Integrating specific and general information in decision making: An instance-based learning account

David Peebles and Chris N. H. Street

University of Huddersfield, Huddersfield, UK.

July 17, 2018

Author Note

Correspondence concerning this article should be addressed to David Peebles, Department of Psychology, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, UK. E-mail: d.peebles@hud.ac.uk

Abstract

We present a cognitive process model of adaptive decision making that provides a detailed, mechanistic account of the integration of individuating and context-general information demonstrated in the experiment of Street, Bischof, Vadillo, and Kingstone (2016). The model is grounded in instance-based learning theory and instantiated in the ACT-R theory of human cognitive architecture. The model provides a close fit to the data from the experiment, an accurate prediction of new data, and a parsimonious explanation of adaptive decision making based on a theory of human cognitive architecture and basic learning mechanisms. This is the first computational account of the processes underlying lie-truth judgement formation and it provides a rigorous test of adaptive lie detector (ALIED) theory. Consistent with ALIED, the model demonstrates that it is possible to consider the lie bias and truth bias as functionally equivalent.

Keywords: ALIED, ACT-R, Instance-Based Learning Theory, Lie Detection

Integrating specific and general information in decision making: An instance-based learning account

1. Introduction

1.1 Individuating and context-general cues in lie detection

Ordinarily, people are able to understand what others are thinking with a surprisingly high degree of accuracy. Allusions in speech and eye direction can be sufficient for understanding what another is thinking (Clark, 1996; Clark, Schreuder, & Buttrick, 1983; Tomasello, 1995). But when someone wants to deliberately hide their true thoughts, the ability to determine whether a person truly holds an asserted belief or not drops to near chance accuracy (Bond & DePaulo, 2006; Kenny & DePaulo, 1993). How do people assess whether someone is being honest in their claims?

One answer comes from a recent account of lie detection, the *Adaptive Lie Detector* theory (ALIED: Street, 2015). According to this account, people's judgements are informed by two types of knowledge: specific information relating to the particular circumstances under consideration and relevant general information that constitutes the background context and which informs important estimates such as the base rate or likelihood of the situation.

For example, Stephen may believe that Gill is lying when she claims that she did not cheat on a test. The belief is based on knowledge specific to this statement, which might be based on Gill's verbal or nonverbal behaviours, for example, or CCTV footage showing her cheating. This information individuates this statement, and as such is called *individuating* information (see Koehler, 1996). But Stephen may also believe that, in general, when people deny cheating they are more likely to tell the truth because most people do not cheat on tests. This generalised knowledge is not directly related to Stephen's current judgement about whether a specific statement is a lie, but rather generalises across statements in this situation. ALIED refers to this as *context-general* knowledge. When judging whether Gill is lying about past cheating, Stephen will combine this context-general knowledge with his consideration of specific individuating cues.

Individuating cues vary in their degree of diagnosticity however and if a particular cue is highly diagnostic (think of Pinocchio's nose) people can use this information to attain high accuracy. There is experimental evidence that people can indeed make good use of highly diagnostic information (Blair, Levine, & Shaw, 2010; Bond, Howard, Hutchison, & Masip, 2013; Levine & McCornack, 2014). But high accuracy rates are rare because individuating cues to deception are typically either unavailable (Levine, 2010) or when they are available, are weak and unreliable (B. M. DePaulo et al., 2003; Hartwig & Bond, 2011; Sporer & Schwandt, 2006, 2007).

In the absence of diagnostic individuating cues, ALIED argues that people will turn to their context-general knowledge to make an informed but overgeneralised judgement. This explains the frequently observed *truth bias* in people's judgements (Bond & DePaulo, 2006); because individuating cues typically have low diagnosticity, and because people predominantly tell the truth (B. M. DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Halevy, Shalvi, & Verschuere, 2014; Serota, Levine, & Boster, 2010), in most situations it is rational to make a truth judgement. Conversely, in situations where lying is more prevalent or where there is a widespread belief that people lie (e.g., Carr, 1968; P. J. DePaulo & DePaulo, 1989; Masip, Alonso, Garrido, & Herrero, 2009; Masip & Herrero, 2017; Meissner & Kassin, 2002; Millar & Millar, 1997), the bias is to judge statements as being untruthful. The truth bias found in normal situations is not a cognitive disposition therefore, but rather an adaptive judgement in the absence of specific information.

1.2 Investigating individuating and context-general cues

The interactive effect of individuating and context-general cues on lie detection has been investigated empirically (Street et al., 2016) where it was demonstrated that people's judgements are a function of both the diagnosticity of individuating cues and the nature of the context-general information they are given. As this experiment forms the basis of the research reported here we provide a brief summary below.

Experiment participants were told about a (fictitious) study in which players took part in a trivia game. The story described how players had an opportunity to cheat while the

experimenter was not looking but that each player was asked at the end of the game whether they had cheated or not. All players denied cheating but some told the truth (i.e., they actually did not cheat) while others denied it by lying (i.e., they cheated and later denied cheating). Players were video recorded while answering the question and the videos were subsequently coded for four behaviours: pitch of *voice*, degree of emotional expressiveness of the *face*, the number of periods of *silence* in their response, and the number of *self* references such as "I" and "me". After hearing the story, the experiment participants were told that their task was to learn the extent to which each behavioural cue indicated whether the person was lying or telling the truth (i.e., how diagnostic of the two behaviours each cue was).

The computer-based experiment consisted of two phases, a *training* phase of four forty-trial blocks followed by a *test* phase of two forty-trial blocks. On each trial of the training phase, participants were presented with one of the four cues and were required to respond whether they believed the cue indicated that the trivia game player was telling the truth or lying (by pressing one of two keys on the keyboard). Participants were then provided with feedback on the accuracy of their response and presented with the next trial. Participants saw four consecutive blocks of forty trials, one block for each cue. The key manipulation in the training phase was the diagnosticity of each cue (i.e., the percentage of truth/lying trials in a cue's block) with the proportions being 20/80, 30/70, 40/60, and 50/50. Each participant saw four different proportions (one per block) selected at random.

After the training phase participants were informed about the difficulty of the game players had taken part in. The purpose of this additional context-general information was to bias participants by shifting their baseline assumption regarding the likelihood of people lying and telling the truth. Some were told that the game had been easy and that most of the players could achieve their accuracy levels without having to cheat, and so most people told the truth when they denied cheating. Other participants were told that the game had been hard and that therefore if the players scored as well as was indicated most of them would have had to cheat and therefore must have been lying when they denied cheating.

After the manipulation, participants underwent the test phase of two blocks of forty

trials. The trails consisted of the four cues (twenty of each), presented individually in random order. Again, participants were required to respond whether they believed the cue indicated that the player was lying or being truthful but this time no feedback on the response was provided.

The key dependent measure in the experiment was the proportion of truth judgements (PTJ calculated as the number of truth judgements divided by the total number of judgements made) for each diagnosticity and these are shown for both the *easy* and *hard* conditions in Figure 1.

Analysis of the data revealed a significant main effect of cue diagnosticity, showing that as cues became more diagnostic of honesty the proportions of truth judgements increased. There was also a significant main effect of context, with more truth judgements being made in the easy condition compared to when people were told that the game was difficult. The analysis also revealed a significant interaction between cue diagnosticity and context in that the degree to which cue diagnosticity affected participant's responses differed between the easy and hard conditions.

The difference between the easy and hard conditions can be seen in the progression of the respective distances of the plotted data of the two conditions from the (dotted) line of test/training probability equality in Figure 1. As diagnosticity of the individuating cue increases, the distance generally increases for the easy condition but decreases in the hard condition. However the patterns of responses in both context conditions (easy, hard) are in fact remarkably similar if one is transformed by reflecting along the equality line and the .5 points and then superimposed over the other (as shown in Figure 2 where the hard condition data is transformed). The correlation and mean deviation between the easy and transformed hard data are .97 and .06 respectively. An explanation for this similarity will be provided below.

The experiment provided a clear, quantitative demonstration of how people's judgements result from an interaction between knowledge formed from the history of experience underlying the diagnosticity of individuating cues and context-general knowledge about the prevalence of lying. Questions remain however concerning how the

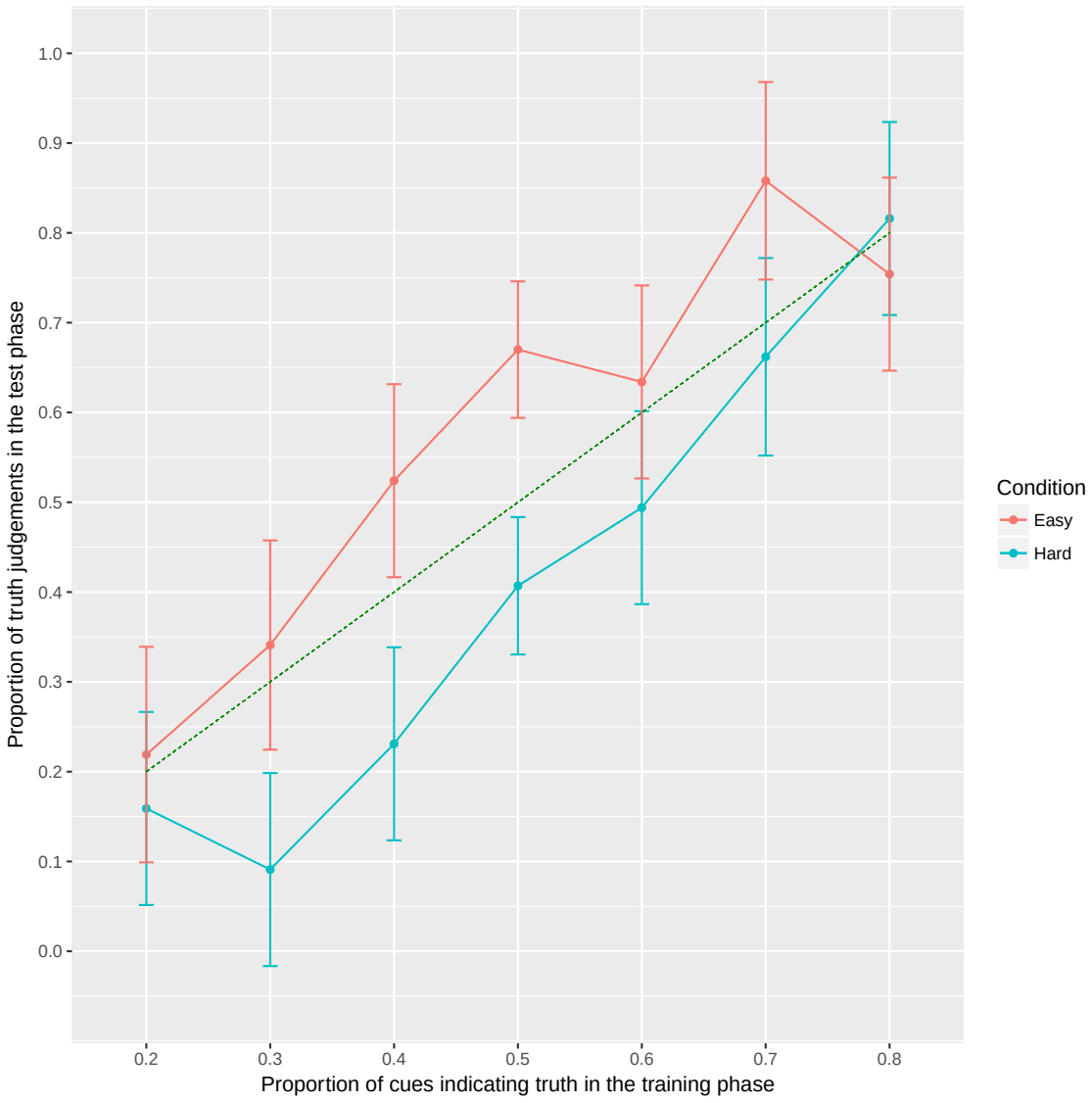

Figure 1. Proportion of truth judgements in the test phase for each proportion of truth cue conditions in the training phase, easy and hard game conditions, from the experiment reported in Street, Bischof, Vadillo, and Kingstone (2016). The dotted line indicates where the two proportions are equal. Error bars denote 95% confidence intervals.

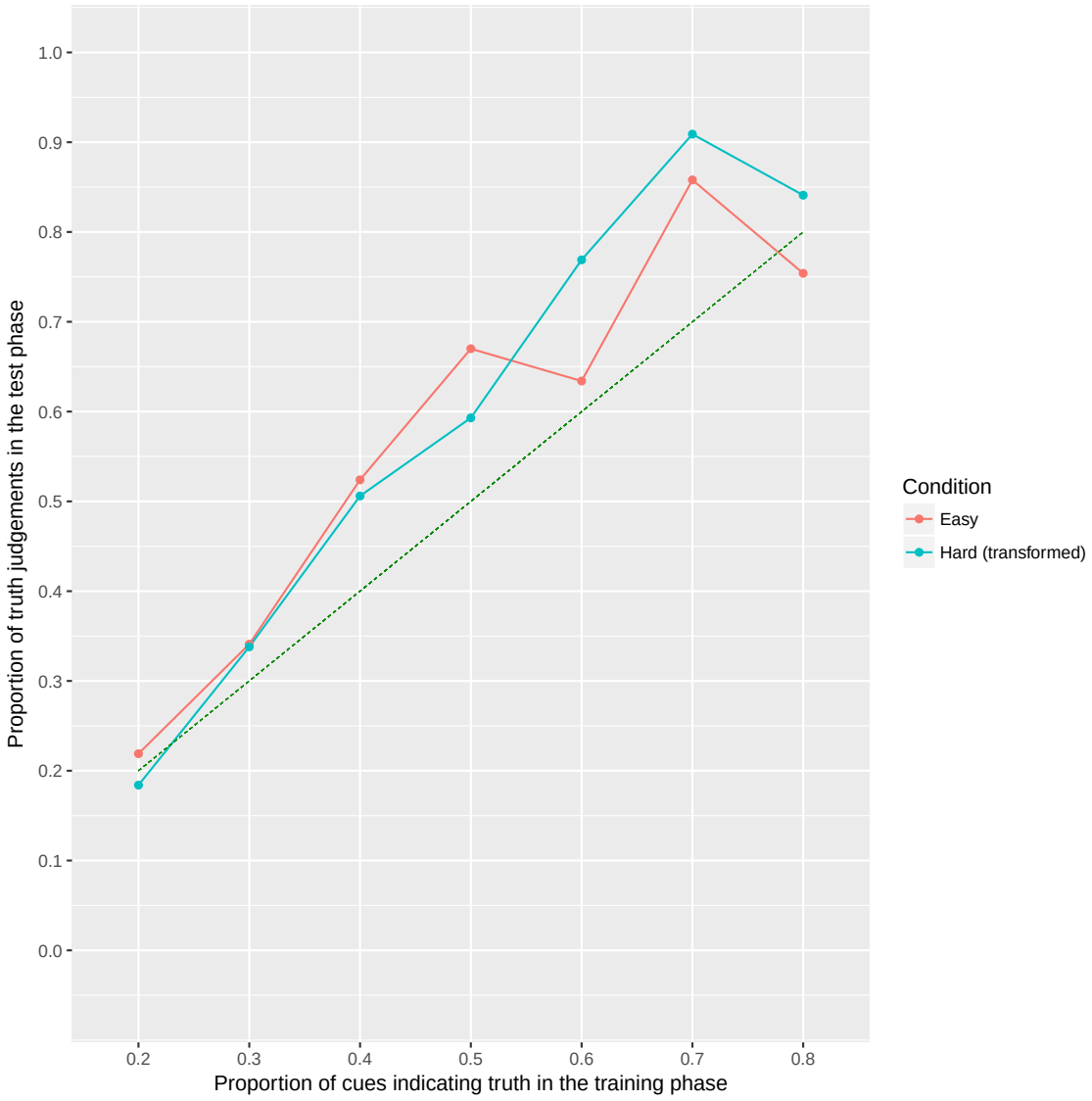

Figure 2. Data from Figure 1 with the hard game condition data transformed and superimposed on the easy condition data.

two types of knowledge are learned and cognitively represented and what the mechanisms of interaction that produce the observed adaptive behaviour may be.

In the rest of this paper we provide an answer to these questions in the form of a cognitive process model of adaptive decision making that provides a detailed, mechanistic account of the interaction between individuating and context-general cues grounded in instance-based learning theory (IBLT: C. Gonzalez, Lerch, and Lebiere, 2003) and instantiated in the ACT-R theory of human cognitive architecture (Anderson, 2007). In the following sections, we first outline the key assumptions of IBLT and then introduce the

ACT-R cognitive architecture, detailing the mechanisms relevant to the model. We then describe the model of the data from the experiment of Street et al. (2016).

1.3 Instance-based learning theory

Instance-based learning theory (C. Gonzalez et al., 2003) is an example of a general theoretical approach to understanding human cognition as being shaped by the learning and subsequent recall of individual experiences. Instance theories are prevalent in cognitive science and various forms of the theory have been used to account for a diverse range of functions, including automatisation (Logan, 1988), categorisation (Medin & Schaffer, 1978; Nosofsky, 1984), attention (Logan, 2002) and decision making (Gilboa & Schmeidler, 1995).

IBLT claims that during problem solving and decision making, knowledge is generated through the creation of *instances*, representations in long-term memory that record three elements pertaining to each event: the conditions (a set of contextual cues), the action or decision made, and the utility of the action (an evaluation of the outcome of the action or decision under those conditions). In subsequent problem solving or decision making, IBLT assumes that people attempt to retrieve past instances of similar situations from long-term memory and act accordingly, or otherwise employ some heuristic (e.g., random choice, loss minimisation or gain maximisation) if appropriate instances are unavailable.

A crucial element of IBLT which differentiates it from many other instance theories is that it specifies in detail the representations, learning mechanisms and decision processes underlying the theory. These are taken from the ACT-R cognitive architecture[1] and using this approach IBLT has been successful in accounting for a range of dynamic and intuitive decision making phenomena (e.g., Dutt, Ahn, & Gonzalez, 2013; C. Gonzalez & Wimisberg, 2007; Lejarraga, Dutt, & Gonzalez, 2012; Thomson, Lebiere, Anderson, & Staszewski, 2015). In the next section we describe ACT-R and the specific mechanisms utilised by IBLT.

---

[1] The mechanisms can be specified outside of ACT-R however. See C. Gonzalez, Dutt, and Lebiere (2013).

1.4 The ACT-R cognitive architecture

ACT-R is a theory of the core components of the human cognitive system (including perceptual, cognitive and motor processes) and how these components work together to produce intelligent behaviour. Like many other cognitive architectures, ACT-R aims to provide a rigorous, formal, information processing level of description that bridges the gap between the theories of neuroscience and those of the behavioural sciences (Cooper & Peebles, 2015). It incorporates well established theories of declarative and procedural knowledge representation, cognitive control, and learning to create complete, integrated processing models of cognition. An important feature of ACT-R is that it is implemented as a software system so that cognitive models can be built, run and compared with human performance data. In addition, it incorporates models of vision and motor control which can be connected to—and interact with—external task and simulation environments.

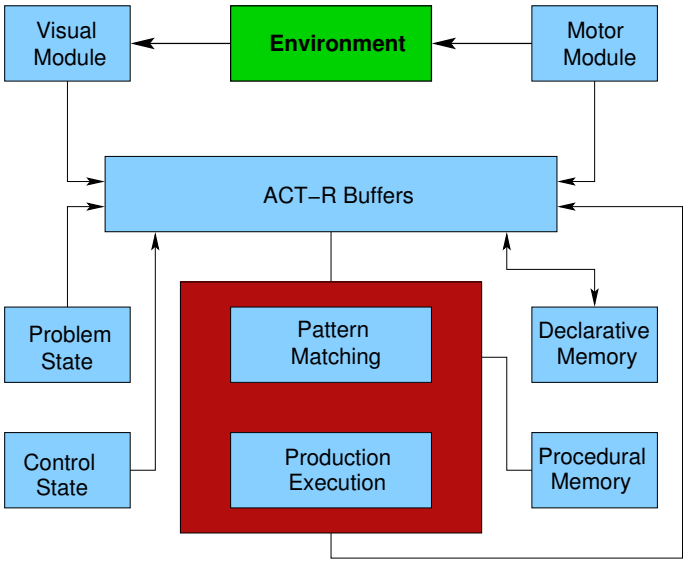

Figure 3. Functional components of the ACT-R cognitive architecture

ACT-R consists of a set of modules (shown in Figure 3) for declarative memory (a network of knowledge *chunks*), procedural memory (represented by sets of production rules that encode knowledge about behaviour), vision, and motor control. Cognition is goal driven and proceeds in ACT-R via a pattern matching process that attempts to find production rules with conditions that match the current state of the system. For example, in order to be executed, a particular production rule may require that a word has been read from the

computer screen, a response associated with that word has been retrieved from long-term declarative memory, and that the mouse pointer is located at the response button. If all of these conditions are true, then the rule "fires" and its actions will take effect (e.g., if the mouse is clicked, then the search for another word is initiated). Tasks are performed through the successive sequential actions of production rules.

ACT-R is a hybrid architecture that combines the symbolic knowledge level system with a subsymbolic system which computes numerical values related to the symbolic level entities in long-term memory. Two key values are knowledge chunk *activation* which determines the probability and speed of chunk retrieval and production rule *utility* which is used to resolve conflicts in production selection when two or more production rules match the state of the cognitive system.

Learning in ACT-R occurs in two ways. The first is in the creation of new knowledge (both chunks and production rules) in long-term memory during the course of task execution. The second is through the gradual adjustment of activation and utility values over time which affects the likelihood of symbolic entities being utilised in the future. As this latter form of learning for chunks is crucial to the operation of the current model, declarative memory and learning is now described in detail.

1.4.1 Declarative memory. The three mechanisms underlying IBLT and the cognitive model we describe below are: (a) the activation of declarative knowledge and how that activation affects the probability of choices over time, (b) the learning mechanism that adjusts a chunk's activation to reflect its history of use, and (c) partial matching during retrieval.

When chunks are created in declarative memory, they have an initial level of activation which decays over time and which determines the probability that they can be subsequently retrieved for future processing. In the case of the current experiment, chunks represent whether an individuating behavioural cue is indicative of honesty or deception. Memory retrieval in ACT-R occurs when a production rule contains a retrieval request to the declarative memory module containing one or more cues. For example, if a production contains the features of an object (e.g., [object = umbrella, colour = red]) then it may probe

declarative memory for a chunk using these features to retrieve additional previously stored items of information related to this object (e.g., [location = hallway]), if it has sufficient activation to be retrieved.

The activation of a chunk is the sum of three components: a *base-level* activation, reflecting the history of the chunk's use, an *associative* activation, reflecting its relevance to the current context, and a noise component, ϵ, generated from a logistic function with mean zero and variance s. The activation of a chunk i, A_i is defined as

$$A_i = B_i + \sum_{j \in C} W_j S_{ji} + \epsilon \tag{1}$$

where B_i is the base-level activation of chunk i, C is the context (i.e., the set of elements, j currently in the ACT-R buffers which constitute the current state of the system), W_j is the attentional weighting given to element j, and S_{ji} is the strength of association between element j and chunk i.

The retrieval probability of each chunk i, P_i is a function of its activation, A_i with increasing activation resulting in increased probability of recall. Retrieval probability is defined as

$$P_i = \frac{1}{1 + e^{\frac{-(A_i - \tau)}{s}}} \tag{2}$$

where s is a noise parameter that tempers the relationship between activation and recall probability and τ is the threshold activation below which chunks will not be retrieved. For a set of chunks that match a retrieval request, the probability of chunk i being selected, P_i, becomes a function of its activation, A_i relative to the activations of the others according to the Boltzmann *softmax* equation

$$P_i = \frac{e^{A_i/s}}{\sum_j e^{A_j/s}} \tag{3}$$

The chunk with the highest activation among those that match the request is the one most likely to be retrieved. If no chunk has an activation greater than the retrieval threshold then no chunk will be retrieved and a retrieval failure will be signalled.

1.4.2 Base-level learning. A chunk's base-level activation decays as a power function of time but is increased with each 'presentation' (i.e., when a chunk initially enters into declarative memory or when an existing chunk's activation is increased by each

additional experience of that chunk). The learning of base-level activation for a chunk i is described as

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{4}$$

where n is the number of presentations for chunk i, t_j is the time since the j^{th} presentation, and d is the parameter determining the activation decay rate.

Base-level activation learning is the mechanism by which ACT-R is able to produce behaviour that is governed by past experience and the accumulation of evidence and explain a wide variety of memory phenomena including the power laws of learning and forgetting, and the effects of frequency and recency on recall probability (e.g., Anderson & Schooler, 1991; Pavlik & Anderson, 2005).

1.4.3 Partial matching. As described above, the declarative retrieval mechanism is rather rigid in that if no chunk is an exact match to the items in the retrieval request, then no retrieval will occur and a retrieval failure will be signalled. This is the simplest option but ACT-R has more sophisticated mechanisms to capture the flexibility of human memory and also common human retrieval errors, for example when an incorrect (but quite similar to the probe) chunk is recalled.

One approach is to not treat cue-chunk similarity as a binary all or none affair but to use a *partial matching* mechanism to compute the similarity between the probe and memory chunks. With partial matching all chunks of the same type as the probe are taken into consideration and the activation of each chunk, i is modified in proportion to its similarity to the probe according to ACT-R's partial matching equation

$$P_i = \sum_k PM_{ji} \tag{5}$$

In this equation, the partial matching value of chunk i, P_i is computed as the sum of the similarity between each of its slots with the corresponding slot j in the probe, M_{ji} multiplied by a *mismatch penalty* value, P (which is constant over all slots). M_{ji} is typically set to 0 when slot values are equal and -1 when they are not so that the reduction in a chunk's activation is proportional to the number of mismatching slots. The effect is that when a chunk in declarative memory does not match information in the retrieval request, a penalty

is applied to its activation, the severity of which is proportional to the number of mismatching slots. This reduces the likelihood of the chunk being subsequently recalled.

2. A model of the experiment

Having laid out the theoretical assumptions and computational mechanisms underlying our approach, we now describe the proposed model in detail[2]. The control flow of the model executing a single trial (in either the training or test phase) of the experiment is illustrated in Figure 4. Reflecting the simplicity of the task, the model is relatively small and straightforward, consisting of 15 production rules and two initial declarative chunks to represent the general knowledge (that we assume the experiment participants had and which was explicitly reinforced during the experiment) that easy games are associated with truth telling and hard games with lying.

For each training trial of the experiment, the model reads one of four cues ('voice', 'face', 'silence' and 'self') on the computer screen and uses it to probe its long-term memory for a stored fact concerning whether the cue is indicative of telling the truth or not. If the retrieval is successful, then the model responds "truth" or "lie" accordingly (by pressing either the "t" or "l" key on the keyboard). If no fact is retrieved then the model just 'guesses' by selecting one of the two responses at random.

On receipt of the model's response, the experiment software provides feedback ("Correct" or "Incorrect") on the computer screen which the model reads and uses to update its knowledge about the cue by encoding (or strengthening) the correct response in declarative memory. After a two-second delay, the next trial starts.

When the 80 trials of the training phase are complete the model is provided with the additional context-general information regarding the difficulty of the game (either "easy" or "hard" depending on the condition the model is being run on) and then given the 80 test phase trials to complete. The test phase is nearly identical to the training phase in that the model is provided with the cues again (but this time only 20 of each in random order) and must respond whether the cue is diagnostic of truth-telling or lying. As in the experiment,

[2] The model code is available on GitHub: https://github.com/djpeebles/act-r-lie-detection-model.

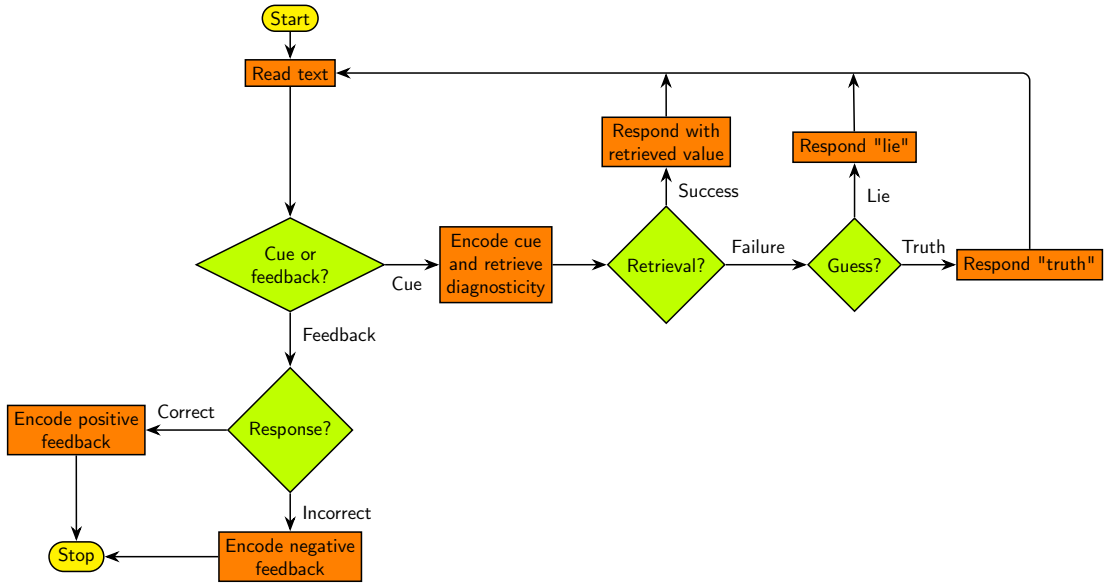

Figure 4. Control flow of the ACT-R model carrying out a trial of the experiment. Rectangles correspond to actions carried out by production rules, diamonds represent junctions in the experiment or model's behaviour.

no feedback was provided to the model in the test phase.

2.1 Comparing human and model performance

The model was evaluated by running it 150 times (to simulate 150 experiment participants) for each experiment condition and the data for each condition averaged. To fit the model to the human data, three parameters were adjusted: the retrieval threshold parameter (τ in Equation 2) was set to 0.4, the activation noise parameter (s in Equation 2 and the variance of the logistic function of ϵ in Equation 1) was set to 0.21, and the mismatch penalty parameter (P in Equation 5) was set to 0.65. All of these values are well within the typical ranges for ACT-R models and the same parameter values were used for both experiment (easy and hard trivia game) conditions.

The mean proportion of truth judgements as a function of cue diagnosticity in the training phase for ACT-R model and the human participants are compared for the easy and hard experiment conditions in Figure 5 and Figure 6 respectively. In both figures, the blue line plots the proportions produced by the model after the training phase and the dotted line

indicates where experiment condition proportions and response proportions are equal. The fit of the model to the human data for both difficulty conditions was very close, R^2 (easy) = 0.92, *RMSD* (easy) = 0.08, R^2 (hard) = 0.98, *RMSD* (hard) = 0.04.

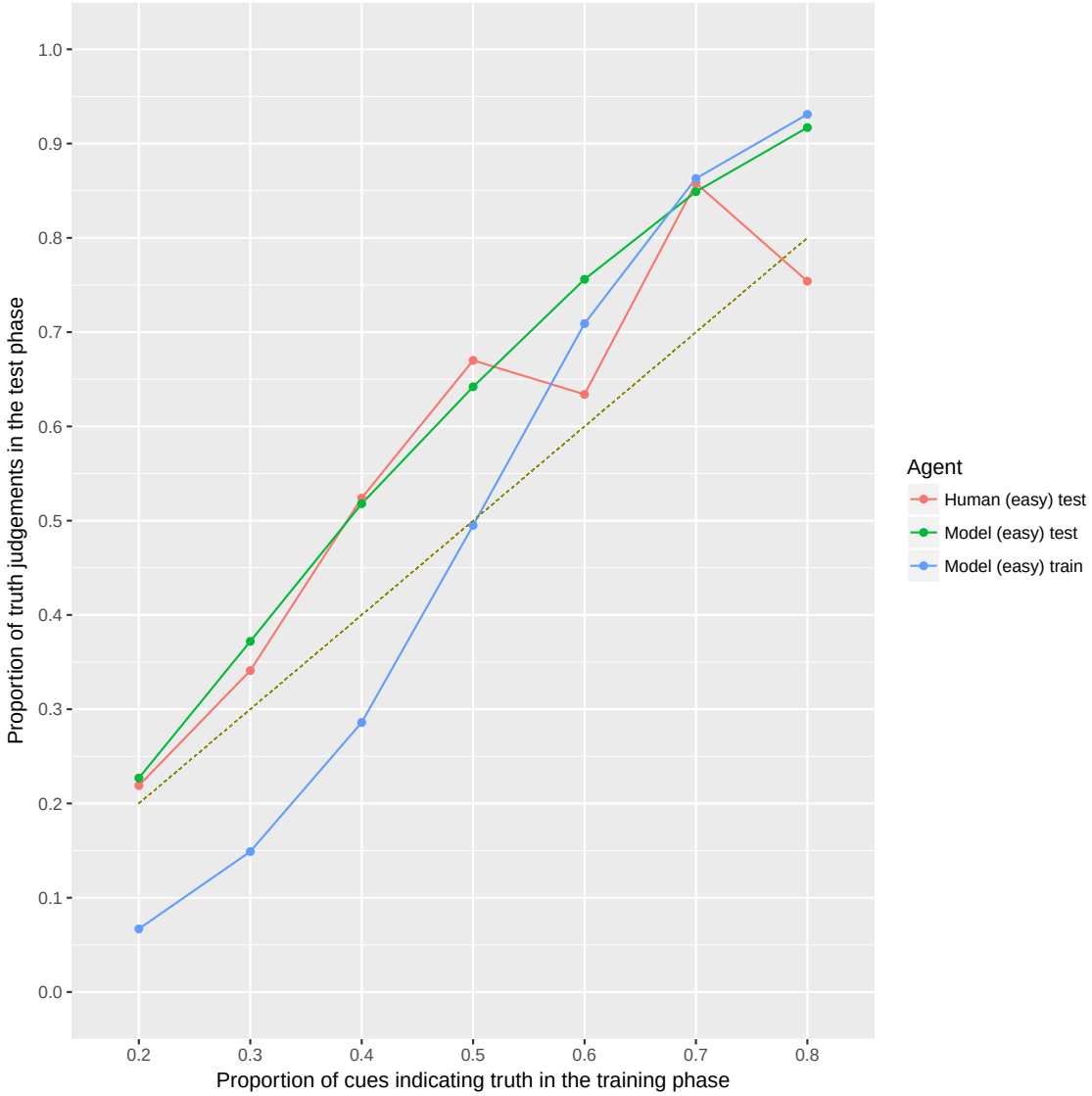

Figure 5. Proportion of test phase truth judgements for each proportion of truth cue conditions in the training phase, human and model, "easy" condition.

2.2 Explaining the model's learning during the training phase

To recap, the model's performance depends primarily on the chunks in declarative memory that are created during the training phase and which represent the *instances* or experiences of the four cues and their association with the truth and lie responses. As the

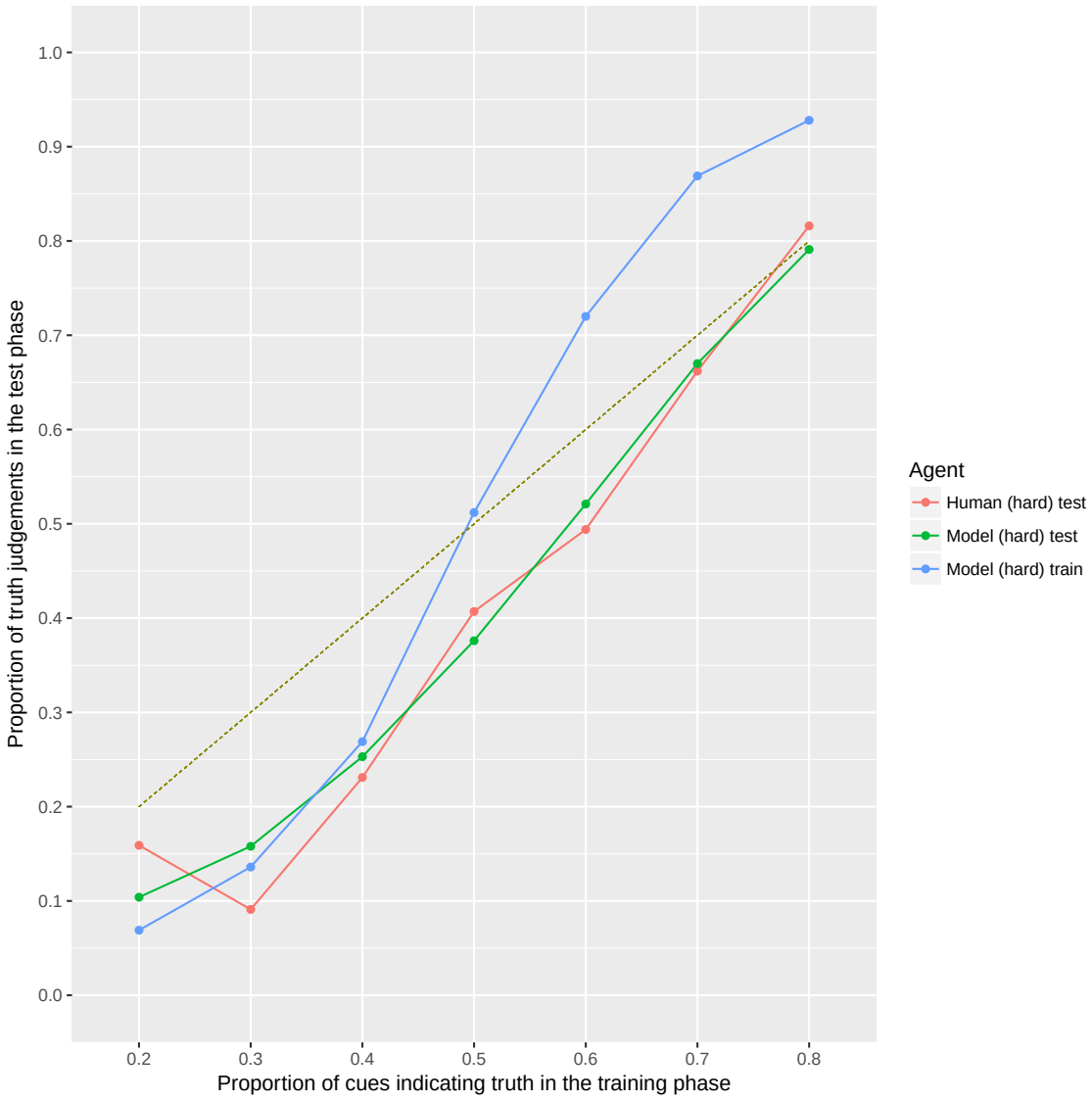

Figure 6. Proportion of test phase truth judgements for each proportion of truth cue conditions in the training phase, human and model, "hard" condition.

model proceeds through the training phase, eight chunks are created—two for each cue—that represent (in the form of activation) the model's current beliefs regarding strength of association between each cue and the truth and lie responses. These activations are updated through base-level learning on each trial and the activation determines the likelihood of truth and lie response retrievals at each new cue presentation in subsequent trials.

Figures 5 and 6 reveal that, apart from when the cue was perfectly non-diagnostic (i.e., when the cue was equally associated with telling the truth and lying during the learning phase) the model did not match the experimental truth proportions exactly but

systematically over-estimated (in the case of a truth-diagnostic cue > 0.5) or under-estimated (in the case of a lie-diagnostic cue < 0.5) the proportion of truthful statements as the s the cue diagnosticity increased (i.e., as the experimental proportion of truth cues moved further away from 0.5 in either direction). This discrepancy is approximately equal on either side of the 0.5 level and reaches a maximum at the proportions 0.3 and 0.7 indicative of honesty.

The sigmoid shape of the curve is due to a number of nonlinearities in the system, for example in the choice probability function (Equation 3) and the differences in activation between the competing chunks. The profile of this curve is affected by the retrieval threshold and activation noise parameters and explaining how so may clarify the mechanisms somewhat.

For example, the optimal retrieval threshold parameter for the model was found to be 0.4. Gradually increasing this parameter flattens the curve up to a value of 1.32 where the model's responses closely match the experiment proportions. Why does this occur? Remember that on each block of 40 trials the model is presented with the same cue. When the proportion of truth trials is very low (e.g., 0.2) the lie response chunk will be relatively active compared to the truth response chunk but a high threshold will result in more trials where no chunk is active enough to be retrieved so more guessing will occur. Increasing the threshold above 1.32 continues the trend until the value of 2.0 where no chunks are ever retrieved. Guessing then occurs for each level of cue diagnosticity (0.2 to 0.8). Put simply, setting the threshold too high results in individuating cues not being retrieved and so the model selects a response at random.

The optimal activation noise parameter for the model was found to be 0.21. Increasing this value to 0.53 flattens the curve to obtain close probability matching whereas decreasing the value accentuates the steepness of the curve. To see why this occurs, consider again the case where the proportion of truth trials is very low (e.g., 0.2). As activation noise is increased, the chance that the relatively low activation of the truth response chunk will be boosted by the noise to be higher than both the lie chunk and the threshold and so the chance of being retrieved will increase. Increasing the noise further continues the trend until

eventually the probability of both chunks being retrieved is dominated by noise and again the model's proportion of truth responses are approximately 0.5 for each proportion condition.

Conversely, decreasing the noise results in the retrieval probability being dominated by the chunk's activation. That is, when too much noise is added to the retrieval process, the individuating cue is equally likely to evoke a lie as a truth response. With too little noise, responses will become deterministic: a lie or truth response will be made by the model far more often than humans make those responses because the individuating cue will be retrieved with a very high probability.

Street et al. (2016) did not record the participants' responses during the training phase of their experiment so it is not possible to determine to what extent they learned the various proportions before being given the context general information at the start of the test phase. However a second experiment has since been conducted by Street which has a different test phase but a very similar training phase. As with the earlier experiment, in this new study, 81 participants were required to learn the extent to which each of four behavioural cues indicated whether the person was lying or telling the truth (i.e., how diagnostic of the two behaviours each cue was). The procedure of the new study was identical to that of the original experiment apart from the diagnosticity of the cues; in the new experiment the proportions were 20/80 and 35/65.

To compare the learning of the participants in the second experiment with that of the model of the first experiment, the proportion of truth judgements on the last trial of the learning phase for both are plotted in Figure 7. The data from the new experiment show that people are—as the model predicts—maximising their responses during the learning phase. This result is in accord with previous evidence that people do not simply match environmental probabilities during probability learning tasks when provided with feedback (e.g., Barron & Erev, 2003; Shanks, Tunney, & McCarthy, 2002).

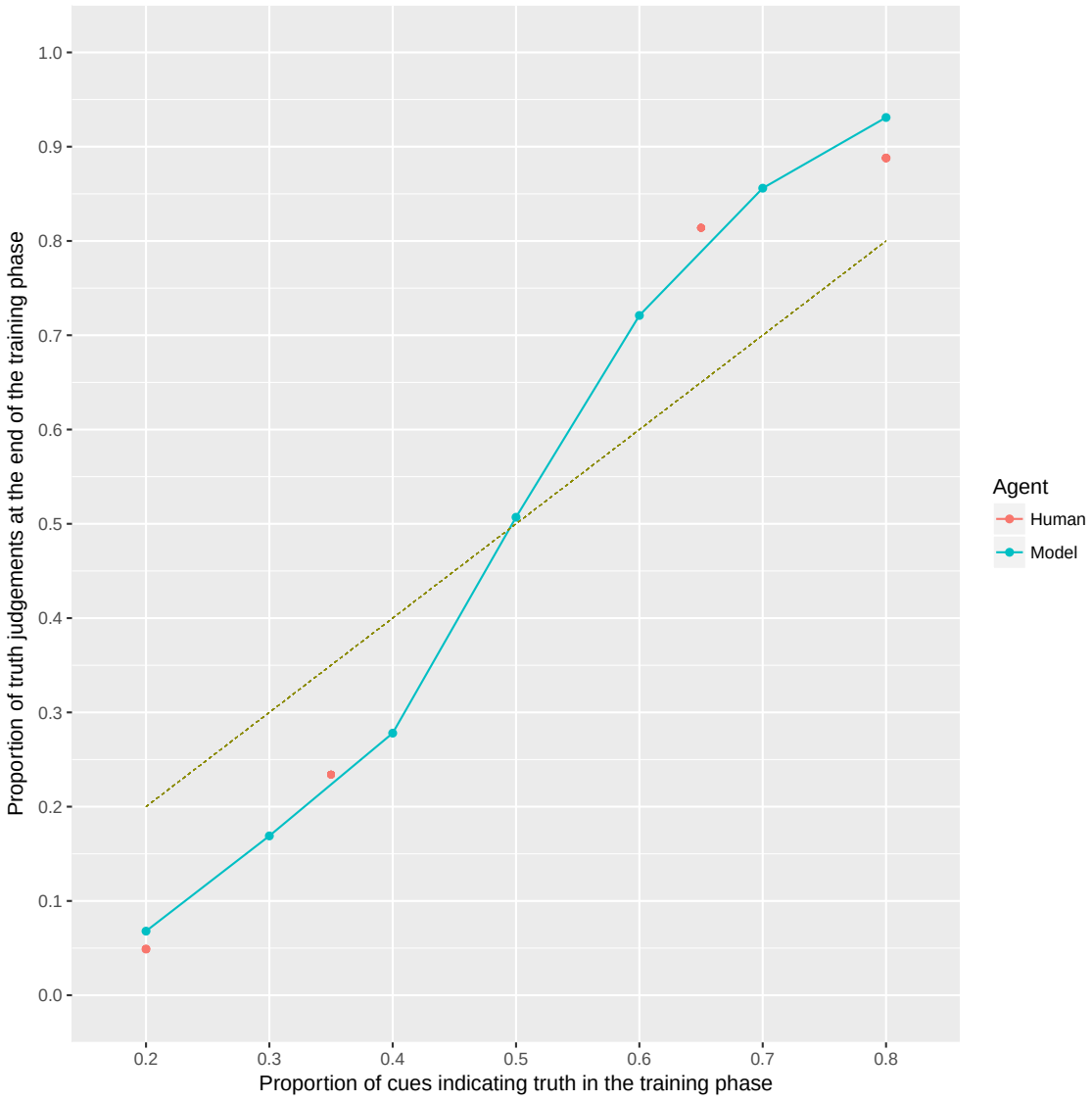

Figure 7. Proportion of truth judgements for proportions of truth cue conditions on the last trial of the training phase, humans (Experiment 2) and model (Experiment 1).

2.3 Explaining the model's performance during the test phase

After the training phase the model is provided with the experiment condition information, "easy" or "hard", and before starting the test trials the model retrieves from memory the context-general response bias associated with each ("truth" or "lie" respectively[3]). Once retrieved, this response bias then becomes an element of the goal which is used as an additional probe in subsequent memory retrieval requests for individuating

[3] Note that it is the context-general information that is being retrieved from memory at this stage. This should not be confused with retrieving an individuating cue response.

cues.

The effect of this additional probe is to change the dynamics of the retrieval process and this is where the partial matching mechanism plays a crucial role. To simplify the model, partial matching was nullified in the training phase by setting the mismatch penalty (P in Equation 5) to 0. This had the effect of eliminating the possibility of retrieval errors in the training phase. Partial matching was allowed to operate in the test phase however which ensures that all eight cue chunks in declarative memory (the lie and truth response associated to each of the four cues) are entered into the retrieval process.

Each chunk's mismatch score is computed according to its similarity to the probe on the two elements (cue name and context-general response bias). This mismatch score is then used to reduce the activation of chunks according to their degree of dissimilarity. The outcome of this process is that the chunk with the highest activation is still the one retrieved, but that chunk may not be an exact match to the elements specified in the retrieval request. It also biases the retrievals of individuating cues in favour of those chunks containing the same response as that associated with the context-general condition, increasing and decreasing the probability of a "truth" response in the easy and hard conditions respectively. The additional piece of context-general knowledge informing memory retrieval in the test phase aptly represents the bias towards truth responses in the easy condition and towards lie responses in the hard game condition.

Figures 5 and 6 reveal that the effect of the additional context-general information differs across the individuating cue diagnosticity levels, with the context-general information having an increasingly greater effect as the proportion of truth cues (i.e., the proportion of times the cue was associated with honesty) decreases in the easy condition and increases in the hard condition (as shown in Figure 2 the curves are very similar). That is, the difference between the model training phase (blue) and model test phase (green) becomes greater. To explain why this happens, consider again the easy condition shown in Figure 5 where the context-general information suggests that most people will tell the truth. When the cue is present on only 20% of the occasions where speakers tell the truth, only eight of the 40 cues in a block indicate truth telling whereas 32 experiences of the cue indicate lying.

During the training phase this results in the lying chunk being highly active compared to the truth chunk and consequently being retrieved approximately 93% of the time.

During the test phase, the extra "truth" context-general cue is included in the retrieval request and consequently the partial matching mechanism reduces the activation of the "lie" individuating cue chunks, which increases the chance of "truth" chunks being retrieved. As the proportion of trials where the individuating cue indicates honesty gradually increases to 0.8 however, the effect of the additional context-general "truth" knowledge decreases because the activation of the "truth" individuating cue chunk in declarative memory is already increasing relative to the "lie" chunk. As such, the effect of reducing the activation of the "lie" individuating cue chunk through partial matching diminishes as the proportion of "truth" trials increases.

This explanation works in reverse for the hard condition shown in Figure 6. This time, consider the 0.8 proportion condition where only eight of the 40 cues in a block indicate lying and 32 of the cues indicate truth telling. As with the easy condition, the training phase results in the "truth" chunk being highly active compared to the "lie" chunk and consequently being retrieved approximately 93% of the time. For the hard condition however, during the test phase, the extra "lie" context-general cue in the retrieval request reduces the activation of the "truth" individuating cue chunk, decreasing the chance of its retrieval. As the proportion of "truth" trials gradually decreases to 0.2, the effect of the additional lie context-general cue decreases because the activation of the "truth" individuating cue chunk in declarative memory is low relative to the lying chunk.

Reducing the activation of the "truth" chunk through partial matching has a diminishing effect on the probability of "truth" responses as the proportion of "truth" trials decreases. This is in fact the exact opposite of what happens in the easy condition and accounts for the similarity between the easy and hard conditions in the human data revealed in Figures 1 and 2. That is, the process underlying the response in both the easy and hard context-general conditions can be explained as relying on the same process of integrating context-general information with individuating cues.

One feature of the human data not matched by the model is the drop in the PTJ when

a highly diagnostic truth cue (0.8) is presented in the easy condition, and the increase in the PTJ when a highly diagnostic lie cue (0.2) is presented in the hard condition. In fact, the human data show that for highly diagnostic individuating cues (0.8 and 0.2), experiment participants produced responses very close to the actual proportions, regardless of the context-general condition.

This pattern is consistent with ALIED's position that context-general information has relatively little effect when highly diagnostic cues are present, although ALIED does not suggest a mechanism for this effect. One plausible explanation is that it reflects findings that people do not usually treat probabilities linearly but are increasingly sensitive to changes in probability as they move towards the two endpoints zero and one (R. Gonzalez & Wu, 1999; Tversky & Kahneman, 1992). In the case of current experiment, this manifests as a floor/ceiling effect that occurs when the differences in proportions becomes sufficiently large. Perhaps when environmental proportions reach a certain critical size, people are unwilling to respond beyond a certain proportion (i.e., towards 100% as they know that was not the case during training) and become more sensitive to the smaller proportion so that the effect of the additional context information becomes negligible—or even reverses by making people more aware of the actual environmental proportions.

More broadly speaking, perhaps the context-general effect that serves to bias responses has less of an effect (or maybe reverses) when individuating cues have high diagnosticity. In terms of the ACT-R model, if an individuating cue had a diagnosticity of 0.95 (i.e., very highly predictive of honesty), the activation of that chunk would typically be so high that only a very high mismatch penalty and/or noise in the retrieval process would be sufficient to prevent the individuating cue chunk from being retrieved accurately. As such, the individuating cue chunk would ultimately drive the decision.

3. Discussion

The central claim of the ALIED theory of lie detection is that when making judgements, people combine knowledge from two sources: specific *individuating* information relating to the current situation and relevant *context-general* information which

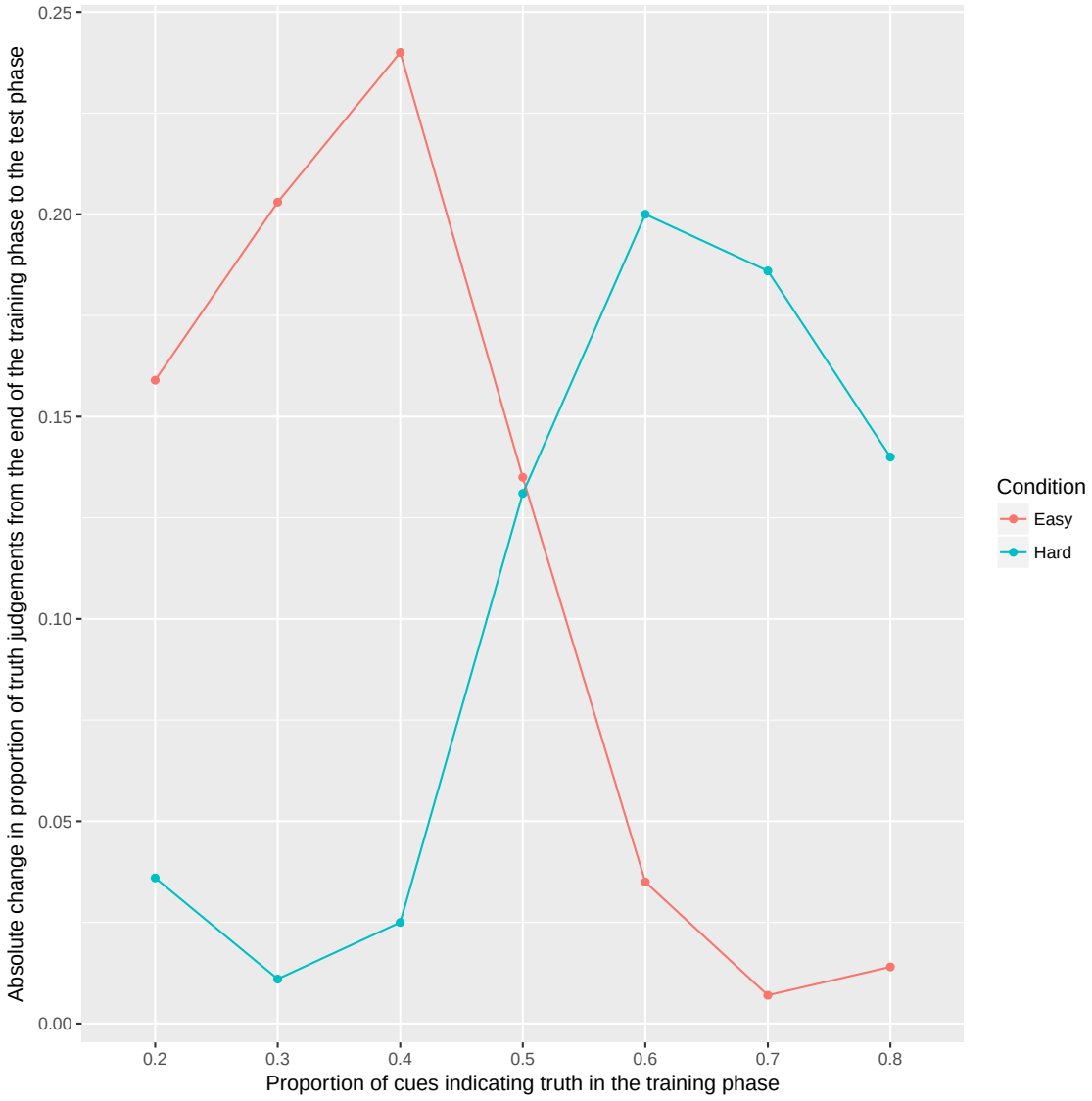

Figure 8. Absolute change in the model's proportion of truth judgements from the end of the training phase to the test phase.

constitutes the base rate or likelihood of the situation. According to ALIED, the relative influence of each of these knowledge sources depends upon the degree of diagnosticity of the individuating cues, with context-general knowledge becoming more influential as individuating cue diagnosticity decreases.

The computational model presented here provides a mechanistic account of how the integration of individuating and context-general information can occur in terms of the storage and subsequent retrieval of *instances* of experience. In so doing, the model brings to bear the many theoretical constraints of the ACT-R cognitive architecture and the

assumptions of instance-based learning theory to provide a close fit to the human data reported by Street et al. (2016) and a precise formal explanation of the phenomena described by ALIED. As such it is an illustrative example of how formal modelling can provide a rigorous and cognitively grounded explanation of a relatively loosely specified verbal theory (see Lewandowsky & Farrell, 2010). In addition, the ACT-R model accounts for all aspects of human behaviour by carrying out a precise simulation of the experimental task, including vision, attention, memory, goal structure and motor control. This is crucial as the timing of the experiment trials influences the activation of memory chunks and, in turn, recall probabilities.

It is important to note that the model did not just fit existing data but also predicted that participants would over- and under-estimate the proportion of truth cues during the training phase (data which were not collected in the original experiment). Data from a subsequent similar experiment were remarkably close to this prediction, again providing strong support for the assumptions of the model.

A key issue raised by the model is the order in which processes operate. While this may sound like a minor point, it is fundamental to discussions around whether or not people default to believing information. It is not clear whether people begin with a presumption that others tell the truth, for instance, and then look for individuating cues to confirm that presumption, or whether people begin in an unbiased way and search for the most decisive information, making use of both context-general and individuating information to inform that judgement. That is, it is not clear whether people post-hoc select the individuating cues that are consistent with their presumption, or whether they have no presumption and instead engage in a search that ultimately leads to a biased response.

The current model takes the former position: It assumes context-general information will always bias the search of memory for individuating cues. Evidently, this approach explains the past data well and also makes a prediction which was met by new data. This should be seen as a challenge to ALIED theory insofar as the theory is not sufficiently precise to be able to commit to either position. The ACT-R model demonstrates that ALIED theory should commit to a position where context-general information will always bias the

information search.

This may appear to suggest that there is a cognitive default that people always engage with—i.e., always use the context. And in one sense, we would agree. But the default here is not a default in a particular belief—lie or truth. Instead, it is a default of always taking into account the current context. If the context were to change from one trial to the next, so too would the observed bias in participants' responses (see Figures 5 and 6). As communicated in the original theory article, "the bias must reflect the current context if it is to be adaptive, in the sense of both functional and flexible" (Street, 2015, p. 337). The predictive success and fit of the model, we believe, support this position.

This discussion dovetails with the issue of compensatory and non-compensatory decision strategies. The ACT-R model relies on IBLT, which has been successful in predicting a wide range of cognitive functions. Interestingly, this approach means that the context-general information is always used in the judgement process. ALIED theory offers two possible ways in which context-general information may be used (Street et al., 2016). First, it may be used in a non-compensatory fashion whereby people either use context-general information or individuating information, but not both. Second, a compensatory approach may be taken whereby people use both individuating and context-general information but trade off the degree of influence on the final judgement depending on how diagnostic the individuating cue is. Street et al. (2016) sought to distinguish these two accounts with the use of optimisation models, but they were unable to distinguish them. The ACT-R model takes a compensatory approach, applying a penalty based on the mismatch between the to-be-retrieved individuating cue and the context-general information.

The model was also able to demonstrate a high degree of fit to the data by assuming that a lie bias and a truth bias arise from the same processing operations. That is, it is not necessary to consider the truth bias as arising from a default form of processing and a lie bias as something that may result from an additional trigger or further processing (cf. Gilbert, Krull, & Malone, 1990; Levine, 2014). A key claim of ALIED theory is that the truth and lie biases can be explained as a behavioural response resulting from the same

underlying processing mechanisms. The cognitive model developed here that was both successful in fitting past data and in having its predictions supported with new data lends credibility to this claim.

The model supports the claim that changing people's expectations regarding the likelihood of a particular behaviour (i.e., context-general information) only changes the response for that behaviour relative to their current beliefs about the prevalence of the behaviour. This is illustrated in Figure 8 which plots the absolute change in the model's proportion of truth judgements from the end of the training phase to the test phase for each proportion of truth cue conditions in the training phase. If the initial probability of the behaviour is low then the effect of the context-general information will be greater than if the initial probability of the behaviour is already high (consistent with ALIED theory).

According to the model, if you have learned that a cue is strongly diagnostic of a particular behaviour, then receiving information that attenuates the association will have a great effect on your subsequent judgements (particularly if you have maximised your learning). In contrast, if you have learned that a cue is strongly diagnostic of a particular behaviour and you receive information that further supports that knowledge, then it will have only a very small effect on your subsequent judgements because the association is already very strong. For example, if you have been trained to associate a particular cue with truth telling only 30% of the time then subsequently hearing that the game was hard and that therefore people are more likely to lie is going to have a smaller effect on you than when you have seen the cue associated with truth telling 70% of the time. When a cue is diagnostic of a particular behaviour on only 50% of your experiences of it however, then subsequent information either reinforcing or undermining that belief will shift your opinion in the appropriate direction to the same extent (see Figure 8).

In its demonstration of the varying effect of new information on learned response patterns, the model also represents a novel development of the IBLT approach to explaining dynamic decision making and decisions from experience (C. Gonzalez, 2017). It also presents a credible example of how an approach based on a theory of human cognitive architecture and basic learning mechanisms can account for decision making behaviour and predict

learning patterns which will provide a starting point for future investigations.

Acknowledgements

The authors would like to thank Dan Bothell for his invaluable advice in developing the model.

References

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* New York, NY: Oxford University Press.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*(6), 396–408.

Barron, G. & Erev, I. (2003). Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, *16*(3), 215–233.

Blair, J. P., Levine, T. R., & Shaw, A. S. (2010). Content in context improves deception detection accuracy. *Human Communication Research*, *36*(3), 423–442.

Bond, C. F. & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and social psychology Review*, *10*(3), 214–234.

Bond, C. F., Howard, A. R., Hutchison, J. L., & Masip, J. (2013). Overlooking the obvious: Incentives to lie. *Basic and Applied Social Psychology*, *35*(2), 212–221.

Carr, A. Z. (1968). Is business bluffing ethical? *Harvard Business Review*, *46*(1), 143–153.

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground at the understanding of demonstrative reference. *Journal of verbal learning and verbal behavior*, *22*(2), 245–258.

Cooper, R. P. & Peebles, D. (2015). Beyond single-level accounts: The role of cognitive architectures in cognitive scientific explanation. *Topics in Cognitive Science*, *7*(2), 243–258.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of personality and social psychology*, *70*(5), 979–995.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological bulletin*, *129*(1), 74.

DePaulo, P. J. & DePaulo, B. M. (1989). Can deception by salespersons and customers be detected through nonverbal behavioral cues? *Journal of Applied Social Psychology*, *19*(18), 1552–1577.

Dutt, V., Ahn, Y.-S., & Gonzalez, C. (2013). Cyber situation awareness: Modeling detection of cyber attacks with instance-based learning theory. *Human Factors*, *55*(3), 605–618.

Gilbert, D. T., Krull, D. S., & Malone, P. S. (1990). Unbelieving the unbelievable: Some problems in the rejection of false information. *Journal of personality and social psychology*, *59*(4), 601–613.

Gilboa, I. & Schmeidler, D. (1995). Case-based decision theory. *The Quarterly Journal of Economics*, *110*(3), 605–639.

Gonzalez, C. (2017). Decision making: A cognitive science perspective. In S. E. F. Chipman (Ed.), *The oxford handbook of cognitive science*. New York, USA: Oxford University Press.

Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of ACT-R. *Journal of Computational Science*, *4*(4), 262–268.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science*, *27*, 591–635.

Gonzalez, C. & Wimisberg, J. (2007). Situation awareness in dynamic decision making: Effects of practice and working memory. *Journal of Cognitive Engineering and Decision Making*, *1*(1), 56–74.

Gonzalez, R. & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166.

Halevy, R., Shalvi, S., & Verschuere, B. (2014). Being honest about dishonesty: Correlating self-reports and actual lying. *Human Communication Research*, *40*(1), 54–72.

Hartwig, M. & Bond, C. F. (2011). Why do lie-catchers fail? a lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*, 643–59.

Kenny, D. A. & DePaulo, B. M. (1993). Do people know how others view them? an empirical and theoretical account. *Psychological Bulletin*, *114*(1), 145–161.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17.

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making*, *25*(2), 143–153.

Levine, T. R. (2010). A few transparent liars. In C. Salmon (Ed.), *Communication yearbook* (Vol. 34, pp. 40–61). CA: Sage.

Levine, T. R. (2014). Truth-default theory (TDT): A theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378–392.

Levine, T. R. & McCornack, S. A. (2014). Theorizing about deception. *Journal of Language and Social Psychology*, *33*(4), 431–440.

Lewandowsky, S. & Farrell, S. (2010). *Computational modeling in cognition: Principles and practice*. Sage.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review*, *109*(2), 376–400.

Masip, J., Alonso, H., Garrido, E., & Herrero, C. (2009). Training to detect what? the biasing effects of training on veracity judgments. *Applied Cognitive Psychology*, *23*(9), 1282–1296.

Masip, J. & Herrero, C. (2017). Examining police officers' response bias in judging veracity. *Psicothema*, *29*(4).

Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Meissner, C. A. & Kassin, S. M. (2002). "he's guilty!": Investigator bias in judgments of truth and deception. *Law and human behavior*, *26*(5), 469.

Millar, M. G. & Millar, K. U. (1997). The effects of cognitive capacity and suspicion on truth bias. *Communication Research*, *24*(5), 556–570.

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 104–114.

Pavlik, P. I. & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*(4), 559–586.

Serota, K. B., Levine, T. R., & Boster, F. J. (2010). The prevalence of lying in america: Three studies of self-reported lies. *Human Communication Research*, *36*(1), 2–25.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making*, *15*(3), 233–250.

Sporer, S. L. & Schwandt, B. (2006). Paraverbal indicators of deception: A meta-analytic synthesis. *Applied Cognitive Psychology*, *20*(4), 421–446.

Sporer, S. L. & Schwandt, B. (2007). Moderators of nonverbal indicators of deception: A meta-analytic synthesis. *Psychology, Public Policy, and Law*, *13*, 1–273.

Street, C. N. H. (2015). ALIED: Humans as adaptive lie detectors. *Journal of Applied Research in Memory and Cognition*, *4*(4), 335–343.

Street, C. N. H., Bischof, W. F., Vadillo, M. A., & Kingstone, A. (2016). Inferring others' hidden thoughts: Smart guesses in a low diagnostic world. *Journal of Behavioral Decision Making*, *29*(5), 539–549.

Thomson, R., Lebiere, C., Anderson, J. R., & Staszewski, J. (2015). A general instance-based learning framework for studying intuitive decision-making in a cognitive architecture. *Journal of Applied Research in Memory and Cognition*, *4*(3), 180–190.

Tomasello, M. (1995). Joint attention as social co0gnition. In C. Moore & P. Dunham (Eds.), *Joint attention: Its origins and role in development* (pp. 103–130). New Jersey: Lawrence Erlbaum.

Tversky, A. & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.