solo conditions over all training and game rounds, separately considering expert and novice dogs. (Online version in colour.)

top left). It is interesting to note, however, that accuracy over all training rounds is higher for paired participants than for solo participants ($M_{paired} = 0.92$, $s.d._{paired} = 0.27$, $M_{single} = 0.88$, $s.d._{single} = 0.32$, $\beta = -0.37$, $p = 0.049$; see the electronic supplementary material, table S1 top right).

We also verified the classification accuracy during game rounds. First, *expert* dog classification was maintained around 95% and this result is comparable in both treatments ($M_{paired} = 0.96$, $s.d._{paired} = 0.21$, $M_{single} = 0.95$, $s.d._{single} = 0.21$, $\beta = -0.17$, $p = 0.594$; see the electronic supplementary material, table S1 bottom left), which can be appreciated in figure 2*b*,*d*. In figure 2*c*, we observe the development of accuracy for novice dogs, in which paired novices did a comparable job with respect to solitary novices ($M_{paired} = 0.82$, $s.d._{paired} = 0.38$, $M_{single} = 0.8$, $s.d._{single} = 0.4$, $\beta = -0.15$, $p = 0.395$; see the electronic supplementary material, table S1 bottom right). We see that being in a dyad does not contribute to a difference in accuracy with respect to participants that performed the task on their own. However, this result is not conclusive about the benefit of using the extended, rather than the internal, strategy. Indeed, we need to take a closer look at the extent to which the extended strategy was actually used and how well it performed when it was used.

(b) Amount of queries

We examine now the extent to which participants in the paired treatment actually used the extended strategy. Figure 3*a* shows the proportion of dogs of each breed on which the novice made a query to the expert. Participants asked their partners about 40% of their novice dogs, in contrast to less than 1% of their expert dogs. Observe also that the percentage of queries about novice dogs decreased somewhat as the rounds progressed, but novices continued to query their expert partner a significant percentage of times per round, as can be seen in figure 3*b*. While participants' engagement in communication was somewhat modest in their role as novices, the data suggest that, in their role as experts, they were much more active in answering queries. Figure 3*c* shows that over 90% of queries were answered, and of these over 90% were answered correctly. We examined the relationship between the number of queries from both partners within a dyad. In figure 3*d*, we plot each dyad with respect to the number of queries asked by its most and least active player. We can see that the querying activity of both partners is strongly correlated ($r_{18} = 0.948$, $p < 0.001$). Each dyad is size coded according to their average accuracy.

(c) Accuracy and feedback loop for strategy selection

More accurate classifications occurred when players queried their partner than when they did not, as can be seen in figure 4*a*. To check the significance of this comparison, a logistic mixed-effects regression was calculated to predict the participant's accuracy based on whether they queried their partner (queried; fixed effect; see details in the electronic supplementary material, S2.2). A significant main effect of queried was found ($M_{queried = yes} = 0.88$, $s.d._{queried = yes} = 0.32$, $M_{queried = no} = 0.79$, $s.d._{queried = no} = 0.41$, $\beta = 0.707$, $p = 0.003$; see the electronic supplementary material, table S2) showing that the extended strategy of querying the expert seems to provide higher accuracy than the internal strategy of labelling dogs without asking the expert. However,

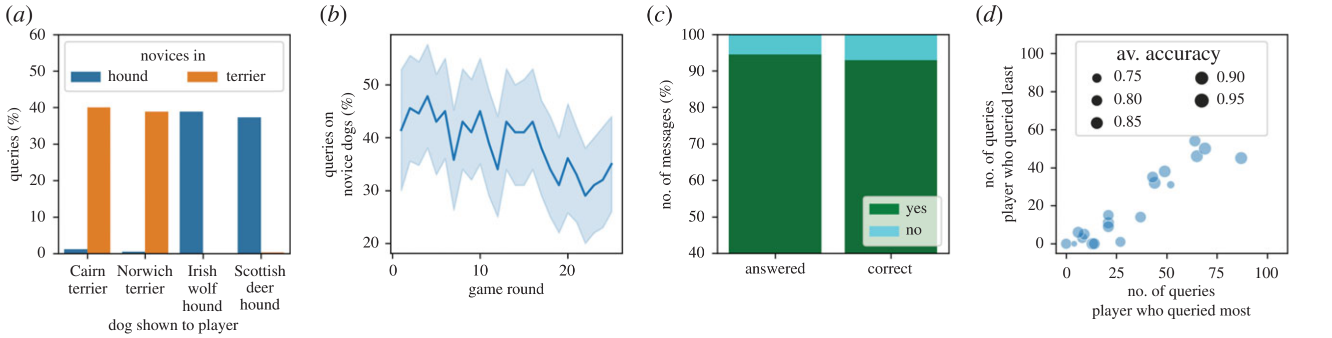

Figure 3. Amount of use of the DLL strategy. (*a*) Percentage of dogs on which players made a query, separated for each of the breeds and by each type of novice dog, indicating that novices query their expert partners about 40% of the time. (*b*) Percentage of dogs that novices queried their partner about per round of play (shaded region corresponds to a 95% confidence interval). (*c*) Percentage of messages answered by experts and accuracy of the answers. (*d*) Relationship of number of queries made by players within a dyad. The horizontal axis indicates the number of queries made throughout all rounds by the most active player in the dyad; the vertical axis corresponds to the least active player. Each data point corresponds to a dyad ($n = 20$) and the size corresponds to the dyad's average accuracy. (Online version in colour.)

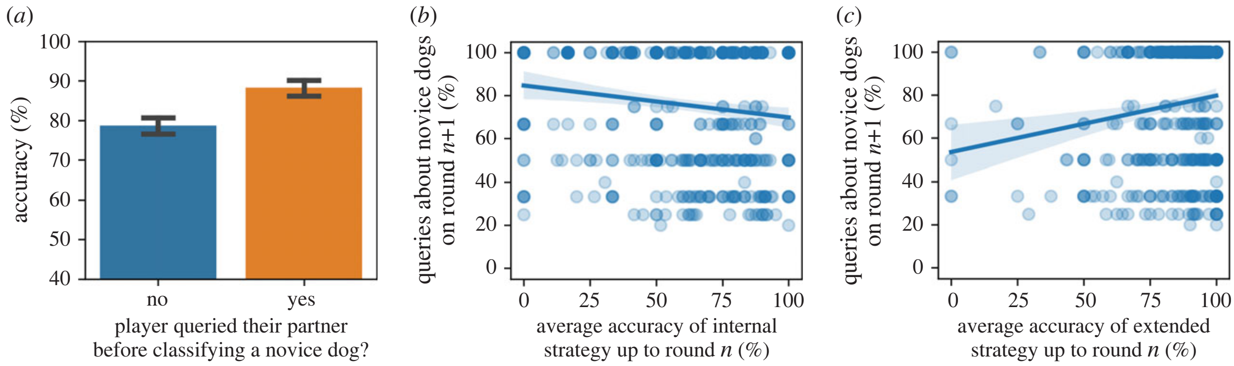

Figure 4. Accuracy of the DLL strategy and feedback loops for strategy selection. (*a*) Player's classification accuracy versus whether the player queried their partner before classifying a novice dog. (*b*) Percentage of novice dogs that player queried their partner about on round $n+1$ versus average accuracy up to round n on rounds when a player used an internal strategy (i.e. not querying their partner). (*c*) Percentage of dogs that a novice queried their partner about on round $n+1$ versus average accuracy up to round n on rounds when a player used an external strategy (i.e. querying their partner). Each data point corresponds to a player's behaviour in a given round ($n = 952$). (Online version in colour.)

it is important to note that participants with different knowledge levels query differently [34], and since we did not control for query in our experiment, we cannot rule out a confounding variable that could be causing both query and accuracy. Further experiments are required to properly control for query.[1]

We also consider here whether participants weigh the costs and benefits of the two strategies on each round. A linear mixed-effects regression was calculated to predict the percentage of novice dogs that a participant queries about on round $n+1$ (query$_{n+1}$) based on the average accuracy of labelling a novice dog without asking the expert (internal strategy) up to round n (acc$_{\text{internal}}$; fixed effect) and player and round (random effects; see the electronic supplementary material, S2.2). A significant regression was found, showing that the percentage of novice dogs the participant will ask about on the next round is lower when the novice previously obtained higher levels of accuracy without querying the expert ($\beta = -0.102$, $p = 0.002$; see the electronic supplementary material, table S3). This influence can be visually appreciated in the centre panel of figure 4. Observe that similar qualms about querying performance not being controlled can arise in the foregoing interpretation.

A second linear mixed-effects regression was calculated to predict the percentage of novice dogs that a participant asks about on round $n+1$ (query$_{n+1}$) based on the average accuracy of labelling a novice dog after asking the expert (extended strategy) up to round n (acc$_{\text{external}}$; fixed effect) and player and round (random effects). A significant regression was found, which shows that participants with higher levels of accuracy with the external strategy tend to re-select it on the next round ($\beta = 0.333$, $p = 0.013$; see the electronic supplementary material, table S4). This influence can be visually appreciated in figure 4*c*. Finally, note that a similar qualm applies in this interpretation as in the previous paragraph.

(d) Influence on the report of understanding

Now we consider the results of the reports on category understanding, according to the questionnaire that participants answered at the end of the experiment. In figure 5*a*, we see the report distributions. Focusing first on experts, we observe that paired participants reported on average a greater understanding of their expert dogs than solitary participants ($M_{\text{paired}} = 6.22$, $\text{s.d.}_{\text{paired}} = 1.11$, $M_{\text{single}} = 5.32$,

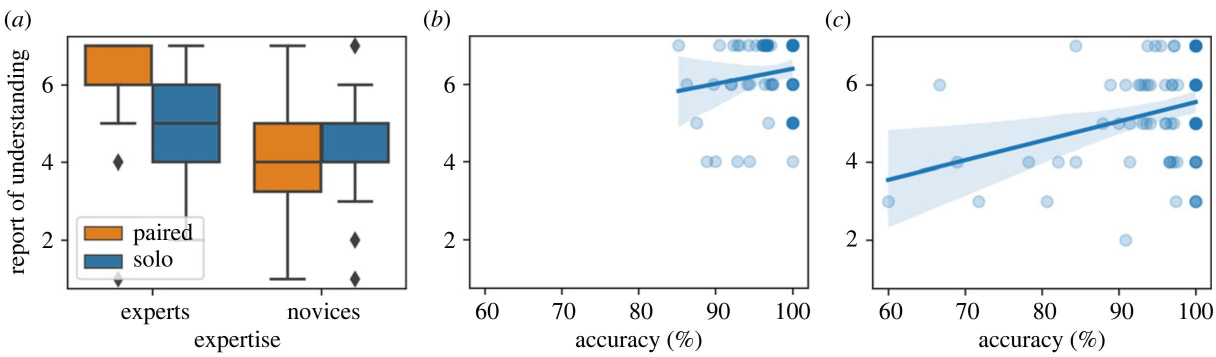

Figure 5. Reports on category understanding. (*a*) Distribution of ratings in self-reports of understanding, grouped according to treatment and separated by experts and novices. (*b*,*c*) Influence of accuracy on reports in the case of experts for paired participants (centre, $n = 77$) and solo participants (right, $n = 88$). Each data point represents each expert dog breed for each player. (Online version in colour.)

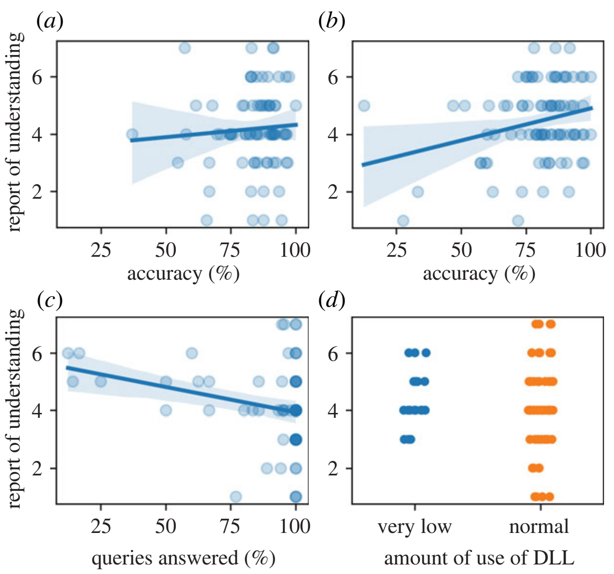

Figure 6. Reports on category understanding by novices. (*a*,*b*) report versus classification accuracy for participants in the paired (left, $n = 88$) and the solitary (right, $n = 78$) conditions. (*c*) In the paired condition, the report versus the expert's response rate to the queries made by the novice ($n = 67$). (*d*) reports versus percentage of novice dogs queried by the player, grouped by 'very low' (first quartile) and 'normal' (rest of data) ($n = 78$). Each data point represents each expert dog breed for each player. (Online version in colour.)

$\text{s.d.}_{\text{single}} = 1.31$, $t_{164} = 4.786$, $p < 0.001$, note: a non-parametric *U*-test agrees with the *t*-test).[2] This is interesting because the difference in experts' performance in both treatments is not statistically significant, as apparent in figure 2*b*. Observe that accuracy does not make a significant contribution to the self-reported understanding for paired experts ($r_{75} = 0.153$, $p = 0.183$; figure 5*b*),[3] as compared to a significant influence in the solitary condition ($r_{86} = 0.3$, $p = 0.004$; figure 5*c*).

In figure 6*a*, we can see the distribution of the reports of understanding by *novices* in both treatments. The averages are not significantly different ($M_{\text{paired}} = 4.18$, $\text{s.d.}_{\text{paired}} = 1.39$, $M_{\text{single}} = 4.44$, $\text{s.d.}_{\text{single}} = 1.33$, $t_{160} = -1.243$, $p = 0.216$, note: a non-parametric *U*-test agrees with the *t*-test). Explaining the variance for the solitary novices does not seem very difficult, since here the correlation between the report and accuracy is important ($r_{86} = 0.262$, $p = 0.014$; figure 6*a*). However, when we focus on paired novices we find interesting trends. First, when we try to explain the variation of the reported knowledge with respect to accuracy, we find that the correlation is very low and not statistically significant ($r_{76} = 0.07$, $p = 0.544$; figure 6*b*). Additionally, something interesting appears when we consider the relationship between the report with respect to the rate of response to the queries by the expert partner. There is an inverse correlation between the two variables ($r_{65} = -0.267$, $p = 0.029$; figure 6*c*). This means that as the expert responds to fewer of the novice's queries, the report of understanding by the novice increases. This is probably because an uncooperative partner leaves no choice but to learn the correct classifications by oneself, which is reflected in a higher report of understanding at the end of the experiment. Finally, the variance of the report is higher for higher uses of DLL (i.e. percentage of novice dogs queried by the player), as can be seen in figure 6*d* ($F_{20,56} = 0.422$, $p = 0.018$). Observe that extreme values of the report are unlikely to appear for those players who almost never queried their partner about their novice dogs, but extreme values are likely to occur for players who queried their partner.

5. Discussion

We started out by characterizing the relationship between novice and expert as it is used in DLL, and proposed that it can be explained by the relationship between user and enabler of a social, cognitive artefact. The novice is a user of a word, who does not need to understand the concept associated with it. The word is there for the novice to use it; it is part of their social environment. However, for this word to be used (and not just uttered) by the novice, there must be an expert who enables such a use in the pooled ground. Our experiment allowed us to observe the characteristics of novice/user and expert/enabler in a simple, but sufficiently interesting scenario. It succeeded in two respects. It created the conditions for the emergence of DLL involving a freshly minted dog breed expert, and it captured several aspects of the novice's use of words that characterize it as a cognitive offloading strategy according to Risko's & Gilbert's model [16]. Our experiment showed that novices are selective users, who assess the benefits of the extended strategy versus the internal strategy of learning for themselves, and that the amount of word-use is highly reciprocal within a dyad. In turn, experts are enablers who reliably support this word-as-artefact (almost always responding quite accurately to questions) and who receive the 'benefit' of increasing their own perceived understanding.

First, we succeeded in creating experts through training. Training caused participants to become experts at a particular kind of dog breed (terrier versus hound). Discriminating breeds within each kind of dog was difficult (e.g. Cairn terrier versus Norwich terrier), but gradually improved over the course of 25 training rounds. It is interesting to note that, despite their achieving comparable expertise with solo participants by the end of the training, paired participants performed slightly better when we consider all training rounds. The only difference between paired and solo conditions during training rounds was that paired participants knew that they were already paired with another participant, although they could not interact with them yet. Possessing this information seems to have increased their learning rate during training. Additional experiments are required to explore whether players are more diligent in their category learning efforts because they anticipate that their partner may later depend on them, but this result is consistent with findings that participants learn especially well when they need to know the information in order to teach others [35]. Participants' expertise carried over to a testing phase, leading to improved performance on the trained expert dog breeds compared with intermingled novel, untrained breeds (novice dogs). Furthermore, the differential expertise attained by members within a dyad was used by the members themselves. The evidence for this was that participants queried their partners about their unfamiliar dog breeds about 40% of the time (more on this below). Participants were even more likely to actively engage in their role as experts, responding to their partner's queries about 90% of the time. Thus, we were able to create cognitive specializations among the players in dyads, and the players themselves recognized their acquired knowledge differences, which they used to improve their performance.

We observed that the extended strategy proved its worth because it was more accurate for participants to label a novice dog after querying the expert than trying to label it on the basis of their own untrained ability. Moreover, the high correlation between the rates of querying within a dyad indicates that group-specific norms emerge over the course of interactions. If a participant is often asked about dogs by their partner, then they find it more natural to ask them about dogs. This norm for querying leads to large differences across the dyads in their members reliance on each other. For groups with self-reinforcing queries, a kind of 'social capital' emerges in which people learn to depend on each other, both giving and receiving aid.

The question arises as to why paired participants did not always opt for this strategy. Overall, it was used only 40% of the time. There are a number of factors that could account for this, which could be manipulated in future experiments. To begin with, novice players learned over rounds by observing their expert's responses, which clearly lowers novice players' query rates over time. In addition to this, when a player made few queries themselves, that leads to their partner also making few queries. Also, it is likely that a more difficult classification task with a longer training phase needed to achieve good categorization accuracy could raise query rates. Finally, there are factors involved in the cost of the extended strategy. Players needed to interpret the expert's answer. A player might ask 'Is this dog a 'Cairn terrier'?' If the answer is 'yes', the classification is obvious: label the dog as a Cairn terrier. If the answer is 'no', the correct classification is always the other breed in the pair. This might not be so obvious to all players. Moreover, using the extended strategy required extra time and there is also the social cost of showing lack of knowledge and depending on others. It is interesting to note that these costs could be manipulated, for instance, by implementing a treatment in which players always see their partner's classifications without querying them (we are indebted to an anonymous referee for pointing out this manipulation).

We also observed that when either the internal strategy or the extended strategy were less accurate up to that point, participants tended to opt for the other strategy with a higher probability on the next round. Our experiment cannot rule out spurious variables causing this correlation, so a better control is required to confirm whether participants' round-to-round strategy-selection phase is influenced by an evaluation of the costs and benefits of their two options.

As for the influence of DLL on people's meta-cognition, we found that people's self-rated understanding was strongly influenced by their social context—whether they were labelling dogs on their own or with possible help from a peer who was given different training. This influence was different for participants in their role as novices as compared to their role as experts. In their role as novices, we saw that the *average* self-rated understanding of participants is independent of their social context. However, the *variability* of ratings is higher for novices that were allowed and actively engaged with the DLL strategy. That is, paired participants who used the extended strategy gave relatively extreme low and high self-ratings of their understanding of novice dogs. A plausible explanation of these extreme values is two-fold and it warrants closer scrutiny in further experiments. On the one hand, there is evidence that being connected to a knowledgeable other increases one's own sense of knowing [7], so a stronger connection to a trustworthy expert partner could increase the novice's sense of understanding when the novice gave themselves credit for knowing what their partner knows. This seems to be the case only for some of our participants in their role as novices. On the other hand, participants might very well be aware that they do not know the difference between, say, a Cairn terrier and a Norwich terrier. When they resort to asking their expert partner, they come to realize that they are uncertain themselves. That is, a novice feeling forced to ask their expert partner about a dog breed might be equivalent to Rozenblit & Keil's [5] demonstrating to participants that they do not know how a mechanism works by asking them about the mechanism. Only after people have tried to answer a question themselves do they become aware that they are not as knowledgeable as they originally thought. Moreover, when they are given information by their paired expert, then social comparison to that expert makes it clear to the novice that their knowledge was relatively poor. Thus, the high variability of paired novices' rated self-understandings may be owing to their implicitly adopting one of two competing conceptualizations of knowledge—the 'I know what my neighbours know if I am reliably connected to them' versus 'I realize what I don't know when I need to ask others who are clearly more knowledgeable than me'.

Last but not least, we also found that self-rated understanding is higher for experts in dyads than solitary experts. This is apparently not owing to experts believing their partner knows more about the expert dogs than do the experts. Experts very rarely made queries to their partners about breeds for which they were experts. Instead, a plausible hypothesis is that when the novice partner queries the expert for advice, then the expert's confidence in their own knowledge increases. This is consistent with recent empirical results showing that the act of giving advice to others increases one's confidence [36]. In fact, in one study, giving advice raised self-reported confidence more than receiving advice did [37]. This latter result is consistent with our finding that experts giving advice increased their self-reported understanding more than did novices receiving advice from experts. The increase in experts' self-reported understanding when they are put in dyads was not, alas, preceded by actual improvements in their accuracy. Experts in the dyad and individual conditions did not differ in their actual accuracy. Accordingly, the experts' metacognitive judgements (i.e. a judgement about their own cognition) were biased by their social context. This biasing effect of social context is also suggested by the result that for single experts, their self-perceived understanding was significantly related to their actual accuracy, but that was no longer true when the experts were in dyads. This weakened connection owing to dyadic context was also found for our novices. Thus, our results speak to other factors besides accuracy looming larger when people judge their knowledge in dyads compared to when they are isolated. A good candidate for one such factor is that one emphasizes perceived knowledge *relative to others* in social contexts.

Ethics. Before starting the experimental session, each participant signed an informed consent document, which was previously approved by the Ethics Committee (social science room) of the Universidad del Rosario.

Data accessibility. A thorough presentation of the 'Classifying dog breeds' protocol, described in the 'Materials and methods' section, can be found in the protocols repository: https://www.protocols.io/view/classifying-dog-breeds-bvm6n49e. The task was implemented in the nodeGame platform (https://nodegame.org/) and was run in a university computer laboratory. Each participant was seated at a sound- and sight-isolated personal computer running a version of the game. The code implementing the game is freely available. The paired condition can be downloaded from https://github.com/Slendercoder/DLL and the solo condition can be downloaded from https://github.com/Slendercoder/DLL_single. In both cases, the interface is in Spanish. The two datasets used in the analysis are stored in the Github repository: https://github.com/EAndrade-Lotero/SPUoDLL. A thorough explanation of these datasets can be found in https://github.com/EAndrade-Lotero/SPUoDLL/blob/master/README.rtf. The R code to all the statistical analyses is presented in https://github.com/EAndrade-Lotero/SPUoDLL/blob/master/Tests.R. The python code to create the figures is presented in the Jupyter notebook in https://github.com/EAndrade-Lotero/SPUoDLL/blob/master/Figures.ipynb.

Data are also provided in the electronic supplementary material [38].

Authors' contributions. E.J.A.-L.: conceptualization, data curation, formal analysis, funding acquisition, investigation, methodology, project administration, resources, software, supervision, validation, visualization, writing—original draft, writing—review and editing; J.M.O.-D.: formal analysis, writing—original draft, writing—review and editing; J.A.V.-G.: investigation, software, writing—original draft, writing—review and editing; R.L.G.: conceptualization, methodology, resources, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. E.J.A.-L. was partially supported by the following research awards from Universidad del Rosario: 'Becas para Estancias de Docencia e Investigación, 2018' and 'Fondo de Capital Semilla, 2018'. J.M.O.-D. and J.A.V.-G. were partially supported by 'Fondo de Capital Semilla, 2018'.

Acknowledgements. The funding source had no involvement in the research. The authors are grateful to Steven Sloman and Luis Andrade for useful discussions.

Endnotes

[1]We are indebted to an anonymous referee for pointing this out to us.
[2]We should report that, owing to an unexpected data corruption, we lost two data points from one participant.
[3]Observe that, for this analysis, we eliminated one data point corresponding to one expert dog breed of one participant. The data point is clearly an outlier, because the participant reported a score of 1 after performing poorly on the classification task of one of their expert dog breeds. The statistics without removing the outlier are $r_{\text{report vs.accuracy}}(76)=0.55$, $p<0.001$.

References

1. Putnam H. 1975 The meaning of 'meaning'. In *Minnesota studies in the philosophy of science*, vol. 7 (ed. K Gunderson), pp. 131–193. Minneapolis, MN: University of Minnesota Press.
2. Ware R. 1978 The division of linguistic labor and speaker competence. *Phil. Stud.* **34**, 37–61. (doi:10.1007/BF00364687)
3. Burge T. 1979 Individualism and the mental. *Midwest Stud. Phil.* **4**, 73–121. (doi:10.1111/j.1475-4975.1979.tb00374.x)
4. Glenberg AM, Epstein W. 1985 Calibration of comprehension. *J. Exp. Psychol.* **11**, 702–718. (doi:10.1037/0278-7393.11.1-4.702)
5. Rozenblit L, Keil FC. 2002 The misunderstood limits of folk science: an illusion of explanatory depth. *Cogn. Sci.* **26**, 521–562. (doi:10.1207/s15516709cog2605_1)
6. Sloman S, Fernbach P. 2017 *The knowledge illusion: why we never think alone*. New York, NY: Riverhead.
7. Sloman S, Rabb N. 2016 Your understanding is my understanding: evidence for a community of knowledge. *Psychol. Sci.* **27**, 1451–1460. (doi:10.1177/0956797616662271)
8. Hemmatian B, Sloman S. 2018 Community appeal: explanation without information. *J. Exp. Psychol.* **147**, 1677–1712. (doi:10.1037/xge0000478)
9. Fisher M, Goddu M, Keil F. 2015 Searching for explanations: how the internet inflates estimates of internal knowledge. *J. Exp. Psychol.* **144**, 674–687. (doi:10.1037/xge0000070)
10. Sparrow B, Liu J, Wegner DM. 2011 Google effects on memory: cognitive consequences of having information at our fingertips. *Science* **333**, 776–778. (doi:10.1126/science.1207745)
11. Hollingshead AB. 1998 Communication, learning, and retrieval in transactive memory systems. *J. Exp. Soc. Psychol.* **34**, 423–442. (doi:10.1006/jesp.1998.1358)
12. Briscoe RE. 2006 Individualism, externalism and idiolectical meaning. *Synthese* **152**, 95–128. (doi:10.1007/s11229-005-1333-2)
13. Kallestrup J. 2012 *Semantic externalism*. Routledge.
14. Norman DA. 1991 Cognitive artifacts. *Design. Interact.* **1**, 17–38.
15. Clark A, Chalmers D. 1998 The extended mind. *Analysis* **58**, 7–19. (doi:10.1093/analys/58.1.7)
16. Risko EF, Gilbert SJ. 2016 Cognitive offloading. *Trends Cogn. Sci.* **20**, 676–688. (doi:10.1016/j.tics.2016.07.002)
17. Gray WD, Fu WT. 2004 Soft constraints in interactive behavior: the case of ignoring perfect knowledge in-the-world for imperfect knowledge in-the-head.

Cogn. Sci. **28**, 359–382. (doi:10.1016/j.cogsci.2003.12.001)

18. Theiner G, Allen C, Goldstone RL. 2010 Recognizing group cognition. *Cogn. Syst. Res.* **11**, 378–395. (doi:10.1016/j.cogsys.2010.07.002)
19. Goldstone RL, Gureckis TM. 2009 Collective behavior. *Topics Cogn. Sci.* **1**, 412–438. (doi:10.1111/j.1756-8765.2009.01038.x)
20. Haukioja J, Nyquist M, Jylkk¨a J. 2020 Reports from twin earth: both deep structure and appearance determine the reference of natural kind terms. *Mind Lang.* **36**, 377–403. (doi:10.1111/mila.12278)
21. Wrathall MA. 2010 *Heidegger and unconcealment: truth, language, and history*. Cambridge, UK: Cambridge University Press.
22. Clark HH. 1996 *Using language. 'using' linguistic books*. Cambridge, UK: Cambridge University Press.
23. Raczaszek-Leonardi J, Debska A, Sochanowicz A. 2014 Pooling the ground: understanding and coordination in collective sense making. *Front. Psychol.* **5**, 1233. (doi:10.3389/fpsyg.2014.01233)
24. Wegner DM. 1987 Transactive memory: a contemporary analysis of the group mind. In *Theories of group behavior* (eds B Mullen, GR Goethals). New York, NY: Taylor Francis.
25. Clark A. 2006 Material symbols. *Phil. Psychol.* **19**, 291–307. (doi:10.1080/09515080600689872)
26. Fowler CA. 1986 An event approach to the study of speech perception from a direct–realist perspective. *J. Phonet.* **14**, 3–28. (doi:10.1016/S0095-4470(19)30607-2)
27. Borghi A, Cimatti F. 2009 Words as tools and the problem of abstract words meanings. In *Proceedings of the 31st annual conference of the cognitive science society* (eds N Taatgen, H van Rijn), pp. 2304–2309. Amsterdam, The Netherlands: Cognitive Science Society.
28. Di Paolo EA, Cuffari EC, Jaegher HD. 2018 *Linguistic bodies: the continuity between life and language*. Cambridge, MA: The MIT Press.
29. Recanati F. 2000 Deferential concepts: a response to Woodfield. *Mind Lang.* **15**, 452–464. (doi:10.1111/1468-0017.00144)
30. Moreland RL. 1999 Transactive memory: learning who knows what in work groups and organizations. In *Shared cognition in organizations: the management of knowledge. LEA's organization and management series* (eds JM Levine, DM Messick, LL Thompson), pp. 3–31. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.
31. Brandon DP, Hollingshead AB. 2004 Transactive memory systems in organizations: matching tasks, expertise, and people. *Org. Sci.* **15**, 633–644. (doi:10.1287/orsc.1040.0069)
32. Keil F. 2019 How do partial understandings work? In *Varieties of understanding: new perspectives from philosophy, psychology, and theology*. Oxford, UK: Oxford University Press.
33. Balietti. 2017. NodeGame: real-time, synchronous, online experiments in the browser. *Behavior Research Methods* **49**, 1696–1715
34. Siegler RS, Lemaire P. 1997 Older and younger adults' strategy choices in multiplication: testing predictions of ascm using the choice/no-choice method. *J. Exp. Psychol.* **126**, 71–92. (doi:10.1037/0096-3445.126.1.71)
35. Biswas G *et al*. 2005 Learning by teaching: a new agent paradigm for educational software. *Appl. Artif. Intelligence* **19**, 363–392. (doi:10.1080/08839510590910200)
36. Eskreis-Winkler L, Milkman KL, Gromet DM, Duckworth AL. 2019 A large-scale field experiment shows giving advice improves academic outcomes for the advisor. *Proc. Natl Acad. Sci. USA* **116**, 14 808–14 810. (doi:10.1073/pnas.1908779116)
37. Eskreis-Winkler L, Fishbach A, Duckworth AL. 2018 Dear abby: should I give advice or receive it? *Psychol. Sci.* **29**, 1797–1806. (doi:10.1177/0956797618795472)
38. Andrade-Lotero EJ, Ortiz-Duque JM, Velasco-García JA, Goldstone RL. 2022 The division of linguistic labour for offloading conceptual understanding. Figshare. (doi:10.6084/m9.figshare.c.6251589)