# Neurocomputational Models Capture the Effect of Learned Labels on Infants' Object and Category Representations

Arthur Capelier-Mourguy, Katherine E. Twomey, and Gert Westermann

***Abstract*—The effect of labels on nonlinguistic representations is the focus of substantial theoretical debate in the developmental literature. A recent empirical study demonstrated that ten-month-old infants respond differently to objects for which they know a label relative to unlabeled objects. One account of these results is that infants' label representations are incorporated into their object representations, such that when the object is seen without its label, a novelty response is elicited. These data are compatible with two recent theories of integrated label-object representations, one of which assumes labels are features of object representations, and one which assumes labels are represented separately, but become closely associated across learning. Here, we implement both of these accounts in an auto-encoder neurocomputational model. Simulation data support an account in which labels are features of objects, with the same representational status as the objects' visual and haptic characteristics. Then, we use our model to make predictions about the effect of labels on infants' broader category representations. Overall, we show that the generally accepted link between internal representations and looking times may be more complex than previously thought.**

***Index Terms*—Cognitive development, connectionist model, label status, language development, representational development.**

## I. Introduction

THE NATURE of the relationship between labels and nonlinguistic representations has been the focus of recent theoretical debate in the developmental literature. On the labels-as-symbols account [1], [2], labels are symbolic, conceptual markers acting as privileged, top-down indicators of category membership, and label representations are qualitatively different to object representations. In contrast, the labels-as-features (LaFs) view assumes that labels have no special status; rather, they contribute to object representations in the same way as other features, such as shape and color. More recently, Westermann and Mareschal (W&M) [3] suggested a compound-representations (CRs) account in which labels are encoded in the same representational space as objects and drive learning over time, but do not function at the same level as other perceptual features. Rather, they become closely integrated with object representations over learning and result in mental representations for objects that reflect both perceptual similarity and whether two objects share the same label or have different labels. This approach therefore takes a middle ground between the labels-as-symbols and the LaFs views in that labels do not act at the same level as other object features (acknowledging that language is special as in labels-as-symbols), but that an integrated object representation is formed through the association between perceptual object features and labels (as in LaFs). However, despite substantial empirical work (e.g., [3]–[10]) and a handful of computational investigations (e.g., [3], [11], and [12]), there is no current consensus as to the status of labels in object representations, and the debate goes on.

A variety of studies have demonstrated that language does affect object encoding and representations early in development. When and how in development this relationship emerges is less clear. For example, labels can guide online category formation in infants and young children [13]–[15], and previously learned category representations affect infants' online visual exploration in the laboratory [16], [17], but until recently the link between learned labels and category representations had not been directly tested. Gliga *et al.* [5] recently explored electroencephalogram (EEG) neural responses to stimuli in 12-month-old infants presented with a previously labeled object, a previously unlabeled object, and a new object. They found significantly stronger gamma-band activity only in response to the previously labeled object, and this, in line with previous EEG work, was interpreted as a marker of stronger encoding of this object. Twomey and Westermann [8] extended this paper by training 10-month-old infants with a label-object mapping over the course of one week. Specifically, parents trained infants with two objects during 3-min play sessions, once a day for seven days, using a label for one of the objects, but not for the other. After the training phase, infants participated in a looking time task in which they were shown images of each object in silence. Testing the hypothesis that

Manuscript received November 14, 2017; revised September 13, 2018; accepted November 5, 2018. Date of publication November 29, 2018; date of current version June 10, 2020. This work was supported in part by the Doctoral Scholarship Grant through the Leverhulme Trust to ACM, in part by the ESRC International Centre for Language and Communicative Development under Grant ES/L008955/1, in part by ESRC Future Research Leaders Fellowship to KT under Grant ES/N01703X/1, and in part by the British Academy/Leverhulme Trust Senior Research Fellowship to GW under Grant SF150163. *(Corresponding author: Arthur Capelier-Mourguy.)*

A. Capelier-Mourguy and G. Westermann are with the Department of Psychology, Lancaster University, Lancaster LA1 4YF, U.K. (e-mail: a.capelier-mourguy@lancaster.ac.uk; g.westermann@lancaster.ac.uk).

K. E. Twomey is with the School of Health Sciences, University of Manchester, Manchester M13 9NT, U.K. (e-mail: katherine.twomey@manchester.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCDS.2018.2882920

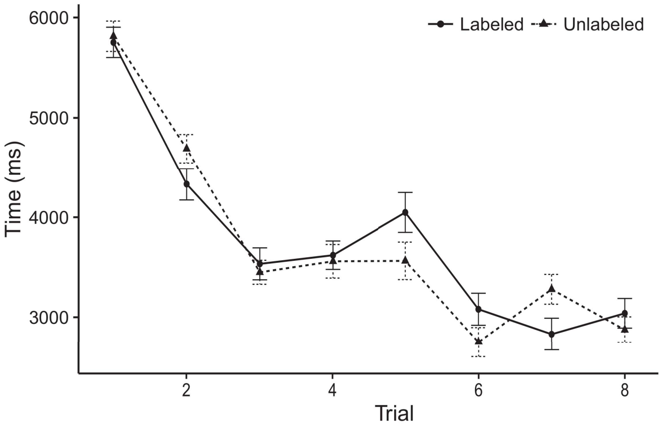

Fig. 1. Looking time results from [8]. Error bars represent 95% confidence intervals.

(previously learned) labels would affect infants' object representations, the authors predicted that infants should exhibit different looking times to the labeled and unlabeled objects. Their predictions were upheld: results showed a main effect of labeling, such that infants looked longer at the previously labeled than the unlabeled object (see Fig. 1 for the original data).

These data shed light on the debate on the status of labels. Specifically, they support both the LaFs and the CRs theories. On the LaFs account, if a label is an integral part of an object's representation, when the label is absent there will be a mismatch between that representation and what the infant sees in-the-moment (equally, a similar response would be expected when another of the object's features, for example color, differed from the learned representation). Since infants are known to engage preferentially with novel stimuli [18], [19], this mismatch will elicit a novelty response, indexed by increased looking times to the previously labeled object. On the CRs view, seeing the previously labeled object would activate the label representation [20]. This active label representation would, in turn, lead to a priming-like increase in looking time toward the previously labeled object [21]–[23].

Importantly, while the behavioral data presented in [8] support either of these views, they cannot differentiate between the two. Computational models, on the other hand, allow researchers to explicitly test the mechanisms specified by these theories against empirical data. Specifically, simple computational models, by stripping back mechanisms to a minimum, allow us to precisely understand these mechanisms and discover which ones are relevant and which ones are not (for similar arguments, see [24] and [25]). Thus, here we implemented both accounts in simple computational models to explore which of the LaFs and CRs accounts best explains Twomey and Westermann's [8] looking time data.

## II. Experiment 1

### A. Model Architecture

We used a dual-memory three-layer auto-encoder model inspired by W&M [3] to implement both the LaFs and the CRs theories. Such neurocomputational models have successfully captured looking time data from infant categorization tasks [3], [26]–[30]. Auto-encoders reproduce input patterns on their output layer by comparing input and output activation after presentation of training stimuli, then using this error to adjust the weights between units using back-propagation [31].

Our model consisted of two auto-encoders coupled by, and interacting through, their hidden units. These two subsystems represented, on an abstract level, a short-term (STM) and a long-term (LTM) memory component. This model has previously been used to simulate the impact of infants' background category knowledge acquired in everyday life (represented in LTM memory) on lab-based looking time experiments involving in-the-moment knowledge acquired in familiarization-novelty-preference studies (represented in STM) [3]. It was therefore well suited to simulate the effects of infants' learning about objects and labels at home on their subsequent looking behavior in the lab as in [8].

The two auto-encoders had different learning rates: the LTM component used a learning rate of 0.001 so that it encoded information relatively slowly; the STM used a learning rate of 0.1 and encoded information relatively quickly. For the interaction between the two networks' hidden units, both hidden layers were updated in parallel, receiving activation from their input layer and the other network's hidden layer until both hidden layers had converged to a stable representational state, with the lateral interaction resulting in no further update in their activation. The weights from the STM to LTM were treated as part of the LTM network and updated with a learning rate of 0.001; similarly, the weights from the LTM to the STM were treated as part of the STM network and updated with a learning rate of 0.1. Thus, the influence of the other memory on each network was updated at the same rate as the rest of the network. Both networks received identical input. The details for all the model parameters and the full code are available online.[1]

*1) Labels-as-Features Model:* Fig. 2(a) depicts the LaF model. To represent the label as a feature that was equivalent to all other features, we included it both at the input and the output level for both components. Thus, the label had exactly the same status as all other features in the model's representation.

*2) Compound-Representations Model:* Fig. 2(b) depicts the CR model. Here, labels are represented only on the output side of the LTM network. Thus, in effect, the model learns to associate the perceptual object description with the label. This approach reflects the empirical finding that presenting an object to infants activates their (learned, LTM) representation of the label for that object [20].

*3) Stimuli:* Our stimuli were encoded as sets of abstract binary features that were designed to reflect the visual, haptic and label characteristics of the 3-D object stimuli used in Twomey and Westermann [8]. Thus, our encoding can be interpreted as a list of dummy variables that could generalize to alternative stimuli, coding for the presence/absence of one particular dimension of the stimuli (e.g., "is made of

[1] https://github.com/respatte/LabelTime

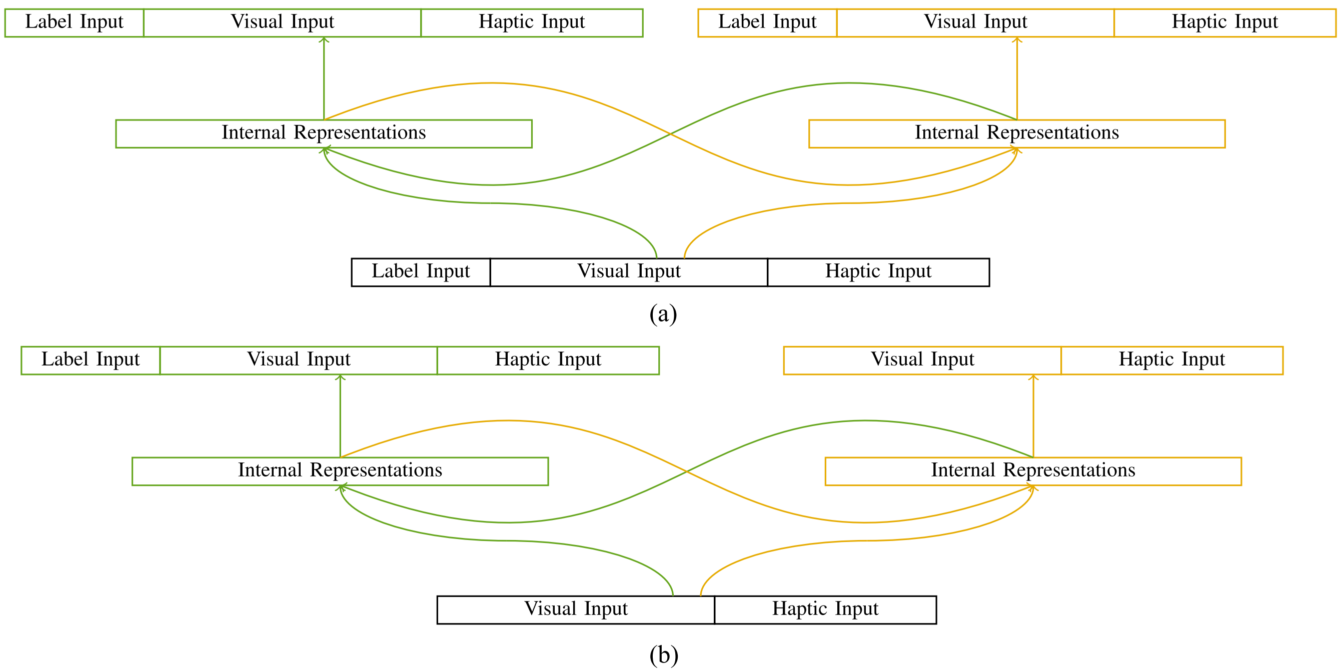

Fig. 2. Structure of the dual-memory network models: the LTM memory is in green (left), and the STM memory in yellow (right). Layer width corresponds to number of units: 5 label, 10 visual, 8 haptic, and 15 hidden units. (a) LaFs model. (b) CRs model.

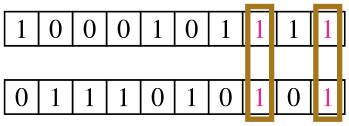

Fig. 3. Encoding of stimuli, with overlapping units highlighted.

wood," "is red," would be plausible dimensions for the stimuli considered here).

*a) Visual input:* Twomey and Westermann's [8] empirical study stimuli were two small wooden toys: a castanet, and two wooden balls joined with a string. One toy was painted red and the other blue, with color counterbalanced across children. Thus, the stimuli were visually dissimilar, but both consisted of two wooden components connected with string/elastic. To reflect the partial overlap in visual appearance of these objects, we encoded the visual component of our stimuli as patterns of activation over ten units; each object had the same number of active units (6), with two out of the ten units active for both objects to represent commonalities between stimuli (see Fig. 3).

*b) Haptic input:* As well as visual experience, infants in [8] received haptic input when handling or mouthing the stimuli. We reasoned that the degree of overlap in this input would vary between infants. Because both objects were wooden and presented simultaneously, infants would have experienced some overlap in haptic experience with the objects. On the other hand, because the objects had different affordances, this overlap would never have been total. Thus, we encoded haptic input over eight units, with overlap varying randomly between two and six units between simulations. Haptic stimuli were presented to the model simultaneously with the visual stimuli and encoded in an identical fashion.

*c) Label input:* Label input consisted of five binary units, activated (set to 1) for the labeled object only. For the unlabeled object, the units were simply set to 0.

## *B. Procedure*

In line with the experimental study in [8], our procedure consisted of two phases. First, to simulate the 3-D object play sessions at home, we trained the models with both objects, one with a label and one without a label (*background training*). Then, we simulated the second, lab-based part of the study by familiarizing the models with both objects without the labels to simulate the silent familiarization phase of the empirical study. Specifically, we ran each architecture in a familiarization phase in which the label units were inactive for both stimuli: the label inputs for the LaF architecture were set to zero, and the label outputs were ignored for both architectures (therefore not contributing to network error nor impacting on further weight updates).

To collect an amount of data consistent with infant studies, we ran a total of 40 model subjects for each architecture.

*1) Play Sessions:* To reflect the likely differences in playing time across children, the total number of iterations for which the model received each stimulus during background training was selected randomly from a normal distribution of mean 2000 and standard deviation 200. Stimuli were presented individually in alternating fashion. Although this does not precisely reflect the rich, combined play with both objects for different times experienced by infants, alternating the stimuli allows the model to learn more efficiently from a purely computational point of view, and should not influence results, as different training orders for the same stimuli asymptotically converge to the same solution.

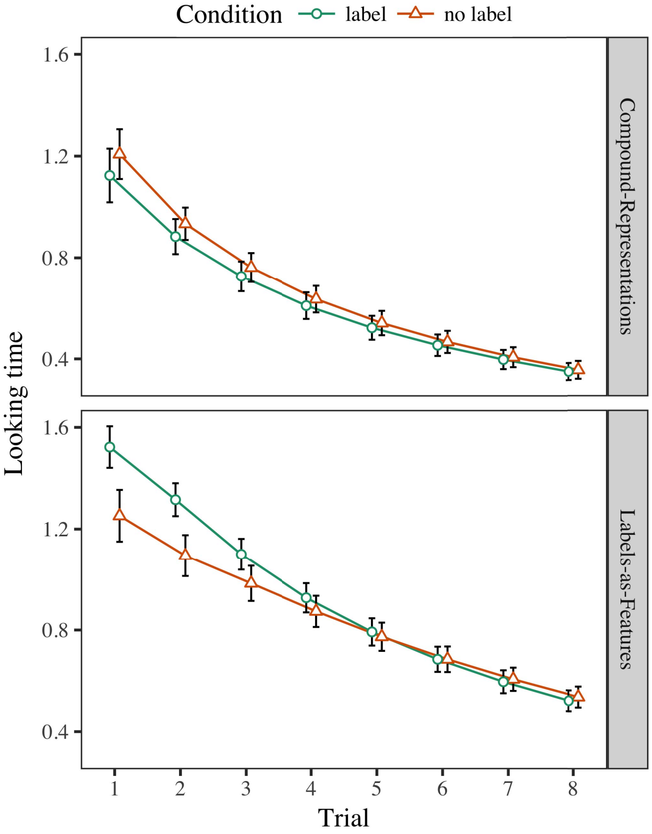

Fig. 4. Looking time results for Experiment 1 simulations. Error bars represent 95% confidence intervals.

*2) Familiarization Training:* Before familiarization training, we added noise to the STM's hidden-to-output weights (by adding a value in the range ±[0.1, 0.3] to the existing weight values) to simulate the likely memory decay from infants' final play session, which had taken place the previous day. Then, the label input units were set to zero, and the output units ignored, not taking them into account when computing network error and back-propagation. Haptic input and output units were also set to zero, to reflect the absence of haptic experiences in the lab experiment.

Familiarization then proceeded as follows: in line with Twomey and Westermann [8], stimuli were presented in alternation for eight trials each. The familiarization phase therefore consisted of 16 trials in total. The initial stimulus was counterbalanced across simulations. In line with previous similar models, we used the network's error on the output of the STM component as an index of infants' looking times [3], [26], [28]–[30].

## C. Results

Results from the familiarization phase for both simulations are depicted in Fig. 4. We submitted STM error (looking time) to an omnibus linear mixed-effects model using the R (3.4.4) package `lme4` (1.1−17) [32] (full code available on GitHub). The model with maximal random-effects structure that converged [33] included fixed effects for trial (1–8), theory (CRs, LaFs), and the trial-by-condition (label, no label), theory-by-condition, trial-by-theory, and trial-by-theory-by-condition interactions; and by-subject random intercepts and slopes for trial and condition. All fixed effects in this final analysis significantly improved model fit according to a likelihood ratio test; a main effect of condition was dropped because it did not contribute to model fit. Full details of the fitted fixed effect parameters are provided in Table I.

To understand the interactions, we submitted looking time for each model to separate mixed effects analyses, constructed in an identical fashion to the omnibus analysis. Full details of the theory-specific analyses' parameters are also given in Table I. Overall, the CR model's looking time decreased rapidly across trials. There was a small but significant improvement in model fit; an interaction between trial and condition, with a slightly slower decrease in looking time in the label condition, but no main effect of condition. Thus, the CR model did not capture the pattern of results in the empirical study, in which infants looked longer at the previously labeled object. The LaF model's looking times also decreased across trials, and this model showed a strong effect of label, with longer looking times toward the previously labeled object. The trial-by-condition interaction also improved the model, with looking time toward the previously labeled object decreasing faster to fall to a comparable level to the looking time to the previously unlabeled stimulus. Although this interaction was not found in the empirical data analysis, it is not uncommon for models to deviate from the precise patterns of empirical data while capturing the overall pattern of interest. This is particularly the case with the additional noisiness found in infant data; the empirical data analysis might have failed to detect this interaction effect between trial and condition, due to the noisiness and smaller sample size of infant studies naturally decreasing statistical power. In the end, the LaF model captures Twomey and Westermann's [8] main empirical results of interest: when all else is held equal, teaching the LaF model a label for one object but not another leads to longer looking times toward the previously labeled object in a subsequent, silent familiarization phase.

## D. Discussion

In Experiment 1, we tested two possibilities for the relationship between labels and object representations using a neurocomputational model to capture recent empirical data [8]. The target data showed that previously learned labels affect 10-month-old infants' looking times in a silent familiarization phase, suggesting that knowing a label for an object directly affects its representation, even when that object is presented in silence. As noted by Twomey and Westermann [8], both the CRs and LaFs accounts predict some effect of labels on object representations, and both theories could explain their empirical data. To disentangle these two accounts, we implemented both theories in simple dual-memory auto-encoder models inspired by [3]. In our CR model, we instantiated labels on the output layer only. This model learned to associate labels with inputs over time such that the presence of visual/haptic input for an object would consistently activate the label, but nonetheless, label information was separate from visual and haptic object

TABLE I
ESTIMATED PARAMETERS FOR EXPERIMENT 1 LOOKING TIMES: FIXED EFFECTS FOR GLOBAL, CR, AND LaF LMER MODELS

| | Global Model | | | LaF Model | | | CR Model | | |
|---|---|---|---|---|---|---|---|---|---|
| Parameter | Estimate | SE | *t*-value | Estimate | SE | *t*-value | Estimate | SE | *t*-value |
| Intercept | 1.025763 | 0.031206 | 32.871 | 1.431984 | 0.031198 | 45.90 | 1.038399 | 0.030776 | 33.740 |
| Trial | -0.107329 | 0.003630 | -29.564 | -0.142704 | 0.003471 | -41.11 | -0.108625 | 0.003743 | -29.018 |
| Condition (no label) | NA | NA | NA | -0.229149 | 0.021846 | -10.49 | NA | NA | NA |
| Trial × Condition | -0.003336 | 0.003193 | -1.045 | 0.042076 | 0.003629 | 11.59 | -0.003156 | 0.003412 | -0.925 |
| Theory (LaF) | 0.406220 | 0.045270 | 8.973 | | | | | | |
| Theory × Condition | -0.229149 | 0.023438 | -9.777 | | | | | | |
| Trial × Theory | -0.035375 | 0.005259 | -6.727 | | | | | | |
| Trial × Theory × Condition | 0.045412 | 0.005085 | 8.931 | | | | | | |

information [3]. In our LaF model, labels were represented on the input as well as on the output layers in exactly the same way as the visual and haptic components of object representations [6], [11]. Only the LaF model captured the longer looking to the previously labeled stimulus exhibited by the infants in Twomey and Westermann's [8] empirical study.

These results offer converging evidence that labels may have a low-level, featural status in infants' early representations. In line with recent computational work [3], [11] we chose to explore such low-level accounts using a simple associative model that could account for the nuances of recent empirical data [8]. Our LaF model offers a parsimonious account of Twomey and Westermann's [8] results, in which looking time differences emerge from a low-level novelty effect [6], [34], [35], without the need to specify qualitatively different, top-down representations [2], [36], [37]. Specifically, as argued in [8], and as implemented in the LaF model, over background training the label is learned as part of the object representation. Thus, when the object appears without the label there is a mismatch between representation and reality. This mismatch leads to an increase in network error for the previously labeled stimulus only, which has been interpreted in the literature as a model of longer looking times [3], [26], [28]–[30]. Further, these results delineate between the two possible explanations for infants' behavior in the empirical task; specifically, our results support accounts of early word learning in which labels are initially encoded as low-level, perceptual features, and integrated into object representations.

## III. EXPERIMENT 2

Overall, then, our LaF model offers a mechanism by which labels affect infants' representations of single objects. However, rather than one-to-one label-object mappings, infants typically learn labels for categories of objects; for example, a child might learn that their brown furry cuddly toy, the spotted animal in their picture book, and the hairy, barking animal at Grandma's are all referred to by the label "dog." A question that Twomey and Westermann's [8] empirical study and the current computational replication leave open, then, is whether the effect seen here would persist when considering richer categories rather than single objects. Thus, in Experiment 2 we extended our LaF model to category learning to make testable predictions for future empirical work. To this end, we trained our model with two object categories, one labeled and one unlabeled, before testing the model on a new exemplar from each category in the same way as in Experiment 1.

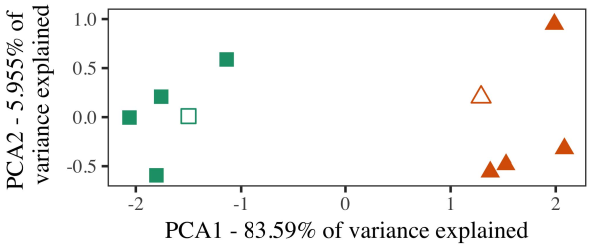

Fig. 5. Example of two categories generated for Experiment 2 [first two dimensions of a principal component analysis (PCA)]. Hollow shapes represent the prototypes, used during the familiarization (lab) phase, around which categories, where constructed, and filled shapes represent exemplars used during background training. We used PCA to reduce the dimensionality of the representational space in order to plot the 10-D exemplars in a 2-D space. The proportion of variance in the original representation explained by each of the plotted dimensions is specified on the axis labels.

As our implementation of the CR model did not replicate the empirical results in Experiment 1, we do not report it in Experiment 2 and instead focuse on the LaF model.

### A. Stimuli

In these simulations, stimuli consisted of two distinct categories with five exemplars each. Four of the five exemplars for each category were used for background training, keeping the remaining one as a novel within-category item for the simulated looking time phase.

To allow for convenient future empirical testing of our predictions (e.g., using pictures in a storybook read at home as in [16] and [38]), we removed the haptic units from the model. We constructed our categories around two exemplars with one overlapping unit (out of the ten visual units), and then randomly adding noise to this exemplar, adding to the prototype values taken from a uniform distribution between $-0.5$ and 0.5. Thus, we ensured that both categories formed distinct clusters in representational space, while making all exemplars within a category distinct from each other (Fig. 5).

TABLE II
ESTIMATED PARAMETERS FOR EXPERIMENT 2 LOOKING TIMES: FIXED EFFECTS FOR LaF LMER MODEL

| Parameter | Estimate | SE | $t$-value |
|---|---|---|---|
| Intercept | 1.347539 | 0.029841 | 45.16 |
| Trial | -0.152832 | 0.004466 | -34.22 |
| Condition (no label) | -0.350483 | 0.029221 | -11.99 |
| Trial × Condition | 0.065942 | 0.005235 | 12.60 |

TABLE III
PARAMETERS FOR EXPERIMENT 2 INTERNAL REPRESENTATIONS: FIXED EFFECTS FOR LaF LMER MODEL

| Parameter | Estimate | SE | $t$ |
|---|---|---|---|
| Intercept | 1.635e-01 | 4.467e-03 | 36.59 |
| Step | 2.054e-03 | 1.321e-04 | 15.55 |
| Condition (no label) | 1.815e-02 | 6.837e-03 | 2.65 |
| Step × Condition (no label) | -2.752e-04 | 8.009e-05 | -3.44 |

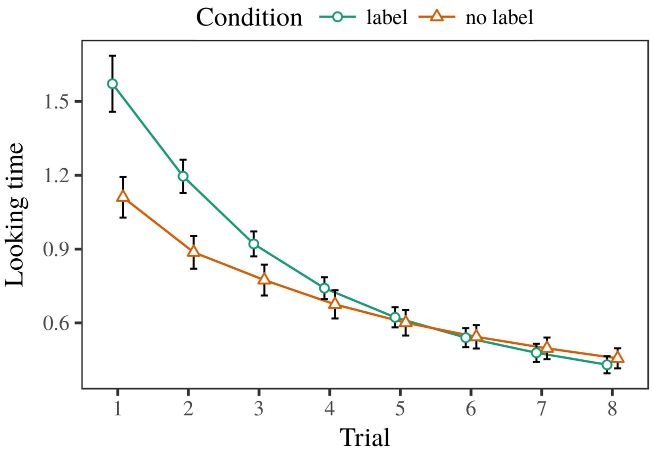

Fig. 6. Looking time results for the Experiment 2 simulations. Error bars represent 95% confidence intervals.

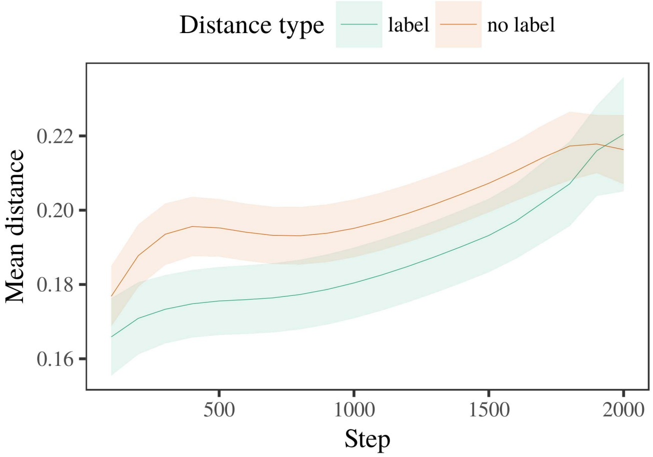

Fig. 7. Evolution of mean distance in internal representations of the LTM during background training for Experiment 2 simulations. Shaded areas represent 95% confidence intervals.

### B. Procedure

Similar to Experiment 1, we first trained the model with exemplars of each category, presented individually in alternating fashion, with timings drawn from a normal distribution of mean 2000 and standard deviation 200. Which category was labeled and which was unlabeled was counterbalanced across simulations.

We then presented the models with a familiarization phase in line with Experiment 1, in which the remaining exemplar for each category was presented without a label. As in Experiment 1, this phase consisted of 16 interleaved trials of up to 40 iterations (eight trials per category).

Again, to collect an amount of data consistent with infant studies, we ran a total of 40 model subjects.

### C. Results

*1) Looking Times:* Using the same procedure as in Experiment 1, we fitted an omnibus linear mixed-effects model to the STM network error (looking time) during familiarization. Results are shown in Fig. 6. The final model included main effects of trial (1–8), condition (label, no label), and a trial-by-condition interaction; the model also included by-subject random intercepts, and random slopes for trial and condition. All fixed effects in this final analysis significantly improved model fit according to a likelihood ratio test. Full detail of the fitted fixed effect parameters are given in Table II.

The model's looking time decreased across trials (main effect of trial), and, as in Experiment 1, the model showed longer looking times toward the previously labeled category (main effect of condition), and a faster decrease in looking time toward this category (trial-by-condition interaction). Thus, the LaF model predicted that when trained with labeled and unlabeled categories rather than individual objects, infants should again show a novelty response when viewing silently presented exemplars of the previously labeled category.

*2) Internal Representations in the Model:* A common way to look at a neural network's "understanding" of the inputs it has received is to examine the activation patterns in the hidden layer following encoding [3], [28], [29], [39]. We recorded these hidden representations for the training stimuli during background training every 100 iterations to investigate the development of memory representations. In our model, the LTM corresponds to representations in memory, whilst the STM corresponds to in-the-moment behaviors and perception; hence, we here examined the hidden units of the LTM network only. The mean within-category distances are displayed in Fig. 7.

We then submitted the mean distance between exemplars of each category to a mixed-effects model. We used the same model building principle as for the looking time results previously discussed.

The final model included main effects of step (iteration number when recording, divided by the recording interval of 100), a condition (label, no label), and a step-by-condition interaction; the model also included by-subject random intercepts and slopes for step and condition. All fixed effects in this final model significantly improved model fit according to a likelihood ratio test. The estimates for the fitted parameters of the fixed effects for this model are displayed in Table III.

The mixed-effects model indicated that the within-category distance increased slowly over time (main effect of step), with the distances between exemplars of the unlabeled category being larger than the distances between exemplars of the labeled category (main effect of condition), and with distances in the unlabeled category growing more slowly than in the labeled category, after a quicker start (step-by-condition interaction). Thus, the presence of a label associated with a category in our LaF model caused exemplars of this category to be represented more closely together, and to be differentiated more slowly than in the unlabeled category.

### D. Discussion

In Experiment 2 we extended our LaF model, which captured the empirical data from Twomey and Westermann [8] in Experiment 1, to a situation simulating infants' learning about object categories. The model predicted similar looking time patterns compared to those observed with single objects; that is, that infants should look longer, in silence, at exemplars that belong to a category for which they know a label.

Examination of the LaF network's hidden representations revealed that the labeled category was more compact than the unlabeled category, making labeled exemplars appear more similar to each other than unlabeled exemplars. The model nonetheless learned to discriminate different exemplars of a same category, making the distance between exemplars increase over time. The prediction that increased similarity between exemplars of a category may be seen together with longer looking times is intriguing. The reduced distances between exemplars of the labeled category in the model suggest that exemplars should be perceived as more similar to each other than those of the unlabeled category. If so, a new exemplar of this labeled category may be perceived as less novel than a new exemplar of the unlabeled category, leading to longer looking times to the latter. In contrast, however, the model predicts longer looking toward the previously labeled category exemplar, despite the reduced distance in internal representations. Our interpretation of this counter-intuitive result is that, despite the labeled category being more compact, the surprise effect of seeing an exemplar of this category without a label is still stronger than the facilitatory effect of a reduced distance in representational space.

Notably, W&M [3] used a CR model to address a related issue, specifically the effect of labeling on children's longer-term category learning. In their model they found reduced looking times to novel category exemplars for which a label was known compared to those with an unknown label. The predictions made by our LaF model in Experiment 2 therefore diverge from those of W&M: although the LaF model, like W&M, predicted that a category label reduces within-category distance in mental representations, it predicted higher instead of lower looking times for novel label-known category exemplars.

The reason for this difference likely relates to differences in stimuli and training between W&M's model and the current simulations. Specifically, W&M aimed more broadly to model the transition from prelinguistic to language-based processing in infant development. W&M provided their model with a relatively rich background knowledge of 208 exemplars drawn from 26 real-world basic level categories from four superordinate categories that were encoded through 18 meaningful features (geometry, object characteristics). In their simulation of label effects on object familiarization, the model first received background training on 202 objects from all 26 categories, including two rabbits. In the no-label condition no objects were labeled, and in the label condition encountered objects were labeled half the time (accounting for the fact that objects are not reliably labeled at every instance in which infants experience them). Then, the models were familiarized on six novel rabbits. Under these circumstances, W&M found that the label model familiarized faster to these stimuli than the no-label model.

In contrast, here we aimed to predict a controlled lab experiment, which involves less naturalistic situations and stimuli, with a single age group. Thus, our current model learned only two categories and saw a single test stimulus for each. During background training, objects from one of the categories were always labeled and objects from the other category were never labeled. Conversely, W&M's categories were perceptually very broad, and overlapped with other categories. The introduction of labels in this environment warped the representational space so that overlapping representations became separated in accordance with the labels. In the simulations reported here, however, the two categories were tight and nonoverlapping, so that the effects of labels were far more subtle. It is possible that the categories considered here are not sufficiently rich and variable for the label to become detached from each object's featural representation across learning. Indeed, our categories are made of a handful of exemplars each, with a limited number of features with low variability defining their belonging to a category, which contrasts with real-world categories defined by more, and more variable features.

Finally, it may be the case that the effect of the label on infants' category representations varies with age, perhaps developing from an LaFs representation to a CRs mechanism over time [34]. From this perspective, our model may simulate an earlier developmental stage (and mechanism), than W&M. It is indeed possible that infants first perceive labels as object features and form categories purely on a similarity basis, then slowly learn that labels are highly reliable predictors of category membership, even for less perceptually similar objects (e.g., "furniture," "animals," or "toys") [3], [34]. Empirical studies with infants are currently underway to address this issue.

## IV. General Discussion

The current simulations demonstrate that an LaFs account can explain empirical looking time data from ten-month-old infants pretrained with one labeled and one unlabeled 3-D object. Further, the LaF model predicted that when trained with labeled and unlabeled simple categories of objects, infants would exhibit longer looking times to a novel exemplar of the previously labeled category presented in silence. Testing this prediction experimentally is crucial; if confirmed, it would shed new light on categorization studies in infants, stressing that the same mechanisms (here compacting the representation of a category) might lead to very different, or even opposite behavioral results depending on the nature and structure of stimuli used.

It is important to note that other computational work has explored the effect of labeling on object representations in infants. Gliozzi *et al.* [11] used a self-organizing map (SOM; [40]) architecture to capture empirical data from a categorization task with ten-month-old children. Given that labels are represented as units in SOMs in the same way as visual features, this model might capture Twomey and Westermann's [8] results for similar reasons to the success of the LaF model. However, the two networks make very different assumptions about learning mechanisms, highlighting an important issue for both infancy research and computational work. Gliozzi *et al.* [11] model learns in an unsupervised way, strengthening associations between units in its SOM using "fire together, wire together" Hebbian learning. In contrast, our model learns by comparing what it "sees" to what it "knows" and updating its representations in proportion to any discrepancy. Thus, the current results are compatible with an error-based learning account to development, in which infants learn by tracking mismatches between representation and environment [41]. Whether unsupervised learning, error-based learning, or some combination of both drives early development is a profound theoretical issue outside the scope of this paper; for now, we highlight the importance of bearing in mind the link between the technical assumptions of a computational model and the implications for (developmental) theory.

In an era of increasing enthusiasm for complex, deep neural networks capable of learning to represent and label images, play (video) games, and many other tasks, it is important to show that simplicity in modeling can be a distinct strength. In particular, the simplicity of the architectures presented here produces a more transparent and interpretable mechanism than a network with many hidden layers. There would, however, be an obvious interest in the future in scaling up this paper to increasingly complex—and therefore realistic—learning environments, ultimately taking our model from the "friendly nursery" of our controlled setup and inputs into the real world. One important question is, for example, if an LaFs network would naturally evolve to give less and less importance to the input labels, effectively becoming a CRs model on the basis of experience with the world. This would support the hypothesis that infants learn through experience that labels are features with a higher predictive value for categorization, and therefore stop experiencing them as input features of object but learn to recall labels when presented with exemplar of known categories.

Finally, our simulations focused on two theories of the effect of labeling on category formation, but did not address the labels-as-symbols theory [1]. This theory assumes that labels are qualitatively different from other object features, and act in a symbolic way to directly shift the attentional focus toward diagnostic features that define a category. It is unclear how this theory could be implemented within the current framework, as our models do not have an explicit attentional component, and the very mechanism by which labels would highlight common features is not clearly defined in the theoretical account. Additional work is needed, on the one hand to define the precise mechanisms underlying this labels-as-symbols theory, and on the other hand to translate them into a computational model that can be tested and evaluated rigorously.

Taken together with Twomey and Westermann [8], however, this paper demonstrates how language can shape object representation and in this way, explain empirical results in infancy research.

## References

[1] S. R. Waxman and D. B. Markow, "Words as invitations to form categories: Evidence from 12- to 13-month-old infants," *Cogn. Psychol.*, vol. 29, no. 3, pp. 257–302, Dec. 1995.

[2] S. R. Waxman and S. A. Gelman, "Early word-learning entails reference, not merely associations," *Trends Cogn. Sci.*, vol. 13, no. 6, pp. 258–263, Jun. 2009.

[3] G. Westermann and D. Mareschal, "From perceptual to language-mediated categorization," *Philosoph. Trans. Roy. Soc. B Biol. Sci.*, vol. 369, no. 1634, 2014, Art. no. 20120391.

[4] S. A. Gelman and J. D. Coley, "Language and categorization: The acquisition of natural kind terms," in *Perspectives on Language and Thought: Interrelations in Development*. Cambridge, U.K.: Cambridge Univ. Press, 1991, pp. 146–196.

[5] T. Gliga, A. Volein, and G. Csibra, "Verbal labels modulate perceptual object processing in 1-year-old children," *J. Cogn. Neurosci.*, vol. 22, no. 12, pp. 2781–2789, 2010.

[6] V. M. Sloutsky and A. V. Fisher, "Induction and categorization in young children: A similarity-based model," *J. Exp. Psychol. Gen.*, vol. 133, no. 2, pp. 166–188, 2004.

[7] V. M. Sloutsky and A. V. Fisher, "Linguistic labels: Conceptual markers or object features?" *J. Exp. Child Psychol.*, vol. 111, no. 1, pp. 65–86, Jan. 2012.

[8] K. E. Twomey and G. Westermann, "Learned labels shape pre-speech infants' object representations," *Infancy*, vol. 23, no. 1, pp. 61–73, 2018.

[9] N. Althaus and D. Mareschal, "Labels direct infants' attention to commonalities during novel category learning," *PLoS ONE*, vol. 9, no. 7, 2014, Art. no. e99670.

[10] N. Althaus and K. Plunkett, "Categorization in infancy: Labeling induces a persisting focus on commonalities," *Develop. Sci.*, vol. 19, no. 5, pp. 1–11, Oct. 2015.

[11] V. Gliozzi, J. Mayor, J.-F. Hu, and K. Plunkett, "Labels as features (not names) for infant categorization: A neurocomputational approach," *Cogn. Sci.*, vol. 33, no. 4, pp. 709–738, Jun. 2009.

[12] M. Mirolli and D. Parisi, "Language as an aid to categorization: A neural network model of early language acquisition," in *Modeling Language, Cognition and Action*, 2005, pp. 97–106, doi: 10.1142/9789812701886_0009.

[13] N. Althaus and G. Westermann, "Labels constructively shape object categories in 10-month-old infants," *J. Exp. Child Psychol.*, vol. 151, pp. 5–17, Nov. 2016.

[14] S. A. Graham and D. Poulin-Dubois, "Infants' reliance on shape to generalize novel labels to animate and inanimate objects," *J. Child Lang.*, vol. 26, no. 2, pp. 295–320, 1999.

[15] K. Plunkett, J.-F. Hu, and L. B. Cohen, "Labels can override perceptual categories in early infancy," *Cognition*, vol. 106, no. 2, pp. 665–681, Feb. 2008.

[16] M. H. Bornstein and C. Mash, "Experience-based and on-line categorization of objects in early infancy," *Child Develop.*, vol. 81, no. 3, pp. 884–897, 2010.

[17] K. B. Hurley and L. M. Oakes, "Experience and distribution of attention: Pet exposure and infants' scanning of animal images," *J. Cogn. Develop.*, vol. 16, no. 1, pp. 11–30, Jan. 2015.

[18] R. L. Fantz, "Visual experience in infants: Decreased attention to familiar patterns relative to novel ones," *Science*, vol. 146, no. 3644, pp. 668–670, 1964.

[19] C. Houston-Price and S. Nakai, "Distinguishing novelty and familiarity effects in infant preference procedures," *Infant Child Develop.*, vol. 13, no. 4, pp. 341–348, Dec. 2004.

[20] N. Mani and K. Plunkett, "In the infant's mind's ear: Evidence for implicit naming in 18-month-olds," *Psychol. Sci.*, vol. 21, no. 7, pp. 908–913, Jul. 2010.

[21] D. A. Baldwin and E. M. Markman, "Establishing word-object relations: A first step," *Child Develop.*, vol. 60, no. 2, pp. 381–398, Apr. 1989.

[22] N. Mani and K. Plunkett, "Phonological priming and cohort effects in toddlers," *Cognition*, vol. 121, no. 2, pp. 196–206, Nov. 2011.

[23] N. Mani, S. Durrant, and C. Floccia, "Activation of phonological and semantic codes in toddlers," *J. Memory Lang.*, vol. 66, no. 4, pp. 612–622, May 2012.

[24] J. L. McClelland, "The place of modeling in cognitive science," *Topics Cogn. Sci.*, vol. 1, no. 1, pp. 11–38, Jan. 2009.

[25] A. F. Morse and A. Cangelosi, "Why are there developmental stages in language learning? A developmental robotics model of language development," *Cogn. Sci.*, vol. 41, pp. 32–51, Feb. 2017.

[26] M. K. Fleming and G. W. Cottrell, "Categorization of faces using unsupervised feature extraction," in *Proc. Neural Netw. IJCNN Int. Joint Conf.*, 1990, pp. 65–70.

[27] G. Westermann and D. Mareschal, "From parts to wholes: Mechanisms of development in infant visual object processing," *Infancy*, vol. 5, no. 2, pp. 131–151, 2004.

[28] D. Mareschal and R. French, "Mechanisms of categorization in infancy," *Infancy*, vol. 1, no. 1, pp. 59–76, 2000.

[29] G. Westermann and D. Mareschal, "Mechanisms of developmental change in infant categorization," *Cogn. Develop.*, vol. 27, no. 4, pp. 367–382, Oct. 2012.

[30] K. E. Twomey and G. Westermann, "Curiosity-based learning in infants: A neurocomputational approach," *Develop. Sci.*, vol. 21, no. 4, Oct. 2017, Art. no. e12629.

[31] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[32] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *J. Stat. Softw.*, vol. 67, no. 1, pp. 1–48, 2015.

[33] D. J. Barr, R. Levy, C. Scheepers, and H. J. Tily, "Random effects structure for confirmatory hypothesis testing: Keep it maximal," *J. Memory Lang.*, vol. 68, no. 3, pp. 255–278, Apr. 2013.

[34] V. M. Sloutsky, Y.-F. Lo, and A. V. Fisher, "How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference," *Child Develop.*, vol. 72, no. 6, pp. 1695–1709, 2001.

[35] V. M. Sloutsky, "The role of similarity in the development of categorization," *Trends Cogn. Sci.*, vol. 7, no. 6, pp. 246–251, Jun. 2003.

[36] S. Waxman and A. Booth, "The origins and evolution of links between word learning and conceptual organization: New evidence from 11-month-olds," *Develop. Sci.*, vol. 6, no. 2, pp. 128–135, 2003.

[37] A. L. Fulkerson and S. R. Waxman, "Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds," *Cognition*, vol. 105, no. 1, pp. 218–228, Oct. 2007.

[38] J. S. Horst, K. L. Parsons, and N. M. Bryan, "Get the story straight: Contextual repetition promotes word learning from storybooks," *Front. Psychol.*, vol. 2, p. 17, Feb. 2011.

[39] T. T. Rogers and J. L. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach*. Cambridge, MA, USA: MIT Press, 2004.

[40] T. Kohonen, "The self-organizing map," *Neurocomputing*, vol. 21, no. 1, pp. 1–6, 1998.

[41] C. Heyes, "When does social learning become cultural learning?" *Develop. Sci.*, vol. 20, no. 2, Mar. 2017, Art. no. e12350.

**Arthur Capelier-Mourguy** received the B.S. degree in applied mathematics and social sciences from the University of Bordeaux, Bordeaux, France, in 2013 and the M.Res. degree in cognitive sciences from the CogMaster in Paris (EHESS), Paris, France, in 2015. He is currently working toward the Ph.D. degree in psychology as a Leverhulme Trust Doctoral Scholar at Lancaster University, Lancaster, U.K.

His current research interest includes understanding and modeling the effect of auditory labels on object perception along development.

**Katherine E. Twomey** received the B.A. degree (honors) in English language, the M.Res. degree in psychological methods, and the Ph.D. degree in psychology from the University of Sussex, Brighton, U.K., in 2008, 2009 and 2012, respectively.

From 2012 to 2014, she was a Postdoctoral Research Associate with the University of Liverpool, Liverpool, U.K. From 2014 to 2017, she was a Senior Research Associate with ESRC International Centre for Language and Communicative Development (LuCiD), Lancaster University, Lancaster, U.K. Since 2017, she has been a Lecturer with the Division of Human Communication, Development and Hearing, University of Manchester, Manchester, U.K. Her current research interests include the interplay between language acquisition and nonlinguistic representations using neural network and experimental techniques.

Dr. Twomey was a recipient of the ESRC Future Research Leaders Fellowship in support of her computational and looking-time-based investigations of curiosity-driven language learning in 2016.

**Gert Westermann** received the Ph.D. degree in cognitive science from the University of Edinburgh, Edinburgh, U.K.

He was with the Sony Computer Science Laboratory, Paris, France, before an academic career, Birkbeck College, London, Oxford Brookes University, Oxford, U.K. Since 2011, he has been a Professor at the Department of Psychology, Lancaster University, Lancaster, U.K. From 2016 to 2017, he was a British Academy/Leverhulme Trust Senior Research Fellow. His research focuses on infant cognitive development with a focus on language and categorization.