Are Mixtures of Experts Psychologically Plausible?

Sébastien Hélie[1], Gyslain Giguère[1], Denis Cousineau[2], and Robert Proulx[3]

[1] Université du Québec À Montréal, Computer Science, C.P. 8888 Succ. Centre-Ville,
Montréal, H3C 3P8, Canada
{Helie.Sebastien, Giguere.Gyslain}@courrier.uqam.ca
http://www.fas.umontreal.ca/psy/GRPLABS/vic/
[2] Université de Montréal, Psychology, C.P. 6128 Succ. Centre-Ville, Montréal, H3C 3J7,
Canada
Denis.Cousineau@umontreal.ca
http://www.mapageweb.umontreal.ca/cousined/home/
[3] Université du Québec À Montréal, Psychology, C.P. 8888 Succ. Centre-Ville, Montréal,
H3C 3P8, Canada
Proulx.Robert@uqam.ca

In this paper, we explore the generality of knowledge partitioning, the empirical equivalent of mixture-of-experts models. In the present study, we varied Lewandowsky et al.'s experimental settings in three ways: the amount of background knowledge given to the participants was increased, the size of the display was diminished and distracting information was added. The results show that increasing the amount of knowledge available to the participants does not qualitatively alter their performance. However, increasing the difficulty of stimulus estimation resulted in non-linear knowledge partitioning, which challenges the adequacy of POLE, a mixture-of-experts model developed to explain KP. Finally, adding distracting information to the display resulted in a smaller proportion of participants using knowledge partitioning to achieve the task. We conclude that mixture-of-experts are adequate psychological models but further research is needed to predict the presence and nature of the experts used.

1 Introduction

Recently, a new learning theory, knowledge partitioning (KP) [1, 2], has been proposed in the field of function learning. Lewandowsky and his colleagues were particularly interested in the use of cues in expertise. In order to understand their original manipulations, we now provide a little background knowledge in the field of firefighting.

Usually, forest fires tend to spread with the wind and uphill. However, when the slope of the terrain and the wind direction are in opposition, the fire spread uphill if the wind speed is low or downhill if the wind speed is sufficient to overcome the fire's propensity to spread uphill. In short, when the wind speed is low, the speed of the fire spread is described by a negative function; when the wind speed is high, it is described by a positive-slope function. Overall, the function relating speed of spread to wind speed is a concave quadratic function where the vertex indicates the point at which the force applied by the wind overpowers the tendency of the fire to spread uphill.

Another important aspect of firefighting is the use of back-burning fires to control the reach of the to-be-controlled fire. Back-burning fires are lit by firefighters to reduce the amount of fuel available to forest fires. Usually, a back-burner is lit when wind speed is low; otherwise the firefighters might lose control of this second fire.

Lewandowsky and Kirsner [1] asked experienced firefighters to estimate the spread of fires, after a short time period had elapsed, as a function of wind speed. On each trial, a specific wind speed was presented along with a context which identified the type of fire (to-be-controlled or back-burner). Results showed that experienced firefighters made different estimations of the fire's spread depending on the context associated with the wind speed. Lewandowsky and his colleague argued that, because back-burners where usually encountered in low wind speed situations and to-be-controlled fires in high wind speed situations, the firefighters' expertise affected their responses: these estimations suggested that back-burning fires moved uphill at a speed which is negatively related to wind speed, while to-be-controlled fires moved downhill, at a speed positively related to wind speed. However, it is obvious that both types of fires obey the same physical rules and should have generated the same estimates, whichever context was presented. Hence, expertise might have encouraged the use of independent parcels of knowledge which are triggered by particular cue values (here the context). This is what Lewandowsky et al. called "knowledge partitioning".

In the preceding experiment, the co-occurrence of particular wind speeds and types of fire was learned through many years of field experience. Lewandowsky et al. [2] designed a paradigm to test whether novices would also use this strategy in a function learning task. In this second experiment, the participants were taught basic firefighting background knowledge and trained in a standard function learning experiment using a quadratic concave function. Every stimulus (wind speed) was also accompanied by a context label, which was associated to a different half of the function during training. This manipulation aimed at recreating the bias present in experienced firefighters' knowledge, for which back-burners are usually encountered in low wind speed situations and to-be-controlled fires in high wind speed conditions. At test, every stimulus was presented twice: once as a back-burner and once as a to-be-controlled fire. Results showed that participants easily achieved the task, but more importantly, that their knowledge led to dramatically different responses to identical stimuli presented in different contexts. Back-burners' speeds were underestimated in high wind speed conditions whereas to-be-controlled fires' speeds were accordingly underestimated in low speed wind conditions. Nevertheless, spreading speeds were almost perfectly estimated when wind speeds appeared in their usual context. Thus, KP might be a useful heuristic to simplify complex relations between stimuli and responses.

These findings constitute empirical evidence favouring the psychological plausibility of mixture-of-experts models [3, 4]. In this type of model, a gating network is used to identify the expert which is best suited to achieve a task. In KP, the context plays the role of gating cue and the local experts are linear. Lewandowsky and his colleagues proposed one such model: Population of Linear Experts (POLE) [5]. In POLE, when a stimulus is encountered, a gating mechanism directs it to the correct expert, which represents one of many linear functions with different slopes and intercepts. There are enough experts to cover the entire stimulus space and, in the version proposed in [5], only the gating system has adjustable weights. Once an expert is chosen, it computes the answer accordingly.

This model [5] possesses three important properties. First, POLE accounts for all past results in the function learning literature by using KP. Second, experts do not blend together. Therefore, the system always commits to a cue-value and chooses an expert accordingly (KP). Third, each expert represents a linear relationship between the stimulus and the response.

In the present study, we tested the robustness of the empirical support [1, 2, 5, 6] to mixtures-of-experts models in general [3, 4] and to POLE in particular [5]. Human data were collected by altering Lewandowsky et al.'s [2] experimental settings. First, each participant was taught extensive firefighting background knowledge. Second, three groups of participants were tested: the first group performed the same task as Lewandowsky et al. [2], the second group was trained in the same task with smaller stimuli and the third group was trained with settings identical to the second, except that distracting information was added to the display (constant visual markers).

2 Experiment

2.1 Method

This experiment is an extension of Lewandowsky et al.'s Experiment 1, systematic condition [2]. Therefore, this section bears on their original methodology.

2.1.1 Participants

Fifty-four undergraduate students from the Université de Montréal participated in this experiment. Eighteen participants were trained in a reproduction of Lewandowsky et al. [2] (control group), eighteen were trained with small stimuli (small stimuli group), and the remaining participants were trained with a smaller display and distracting information (distracting information group). In each group, six participants were assigned to the complete condition, six to the left-only condition, and the remaining six to the right-only condition. Participants in the complete conditions received 7$ as compensation for their time, and those in the left-only or right-only conditions received 5$. The experiment was conducted in French.

2.1.2 Material

Participants were tested individually. All instructions and stimuli were presented on 43 cm (17 inch) monitors connected to PCs. Participants were positioned approximately 60 cm away from the monitor. The experimental task was programmed using Sun Microsystems' Java J2SDK1.4.1. The program was used to present the material and record the participants' answers.

2.1.3 Stimuli

Participants were expected to learn a concave quadratic function in which the fire's spreading speed (F) was related to wind speed (W) in the subsequent manner: $F(W) = 24.2 – 1.8W + 0.05W^2$. Wind direction always opposed slope, and the vertex of the function (W = 18) represented the point at which the force of the wind balanced the effect of the slope. To the left of that point, fire speed decreased with increasing wind, without changing the direction of the fire spread. Lewandowsky et al. [2] referred to these fires as "slope-driven". To the right of the vertex, fires were "wind-driven" and their speed increased as a function of wind speed. During training, 36 stimuli were used, ranging from wind speeds of 0 to 36, omitting the vertex of the function. At test, the omitted wind speed of 18 was included, resulting in a total of 37 transfer stimuli.

On each trial, a horizontal arrow, whose length was proportional to a particular wind speed (henceforth referred to as the stimulus), was shown at the top of the display. The minimal arrow length, associated with the value 0, was approximately 5.8 cm for the small stimuli and distracting information groups and 0.7 cm for the control group. The maximal length, associated with the value 36, was approximately 26 cm for the small stimuli and distracting information groups and 31 cm for the control group. Thus, in the small stimuli and the distracting information groups, the shortest arrow occupied 1/6 of the display and the longest 5/6. In the control group, the arrows spanned the entire monitor. No numerical values for wind or fire speed were shown. Participants were to consider each fire in a context represented both by a textual label and the color of the arrow (blue for *Back-burning* and red for *Firefighting*). In the distracting information group, visual markers were added to the display to indicate the minimum and maximum possible stimulus lengths. The markers were the only difference between the small stimuli group and the distracting information group.

Participants were asked to predict the speed of the fire (notwithstanding its direction of spread) by moving a sliding pointer along a 23.3 cm-scale positioned in the left part of the display. The scale was labeled *slow* at the bottom and *fast* at the top, without any incremental values or tick marks.

After each training trial, the participant's response was followed by a feedback arrow. The arrow was located next to the response scale to indicate the correct speed of spread. Also, a message appeared in a rectangle at the bottom center of the screen to encourage the participant to perform better (yellow rectangle) or to indicate that the response was satisfying (green rectangle). Predictions deviating by 5 or more units (approximately 7.2 cm) from the correct answer were accompanied by the former (yellow message) while acceptable performances were accompanied by the latter (green message). Participants were required to acknowledge feedback by a mouse click. The inter-stimulus interval (ISI) was 2 seconds, and the textual context-label always preceded the stimulus by 1 second. At test, feedback was absent.

2.1.4 Procedure

The procedure was identical for all groups, and differed only according to conditions. In all conditions, each stimulus was presented five times during training. Hence, there was a total of 180 trials for the complete conditions, but only 90 trials for the left-only and right-only conditions (because training was restrained to one half of the function).

In all conditions, 90% of fire speeds occurred in their respective contexts, and the remaining 10% were presented in the opposite context. However, in the left-only and the right-only conditions, all stimuli were presented in the same context (back-burning for left-only and firefighting for right-only). All magnitudes were presented once within each block of 36 trials (18 for the left-only and the right-only conditions), except during the first block, where magnitudes were presented in a blocked manner.

After completion of the training trials, participants in all conditions completed the same transfer test. The transfer test involved predicting the fire speed of all stimuli in both contexts.

2.2 Results

The performance of one participant from the small stimuli group, complete condition, deteriorated with practice ($F(4, 175) = 2.54$, $p < .05$). Therefore, this participant was excluded from the following analyses.

2.2.1 Training

The participants' *Absolute Deviation from Function* (ADF) was used to evaluate learning. The learning curves are shown in Fig. 1. As seen, participants in all conditions from all groups improved their ADF and were thus able to learn the function. Also, Fig. 1 suggests no effects of groups or conditions.

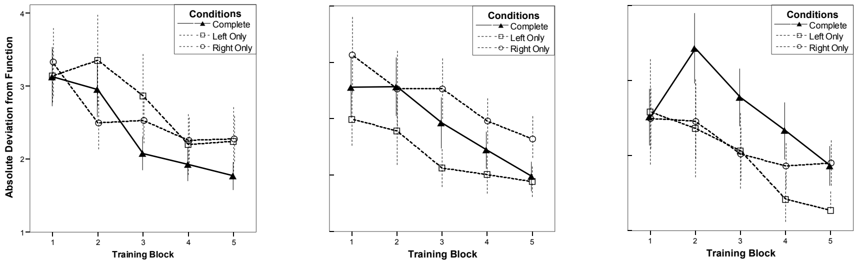

Fig. 1. Participants' ADF during the training phase. The left panel shows performance of participants in the distracting information group, the middle panel participants in the small stimuli group and the right panel shows the control group

A Group (small stimuli vs. distracting information vs. control) × Condition (complete, left-only, right-only) × Block (5, repeated measures) ANOVA was performed on the participants' ADF to corroborate what Fig. 1 hinted. First, the participants were able to diminish their ADF with practice: The mean ADF was 2.33 for the first block and diminished to 1.74 for the fifth block ($F(4, 176) = 14.32$, $p < .01$). However, this effect must be interpreted with care, because the Block × Group interaction was significant ($F(8, 176) = 10.24$, $p < .01$). Thus, the group's effect was further decomposed within each block. The groups significantly differed in the first block of training ($F(2, 44) = 28.42$, $p < .01$) but were similar in all other blocks (all $F(2, 44) < 1.63$, $p > .05$). Tukey A *post hoc* comparisons showed that the control group was significantly better than the other two at the beginning of the task (both differences > 0.96, $p < .01$). However, as suggested by the absence of group effect in the remaining blocks, this difference disappeared with training.

2.2.2 Group performance at test

KP can be detected experimentally by a difference in responses to a given stimulus in different contexts [2]. Fig. 2 shows transfer performances for participants trained in the complete conditions. As seen in the left panel, participants trained with distracting information have learned the function quite well. Further, answers in both contexts matched the quadratic function and were not affected by context.

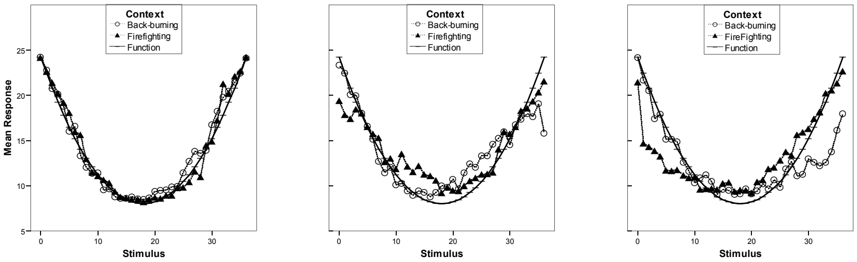

Fig. 2. Mean responses at test in each context. Panels represent the same groups as in Fig. 1

Responses of participants trained with small stimuli are shown in the middle panel. As expected, their responses at test were affected by the context (compare with the left panel), but in a non-systematic way. In comparison, the deviations found by Lewandowsky et al. [2] were systematic: low wind speeds resulted in an underestimation of the speed of fire spread in the firefighting context and high wind speeds accordingly resulted in underestimations in the back-burning context. This is exactly the pattern of results found in the control group (see the rightmost panel). In the middle panel, the underestimations are present (to a lesser extent), but mid-range wind speeds were overestimated.

A better way to highlight the difference in responses to a given stimulus is to compute the signed differences [2]. A signed difference is computed by subtracting the answer given at test to each stimulus in the back-burning context from the answer given to the same stimulus in the firefighting context. Signed differences randomly aggregated around zero would suggest the absence of partitioning, while signed differences systematically deviating in one direction would indicate the presence of partitioning.

The left panel of Fig. 3 shows that participants trained with distracting information are not partitioning their knowledge: their signed differences are randomly aggregated around the abscissa. However, the signed differences of participants trained with small stimuli are more intriguing (middle panel): they are substantially deviating from the abscissa in a sine-like way. Individual analyses might explain these results.

Finally, participants in the control group showed the expected pattern of results: signed differences are negative to the left of the vertex and positive to the right.

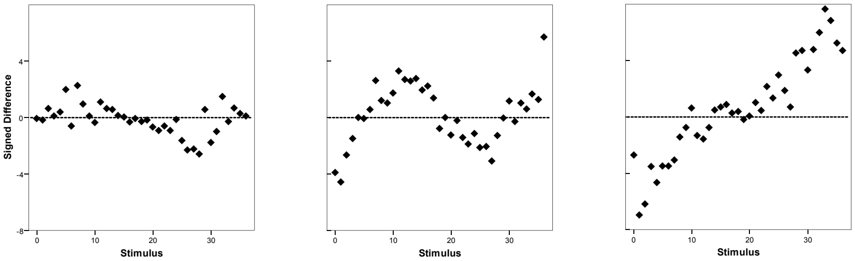

Fig. 3. Signed differences of the participants at test. Panels represent the same groups as in Fig. 1

2.2.3 Individual results at test

The preceding section showed that, at a group level, participants trained with distracting information did not seem to partition their knowledge while those trained without visual markers (small stimuli and control groups) did. However, the preceding analyses may not tell the whole story, considering that Lewandowsky et al. [2] found important individual differences relating to KP. Precisely, previous research found between 13% [6] and 50% [5] of participants who were not partitioning their knowledge. Therefore, it is relevant to verify if the effects found in section 2.2.2 were representative of the entire groups of participants.

One way of classifying the participants as partitioners (P) or non-partitioners (NP) is to individually plot their signed differences and estimate the best-fitting linear model using a linear regression. A slope which is significantly different from zero suggests a systematic effect of context, namely KP. On the other hand, a slope of zero suggests no clear effect of context. Table 1 shows the slope and intercept individually estimated for each participant.

Table 1 shows that, in the group trained with distracting information, all but one participant fit a model with an absolute slope of 0.05 or less. Because these slopes did not differ significantly from zero (p = .05), participants were classified as non-partitioners (NPs). The exact opposite was true of participants in the small stimuli group: All but one participant had an absolute slope greater than 0.20. Hence, these participants were classified as partitioners (Ps). In the control group, the best-fitting slope of two of the six participants was smaller than 0.10: these slopes did not significantly differ from zero (p = .05) and these participants were classified as NP. The four remaining participants were classified as Ps.

The proportion of Ps in the small stimuli group significantly differed from the proportion of Ps in the distracting information group according to a binomial test (B(5, 1/6) = 4, p < .01). The proportion of Ps in the small stimuli group (80%) is well in range with past literature while the proportion of Ps in the group trained with distracting information (16.7%) is below past results. The proportion of Ps in the control group (67%) is similar to Lewandowsky et al.'s results [2] and does not significantly differ from the small stimuli group ($B(6, 4/5) = 4$, $p > .05$). However, this proportion of Ps differs from the proportion found in the distracting information group ($B(6, 1/6) = 4$, $p < .01$).

Table 1. Estimated Parameters for the Best-Fitting Linear Models

		Estimated Parameters			
	Participant	Slope	Intercept	r^2	Classification
Distracting Information					
	110	-0.03	0.63	0.02	NP
	111	0.00	-0.13	0.00	NP
	112	-0.25	5.48	0.34	P
	120	-0.02	0.37	0.01	NP
	121	0.05	-1.45	0.07	NP
	122	0.01	-1.51	0.00	NP
Small Stimuli					
	210	-0.41	8.03	0.58	P
	211	-0.02	0.19	0.00	NP
	212	0.22	-4.19	0.29	P
	220	0.63	-9.04	0.93	P
	221	-0.23	2.68	0.32	P
Control					
	310	0.60	-10.7	0.87	P
	311	-0.02	0.17	0.01	NP
	312	-0.07	2.16	0.09	NP
	320	0.54	-9.45	0.69	P
	321	0.13	-2.61	0.14	P
	322	0.65	-8.88	0.89	P

Note. P = Partitionners; NP = Non-Partitionners

Together, these results suggest that when distracting information is present in the display, fewer participants use the KP heuristic. Also, it is noteworthy that all the Ps in the control group showed positive slopes, which is consistent with the linear experts hypothesis [2, 5]. However, half of the Ps in the small stimuli group and the only P in the distracting information group had negative slopes, which is consistent with the sine-like pattern of Fig. 3. The overestimation of moderate wind speeds is also present in the middle panel of Fig. 2 and further inspection of the middle panel suggests a partitioning of the stimuli in two quadratic functions with skewed vertices. Therefore, diminishing the stimulus' length does not prevent participants from using KP but entails a different, non-linear, type of partitioning, which is not consistent with POLE's predictions [5].

2.2.4 Independence of knowledge parcels

As Lewandowsky et al. [2] pointed out, participants who were uniquely trained on the left or right part of the function represent extreme cases of KP: they possess a single expert, associated with a single context. Therefore, if the knowledge of Ps in each context is truly independent, their responses should be similar to the left-only condition in the back-burning context and the right-only condition in the firefighting context. In the case of non-linear Ps, responses in the back-burning context were similar to responses from participants uniquely trained in this particular context (left-only condition: $r = 0.87$). However, the correlation between partitioners' responses in the firefighting context and those from the right-only condition was smaller ($r = 0.69$). This difference is significant according to Fisher's Z transform test ($Z = 2$, $p < .05$). Therefore, the back-burning parcel of knowledge seems more hermetic than the firefighting parcel. Also, results from Lewandowsky et al. suggested higher correlation coefficients [2].

In the case of linear Ps, responses from knowledge partitioners were similar to responses from participants trained in the left-only ($r = 0.81$) and right-only ($r = 0.83$) conditions (in the back-burning and firefighting contexts respectively). Also, the difference between correlation coefficients is not statistically significant ($Z = 0.25$, $p > .05$). Knowledge about the other half of the function acquired in another context did not affect the participants' responses, suggesting that knowledge was completely partitioned.

3 General discussion

In the Experiment, the usual settings used to assess the presence of KP [2, 5] were varied to check the robustness of this phenomenon. Precisely, three modifications were made: adding details in the cover story, reducing the length of the stimuli, and adding distracting information. First, increasing the level of detail in the cover story did not qualitatively alter performance. However, participants who received detailed cover stories (control group) seemed to have better performed than those in [2]. Second, it is well established that diminishing the span of the stimuli increases discrimination difficulty [7], hence making stimulus estimation more difficult. In the small stimuli group, participants used KP to partition their knowledge but showed negatively-sloped best-fitting linear models (Table 1). This counter-intuitive result was first hinted by sine-like signed differences (Fig. 3) and the use of non-linear expert functions with skewed vertices (Fig. 2). Finally, adding distracting information to the display resulted in fewer participants using KP to achieve the task (distracting information group). However, these participants, who did not use KP to simplify the function, were still able to learn it (as shown by an absence of group effect in the ANOVA).

3.1 Implications for current cognitive modeling

These findings have numerous implications for cognitive modeling. Results from the control group confirmed the adequacy of our reproduction of [2] and suggested that background knowledge, which is not contradictory with task demands [8], do not qualitatively alter the participants' performance: it only helps to perform better. However, results from the small stimuli group are challenging the POLE model [5]: when stimuli are more difficult to estimate, participants still partition their knowledge but non-linear experts are used. This strategy may be adaptive because it minimizes the error resulting from an erroneous choice of expert: if the input is uncertain, the probability of wrongfully gating the input to an inadequate expert is increased. However, if the experts are more complex (in this case quadratic instead of linear), the estimation of a sub-optimal expert results in a smaller error. Therefore, the results from the small stimuli group, while challenging to POLE's predictions, do not invalidate mixture-of-experts models in general [3, 4].

The results from the distracting information group are more problematic to both POLE [5] and general mixture-of-experts models [3, 4], because they show that when distracting information is present in the display, participants do not seem to be using the KP heuristic. Instead, participants are learning the quadratic function by simple associative learning. These findings might still be explained by the degenerate case of the mixture-of-experts, in which a single quadratic expert is used.

Together, these results confirms that KP [1, 2, 5, 6], which is the empirical counterpart to mixture-of-experts models [3, 4], is a strategy used to achieve psychological tasks. However, this heuristic is less ubiquitous than Lewandowsky and his colleagues previously thought [5] and the constraint of using linear experts is too restrictive. Therefore, mixture-of-experts are adequate models of human cognition but further research is needed to detect the presence of experts (to distinguish simple associative learning from the degenerate case of using a single expert) as well as to determine the nature of the experts used to achieve particular tasks.

References

1. Lewandowsky, S., Kirsner, K.: Knowledge Partitioning: Context-Dependent Use of Expertise. Memory & Cognition **28** (2000) 295-305
2. Lewandowsky, S., Kalish, M., Ngang, S. K.: Simplified Learning in Complex Situations: Knowledge Partitionning in Function Learning. Journal of Experimental Psychology: General **131** (2002) 163-193
3. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive Mixtures of Local Experts. Neural Computation **3** (1991) 79-87
4. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, New York (1995)
5. Kalish, M.L., Lewandowsky, S., Kruschke, J.K.: Population of Linear Experts: Knowledge Partitioning and Function Learning. Psychological Review **111** (2004) 1072-1099
6. Yang, L.-X., Lewandowsky, S.: Context-Gated Knowledge Partitioning in Categorization. Journal of Experimental Psychology: Learning, Memory, and Cognition **29** (2003) 663-679
7. Goldstein, E.B.: Sensation & Perception. 5th edn. Brooks/Cole Publishing Company, Pacific Grove (1999)
8. Heit, E., Bott, L.: Knowledge Selection in Category Learning. The Psychology of Learning and Motivation **39** (2000) 163-199