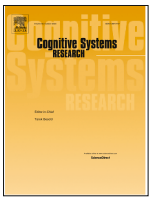
Cognitive Systems Research 96 (2026) 101434

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys



Developing a cognitively plausible model of lie-truth judgements: An adaptive lie detector account

David Peebles [a,*], Chris N.H. Street [b]

[a] Department of Social and Psychological Sciences, University of Huddersfield, Queensgate, Huddersfield, HD1 3DH, West Yorkshire, UK
[b] School of Psychology, Keele University, Keele, ST5 5BG, Staffordshire, UK

ARTICLE INFO

Keywords:
ALIED
Lie detection
ACT-R

ABSTRACT

To date, no account of lie-truth judgement formation has been capable of explaining how core cognitive mechanisms such as memory encoding and retrieval are employed to reach such judgements. One theory, the Adaptive Lie Detector (ALIED: Street et al., 2016) is sufficiently well defined as to be implemented as a cognitively plausible computational model. Here we describe the first cognitively plausible account of lie-truth judgments and test it in three studies. Developed within the ACT-R cognitive theory (Anderson, 2007), the model provides a close fit to past data (Study 1), predicts novel learning data (Study 2), and generalises to more complex environments with multiple cues (Study 3). In so doing, the cognitive model provides strong support for the assumptions of an adaptive theory of lie-truth judgement, and substantiates a unique prediction of ALIED that lie and truth biases can be considered as functionally equivalent.

1. Introduction

The Adaptive Lie Detector account (ALIED: Street et al., 2016) argues that people make reasonable use of the information available to them to reach a lie or truth judgement. While it explains lie-truth judgments and makes novel predictions (see Street, 2015), it has a substantial drawback: it does not explain what cognitive *mechanisms* are involved in the judgement formation process. No explanation of lie-truth judgement formation to date has done so, and yet invariably core cognitive mechanisms such as perception and memory are fundamental to the judgement process. In this paper, we develop a cognitively plausible computational model of lie-truth judgements based on ALIED that can account for data from a previous experiment (Study 1). The model creates a prediction which is tested with novel data in a second experiment (Study 2). Finally, we generalise the model to account for a more complex experiment where multiple cues are present (Study 3).

1.1. Individuating and context-general cues in lie detection

According to ALIED (Street, 2015), people's judgements are informed by two types of knowledge: information relating to the particular statement under consideration (called individuating information) and information that generalises across statements, which can be informed by the base rate of honesty (called context-general information). ALIED argues that when individuating information has high diagnosticity in determining if the statement is a lie or truth (e.g., Pinocchio's nose growing), this has the heavier weight in the final judgement outcome. But when individuating information has low diagnosticity (i.e., it gives little to no reliable information about veracity), ALIED argues that people do not simply guess at random, but rather that their knowledge of context-general information (e.g., "most people tell the truth in this setting") has the heavier weight in the final judgement outcome. This allows for a satisficing judgement in the absence of more diagnostic information about the specific statement being considered. The account is considered adaptive insofar as it reflects an informed decision-making process.

For example, Coby may have seen CCTV footage of Alesha cheating on a test. When Alesha claims that she did not cheat, Coby has highly diagnostic individuating information available to him (the CCTV footage that contradicts Alesha's claim). This information individuates this statement, and as such is called individuating information or an individuating cue (see Koehler, 1996). Individuating information varies in its degree of diagnosticity. If a particular cue has been learnt by the individual to be highly diagnostic (such as Coby's CCTV footage), people can use this information to attain high accuracy (Blair et al., 2010; Bond et al., 2013; Levine & McCornack, 2014). But individuating cues to deception are typically either unavailable (see Luke, 2019) or if they are available, have low diagnosticity (DePaulo et al., 2003; Hartwig & Bond, 2011; Sporer & Schwandt, 2006, 2007).

* Corresponding author.
E-mail addresses: d.peebles@hud.ac.uk (D. Peebles), c.street@keele.ac.uk (C.N.H. Street).

https://doi.org/10.1016/j.cogsys.2026.101434
Received 28 March 2024; Received in revised form 11 July 2025; Accepted 5 January 2026
Available online 20 January 2026

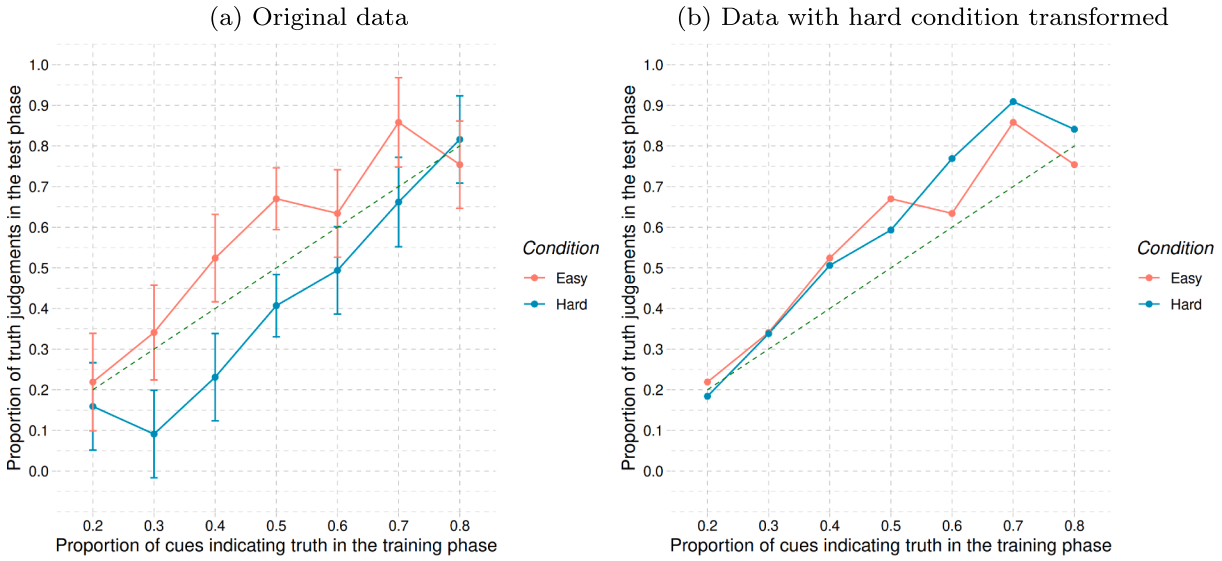


Fig. 1. Proportion of truth judgements on trials for each diagnosticity level of truth cue in the training phase, split by easy and hard game conditions, from the experiment reported in Street et al. (2016) (a). The dotted line indicates where the proportion of judgements matches the frequency of the cue being associated with honesty. Error bars denote 95 % confidence intervals. (b) Data with the hard game condition data transformed (rotated 180 degrees) and superimposed on the easy condition data.

When individuating information has low diagnosticity, Coby could guess randomly. But Coby 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 Alesha's specific statement about not cheating, but rather generalises across statements in this situation (i.e., statements in the context of test cheating). ALIED refers to this as context-general knowledge. When judging whether Alesha is lying about not having cheated, Coby will combine this context-general knowledge with their consideration of specific individuating information. As the learnt diagnosticity of the individuating cues is perceived to reduce, ALIED argues that context-general knowledge will have a heavier weighting in the judgement process in order to make an informed—but overgeneralised—judgement.

Because individuating cues typically have low diagnosticity (DePaulo et al., 2003), and because the context-general information is informed by the learnt experience that people predominantly tell the truth (DePaulo et al., 1996; Halevy et al., 2014; Serota et al., 2010), in most situations it is rational to make a truth judgement. This account explains the frequently observed bias to judge others' statements as truths (known as the *truth bias* Bond & DePaulo, 2006) as resulting from an adaptive decision-making process rather than as an error or default to believe.

In situations where lying is more prevalent or where there is a widespread belief that people lie (that is, context-general information suggests that people will typically lie), ALIED claims that the bias is to judge statements as being untruthful, for which there is evidence (e.g., Bond et al., 2005; Carr, 1968; DePaulo & DePaulo, 1989; Hartwig et al., 2004; Masip et al., 2009; Masip & Herrero, 2017; Millar & Millar, 1997). The observed truth bias found in typical studies is not a cognitive disposition therefore, but rather an adaptive judgement in the absence of reliable specific (individuating) information, according to ALIED (see Levine & Street, 2024).

1.2. A test of ALIED

The interactive effect of individuating and context-general cues on lie detection has been investigated empirically (Street et al., 2016). They demonstrated that people's judgements are a function of both (a) the diagnosticity of individuating cues and (b) 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.

Participants in this study were to judge whether individuals who had recently played a trivia game had cheated (and lied about it) or not (and told the truth about not cheating). This was a cover story: In reality, no trivia game had taken place. The study took the form of a two-choice reinforcement learning paradigm widely used to investigate probability and category learning (Estes, 1972; Medin & Schaffer, 1978; Montag, 2021; Nosofsky, 1986), consisting of a training phase followed by a test phase. This experiment differed from typical designs by introducing additional context-general information prior to the test phase to measure its effect on learned associations.

During the learning phase, participants were presented with a cue that appeared somewhere between 20 % and 80 % (in steps of 10 %) of the time when the trivia players were being honest, which defines each cue's level of diagnosticity (i.e., between 20 % and 80 % diagnostic of honesty). Participants had to judge whether the trivia game player lied or told the truth, and they received feedback on their responses. They learnt about four individuating cues in this way. The cues in question were listed as being present, and were (i) high pitch of voice, (ii) appearing facially expressive, (iii) a large number of silent periods in the middle of sentences, and (iv) a large number of self-references such as 'I' and 'me'. The diagnosticity of each cue was randomised between participants, meaning that a cue which was 60 % diagnostic of honesty for one participant, for example, could be 80 % indicative of deception for another participant.

At the end of the learning phase, participants were given context-general information: either that most trivia game players would lie (because the trivia game is hard and so most people had to cheat and then lie about it) or that most would tell the truth (because the trivia game was easy and so did not need to cheat and thus could tell the truth about not cheating). In a final test phase, participants made lie-truth responses with a single cue presented, as per the learning phase, but were given no feedback on the accuracy of their responses.

The key dependent measure was the proportion of truth judgements in the test phase (PTJ) for each cue diagnosticity. These are shown for both the easy and hard conditions in Fig. 1. The difference between the easy and hard conditions can be seen in Fig. 1(a). 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

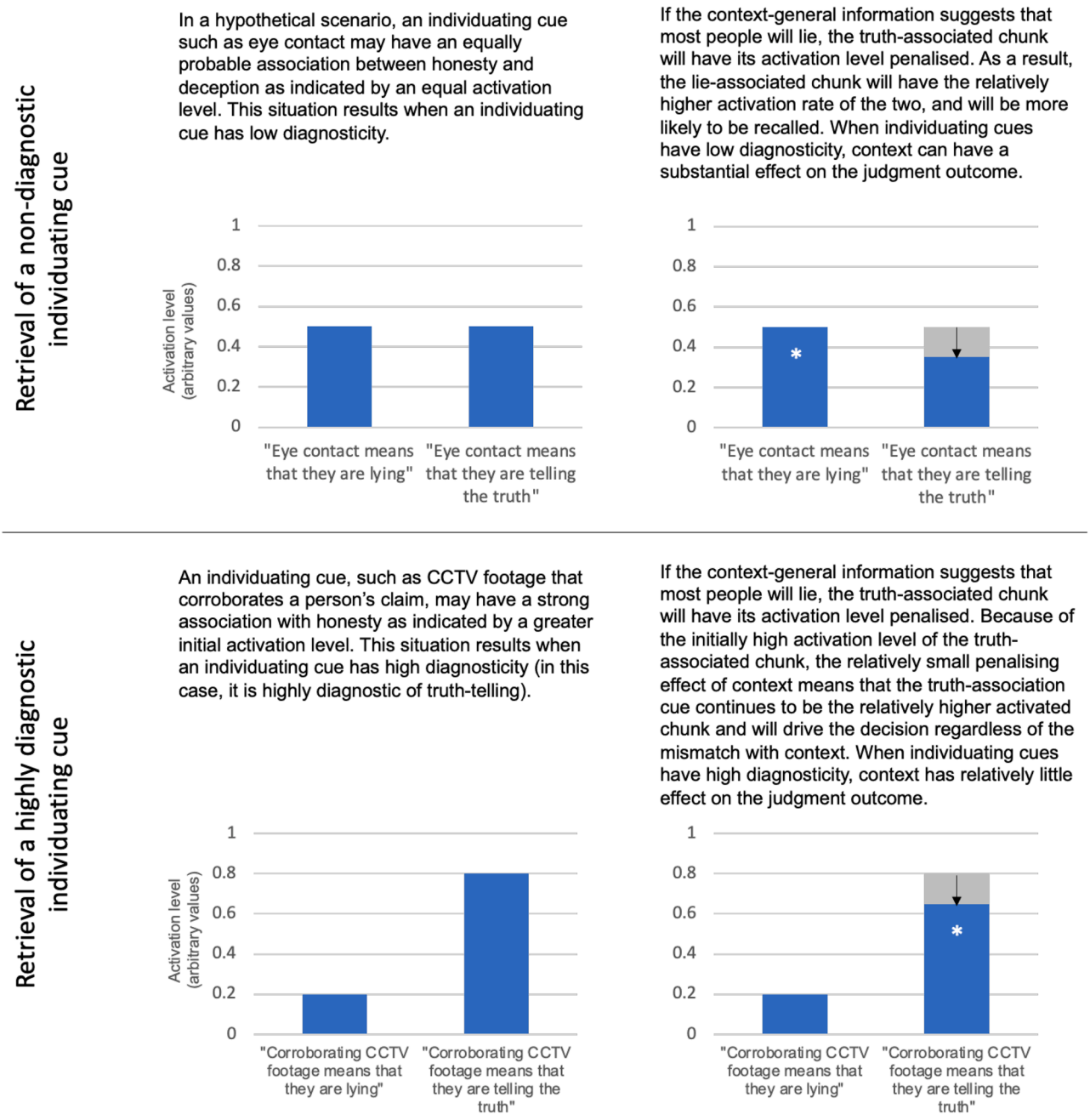


Fig. 2. A conceptual overview of the effect of context-general information on the activation levels of individuating cues in declarative memory.

line and the 0.5 points (as shown in Fig. 1(b)). The correlation and mean deviation between the easy and transformed hard data are 0.97 and 0.06 respectively. An explanation for this similarity is provided below.

Analysis of the data revealed a significant main effect of cue diagnosticity, showing that as cues became more indicative 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 context had a greater degree of impact on the final judgement as the cue diagnosticity reduced.

The experiment provided a 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. The judgements were capable of being modelled as a Bayesian reasoner (Street et al., 2016). Questions remain however concerning how the 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 instantiated in the ACT-R theory of human cognitive architecture (Anderson, 2007).

ACT-R is a theory of the core components of the human cognitive system (including perceptual, cognitive, and motor processes) derived from decades of experimental data and theoretical development. It explains how these components work together to produce intelligent behaviour.

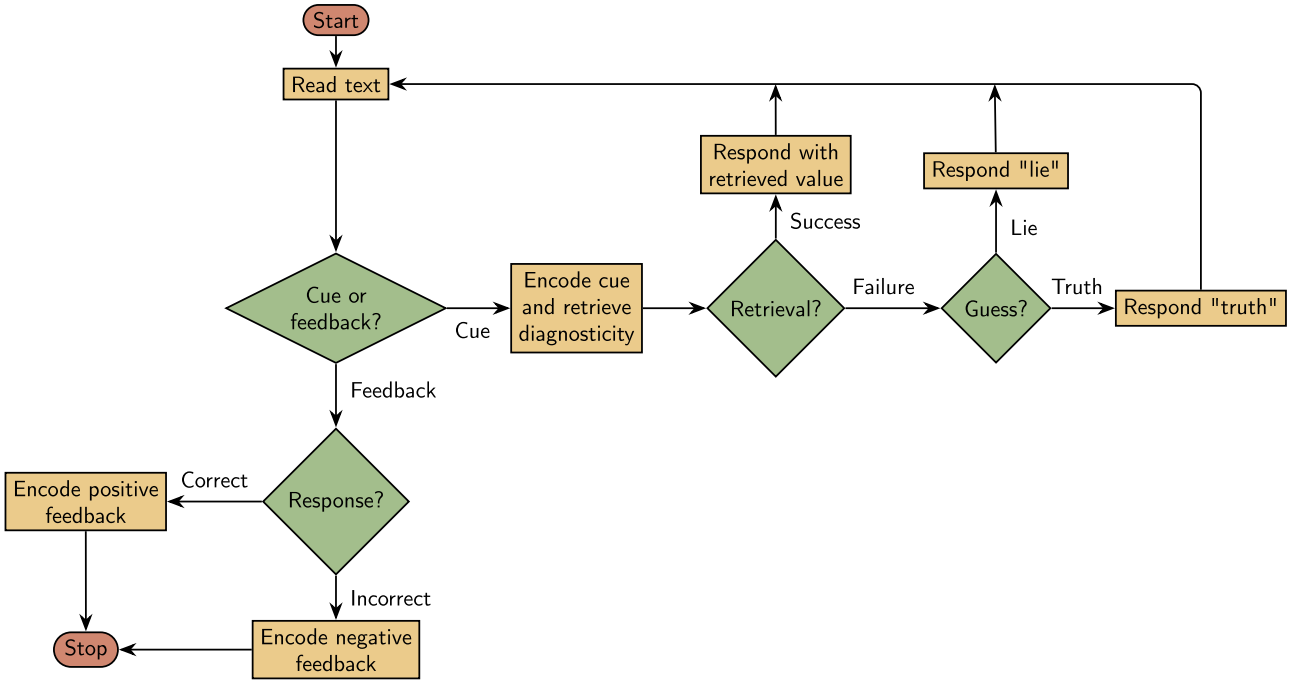


Fig. 3. Control flow of the ACT-R model carrying out a trial of the experiment. Rectangles correspond to actions carried out by production rules, diamonds represent junctions in the experiment or model's behaviour.

A detailed description of the assumptions and functioning of ACT-R is provided in the appendix.

We test the ACT-R model in three simulation studies. In Study 1, a model of the core principles of ALIED is produced and the model is fitted to the data from Street et al. (2016) to assess how well this cognitively plausible model of lie-truth judgements explains human responding. In Studies 2 and 3, a novel prediction that arises from the model is tested with new data: that the lie-truth responses that participants make during the training phase is non-linearly predicted by the cue diagnosticity. Finally, in Study 3 we show that the model can generalise to situations where there are multiple individuating cues available to the decision-maker, demonstrating that the ALIED model can be generalised to more complex environments with multiple cues[1]

2. Study 1: An ACT-R model of lie-truth judgements

2.1. Model control flow

A high-level overview of how the model responds during the test phase is offered first. Upon reading a cue, the model attempts to retrieve from memory whether the cue is associated with honesty or deception. For each cue there exists two chunks in memory representing the possible associations: (i) the cue indicates honesty and (ii) the cue indicates deception. Whichever of the two chunks is the more active (as a result of prior learning and other factors discussed in the appendix) is the chunk that is most likely to be recalled. The recalled chunk forms the basis of the lie-truth judgement. This representation approach is similar to other instance-based and exemplar-based models of cognition (e.g., Gilboa & Schmeidler, 1995; Logan, 1988; Medin & Schaffer, 1978; Nosofsky, 1986), in that individual representations (chunks) exist for the different cue-response pairs and responses are determined according to similarity-based computations on these representations. Unlike exemplar-based models of categorisation however, responses are not a function of the summed similarity to all instances of a category but are determined by the individual instance most similar to the cue.

Context-general information acts somewhat like an attenuator during this memory retrieval process. Any information that is inconsistent with the current context is penalised by lowering its activation (e.g., when the context suggests people will mostly lie, chunks that associate a cue with telling the truth are penalised, as illustrated in Fig. 2). This similarity-based process, known as "partial matching" (see appendix for details) is conceptually similar to the Luce-Shepard rule (Luce, 1959; Shepard, 1962) in that the probability of a chunk being retrieved is a function of its activation strength relative to those of other chunks. For this model, the key effect is that chunks that are aligned with the current context become more likely to be retrieved. Once a chunk is retrieved, the lie-truth judgement is based exclusively on that recalled chunk, e.g., if the retrieved chunk indicates that this cue is associated with honesty, a truth judgement is made. This process is not deterministic but contains a degree of stochasticity so that greater activation increases the probability of retrieval (outlined in the appendix). The result of this process is that context-general information has a stronger effect on the judgement outcome when individuating cues have low (compared to high) diagnostic value, as ALIED claims.

This explanation applies equally to situations where the context-general information either suggests lying or truth-telling will be the more frequent. This symmetry accounts for the similarity between the easy and hard conditions in the human data revealed in Fig. 1. 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, meaning that a truth and lie bias arise from the same processes. This is one of the more unique claims of ALIED that is not made by competing theories of lie-truth judgements. While it is a controversial claim within our field insofar as the truth bias (but not the lie bias) has been observed across many studies (Bond & DePaulo, 2006), it is consistent with a Brunswikian lens model and with Bayes theorem, accounts that have been successful in explaining judgement formation more generally (e.g.,

[1] The ACT-R model code is available to download from GitHub: https://github.com/djpeebles/act-r-lie-detection-model.

Hartwig & Bond, 2011; Madsen & Pilditch, 2018; Reber & Unkelbach, 2010)[2].

A more detailed control flow of the model executing a single trial (in either the training or test phase) of the experiment is illustrated in Fig. 3. 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 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 lying. 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 "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") which the model reads and uses to update its knowledge about the cue by encoding (and thereby 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" (i.e., truth-telling) or "hard" (i.e., lying), 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 original experiment, no feedback was provided to the model in the test phase.

2.2. Results and discussion

2.2.1. Comparing human and model performance

The model was evaluated by running it 150 times (to simulate 150 experiment participants) for each context condition, and the data for each condition was averaged across participants. To fit the model to the human data, three parameters that modify the memory retrieval process were adjusted. The first is the threshold parameter in the chunk retrieval equation (τ in Eq. (A.2), see appendix) which represents the retrieval threshold, a psychological construct related to the accessibility of information in memory. Essentially, τ determines whether a chunk is to be included in the set of chunks being considered for retrieval by setting the minimum amount of activation required for a chunk to be retrieved; chunks with activation levels above this threshold are candidates for retrieval, while those below it are effectively inaccessible at that moment. The τ parameter can be interpreted as representing the amount of effort that participants are willing to spend on a retrieval. This parameter was set to 0.4.

The second is the activation noise parameter (s in Eq. (A.2) and the variance of the logistic function of ϵ in Eq. (A.1), see appendix) which represents the psychological phenomenon of random fluctuations in the accessibility of memory creating trial-by-trial variability in cognitive performance. This parameter was set to 0.21. The final parameter is the mismatch penalty (P in Eq. (A.5), see appendix) which represents the cost or difficulty associated with retrieving information that doesn't perfectly match the current context or cues. To do so, this parameter reduces the activation level of chunks based on how much they mismatch the specified retrieval cues; increasing the penalty makes the probability of retrieving a chunk not completely matching the cues. In the case of this experiment, it is the mismatch penalty working with the context-general information in the retrieval cue that differentially affects the retrieval of chunks in the two conditions. This parameter was set to 0.65. All three parameters were set to values that are well within the typical ranges for ACT-R models and the same parameter values were used for both context conditions.

To compare the model and human performance, the PTJ in the test phase for the easy and hard context conditions are presented in Fig. 4(a) and (b) respectively. 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. That is, the ALIED-inspired computational cognitive model provides a very good fit to the judgement outcomes of humans.

2.2.2. Explaining the model's performance during the test phase

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 (see appendix for greater detail). As the 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 the strength of association (i) between each cue and a truth response and (ii) between each cue and a lie response. Chunk 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. During the test phase these chunks are used as the basis for lie-truth judgements, with the chunk that has the greater activation level being retrieved.

Fig. 5 reveals that the effect of context-general information is greater as the individuating cue diagnosticity decreases, aligning with ALIED's prediction. Given that the model produces lie-truth responses in the training phase, one can see this context effect by observing how the PTJ during the training phase (i.e., before the model had access to context-general information) differs to the PTJ during the test phase.

To explain why this happens, consider the easy condition shown in Fig. 5(a) 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.

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 cognitive 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 0.0 and 1.0 (Gonzalez & Wu, 1999; Tversky & Kahneman, 1992). In the case of the current experiment, this manifests as a floor/ceiling effect that occurs when the difference in proportions becomes sufficiently large.

When environmental proportions reach a certain critical size, people may be unwilling to respond beyond a certain proportion (i.e., towards 100 % as they know that was not the case during training) and become

[2] That the lie bias is rarely observed, ALIED argues, is because lie detection studies tend to use the same methodology: to present participants with low diagnostic information and to allow participants to use their experience-based context-general information ("most people tell the truth") to inform that judgement. When the context changes such that most people will lie (e.g., Hartwig et al., 2004) or cues become highly diagnostic of deception (e.g., Bond et al., 2013), lie biases are observed. Studies such as these are the exception rather than the norm, and hence the lie bias being less regularly observed.

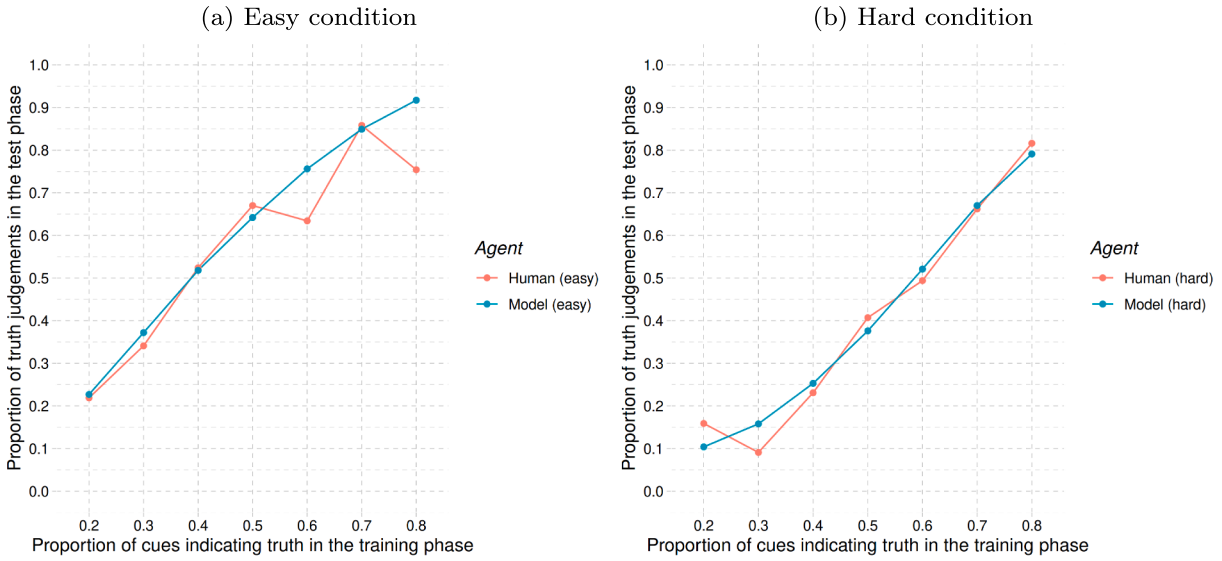


Fig. 4. Test phase PTJ for each proportion of truth cue condition for human and model respondents, separated by (a) easy condition and (b) hard condition. These figures show how both the model (green line) and human participants (red line) respond based on the cue diagnosticity.

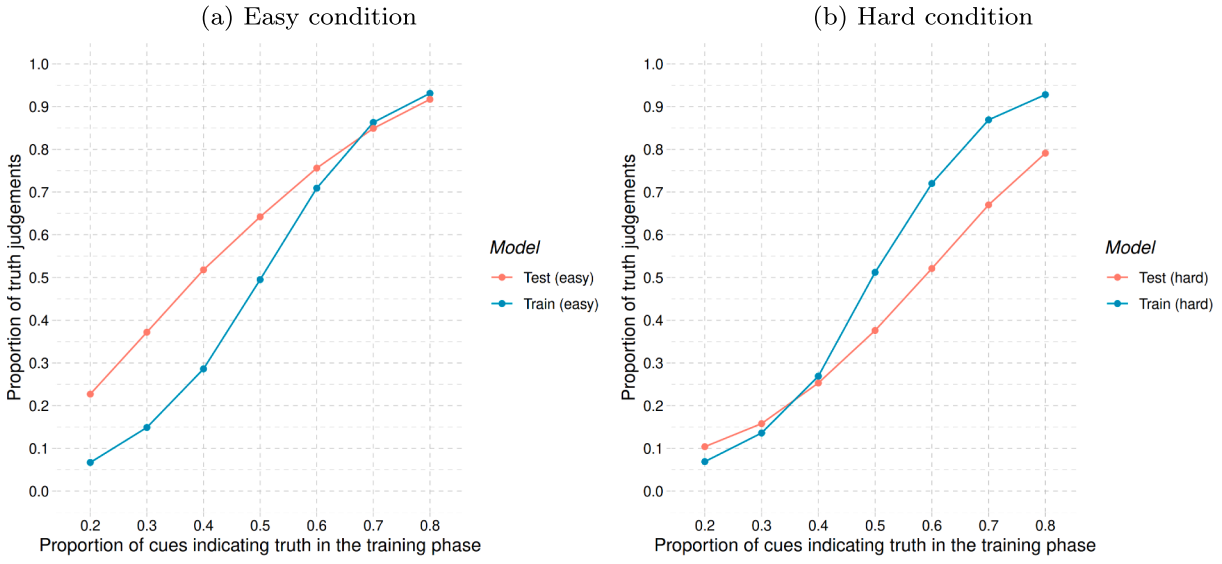


Fig. 5. The PTJ during the training phase (before context-general information is provided) and the test phase (after context-general information is provided) in (a) the easy condition and (b) the hard condition. In the easy condition ("most people will tell the truth"), the context increases the PTJ having the most noticeable effect at low cue diagnosticities, while in the hard condition ("most people will lie"), the context reduces the PTJ most noticeably at lower cue diagnosticities.

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, the context-general effect that serves to bias responses may have 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 95 % indicative of honesty (i.e., very highly predictive of honesty), for instance, 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 (see Fig. 2 for an overview).

2.2.3. *Explaining the model's learning during the training phase*

Fig. 6 reveals that, as the individuating cue became more diagnostic of honesty (i.e., greater than 0.5 on the x-axis of Fig. 6) or more diagnostic of deception (i.e., less than 0.5 on the x-axis of Fig. 6), the model had an increasing tendency to overestimate how likely the trivia game player was telling the truth or lying, respectively, during the training phase. It was only 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) that the model matched the experimental truth proportions. The discrepancy is approximately equal on either side of the 0.5 level, reaching a maximum discrepancy at the proportions 0.3 and 0.7 indicative of honesty.

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 human participants learned the various proportions before being given the context general information at the start of the test phase. To address this, Study 2 replicates the learning phase of Street et al. (2016) and collects data to test the model's prediction regarding the discrepancies between actual cue diagnosticity and participants' responding during the learning phase.

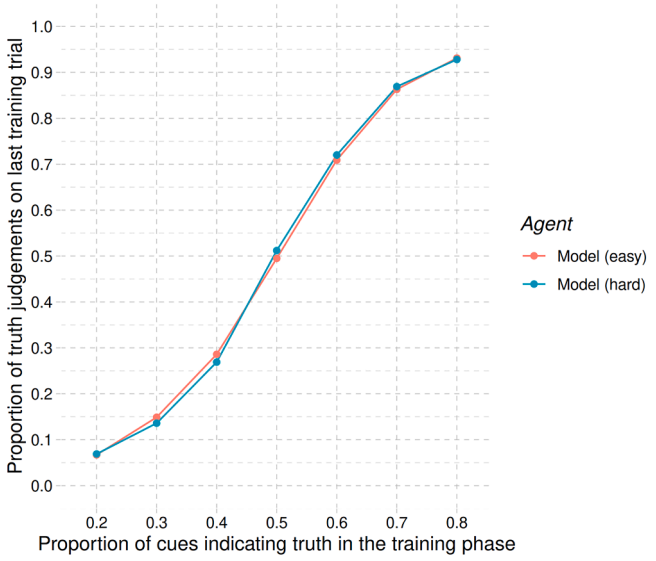


Fig. 6. The PTJ made by the model during the training phase, illustrating an overestimation of honesty or deception when cues were highly diagnostic of honesty or deception, respectively.

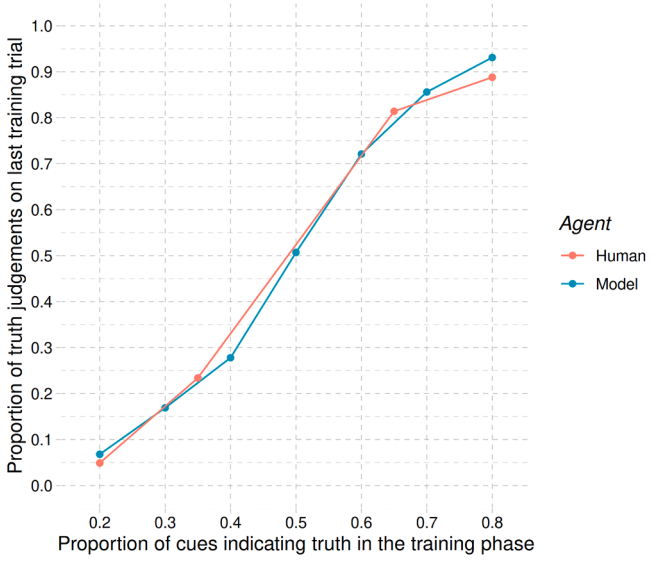


Fig. 7. The predicted PTJ from the model at the end of the learning phase, modelled on the data from Street et al. (2016) (green line), and the actual PTJ from human participants at the end of the learning phase from the current study (red line).

3. Study 2

Study 1 showed that people's judgements can be explained as being affected by both the statistics of the individual cue history combined with context-general information, with a high degree of match between model and human responses. The model represents an important step for the field because it offers novel, quantifiable, and testable predictions of how lie and truth judgements are formed, with the predictions arising from a psychologically informed model that details how memory encoding and retrieval processes result in the lie-truth judgment. For instance, the model predicts a mismatching between (i) the proportion of times that a cue suggesting honesty is present in the environment during learning and (ii) the proportion of truth judgements that are made during learning. This study provides a test of this model prediction. To our knowledge this is an entirely new prediction in this field.

3.1. Method

3.1.1. Participants

84 participants took part and were sampled opportunistically from University of Huddersfield students. Three participants were excluded because 50 % or more of their RTs in the test phase were less than 500ms (which we interpreted as indicating that they were not engaging adequately with the task), resulting in 81 participants (13 male, 68 female, $M_{age} = 21.54$, $SD_{age} = 5.35$) remaining for analysis. No trials were excluded due to overly long responses (defined as greater than 10 seconds). This study was not preregistered.

3.1.2. Design, materials, and procedure

This study had a test phase that was unrelated to the current topic of discussion and so will not be discussed here. The training phase was identical to that of Street et al. (2016) except that there were only four levels of cue diagnosticity indicative of honesty: 20 %, 35 %, 65 % and 80 %.

3.2. Results and discussion

To compare the learning of the participants in this study with that of the model developed in Study 1, the proportion of truth judgements on the last trial of the learning phase for both are plotted in Fig. 7. The data 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 et al., 2002).

This prediction, to our knowledge, has never been made about how deception information is learnt. The model's prediction is novel for the field and is supported both by these data and by the empirical literature from the broader field of judgement and decision-making research.

4. Study 3

The model can also be readily generalised to explain more complex situations such as when there are multiple indicators in the environment. In Street et al. (2016), the aim was to provide high control over the setting to test a theoretical prediction. To do so, the number of cues presented to a participant was limited to a single individuating cue. This offered direct and controlled testing of a lie-truth judgement account that was built on a much noisier previous literature where multiple individuating cues are typically presented to participants via video recordings. Video-based studies with no experimental control over the available cues make the development and testing of accounts with high temporal resolution and mathematical implementation near, if not entirely, impossible. In this study, we also extend the applicability of the model to addressing how decisions are made with multiple individuating cues present.

A key debate in judgement and decision-making research concerns whether people integrate multiple cues when making a judgement (a compensatory strategy) or whether they ignore most cues and use only the more reliable one (a non-compensatory approach: see Pohl (2011) for an overview of the debate). Street et al. (2016) attempted to distinguish these two approaches through optimisation modelling but found that both models performed comparably well in explaining the judgement outcomes. But that study was not designed with the primary aim of separating those two approaches. By providing the participant with multiple individuating cues with which to contend, this study is designed to test whether people are compensatory or non-compensatory decision makers. Both previously (Street, 2015; Street et al., 2016) and presently (Study 1), ALIED has taken the position of a compensatory

decision maker that uses all the information available to it to form a judgement.

If people behave in a non-compensatory way, pairing the most diagnostic cue indicating honesty (80 %) with either (i) a less diagnostic cue also indicating honesty (65 %) or (ii) a less diagnostic cue indicating deception (35 %) should have no effect on the judgement. This is because a non-compensatory decision maker would ignore the less diagnostic cue (35 % or 65 %) and use only the most diagnostic cue available (80 %). However, a compensatory decision maker would use all information, and so when the highly diagnostic honesty cue is paired with a cue that is less diagnostic but suggests deception (80 %–35 % pairing), there should be fewer truth judgements compared to when the highly diagnostic honesty cue is paired with a less diagnostic but also truth-indicative cue (65 %–80 % pairing). Observing significantly fewer truth judgements in the 35 %–80 % pairing compared to the 65 %–80 % pairing would be inconsistent with a non-compensatory decision-making scheme. This research question is also addressed in this study.

The study also demonstrates proof of principle that the model can generalise to situations beyond the presence of a single individuating cue. The principles employed can readily be scaled up to any number of individuating cues, but by limiting to two individuating cues we are able to provide a controlled and parsimonious test of the model's explanatory power. This study provides an in-principle demonstration of the model's ability to be generalised to more typical lie-truth formation studies where multiple cues are presented. Additionally, the model is fit once again to the training data, as per Study 2, for purposes of replicating the effect and demonstrating that the generalisation of the model does not undermine its performance relative to the study-specific modelling approach of Study 1.

4.1. Method

4.1.1. Participants

Participants were 130 undergraduate psychology students at the University of Huddersfield who took part for course credit. Participants' age and (binary) gender were requested but not compulsory and 92 students reported both details while the remainder reported neither. The mean age of those who did provide information was 20.15 ($SD = 3.85$) and 81 were female. This study was not preregistered.

From the 130 participants, data from $n = 25$ were excluded, $n = 12$ because 50 % or more of their RTs in the test phase were less than 500ms (which we interpreted as indicating that they were not engaging adequately with the task), $n = 12$ when it was discovered that they had taken part in the experiment twice, and $n = 1$ because they did not read the instructions. All test phase trials with an RT greater than 10s (which is greater than 4 standard deviations from the mean) were not included in the analysis. This procedure removed 172 trials (1.37 % of the total number of trials) from 66 participants from the test phase with no more than nine trials being excluded from an individual participant. This rule resulted in no exclusions from the training phase data.

4.1.2. Materials and procedure

The experiment was identical to Street et al. (2016) in all but two respects. First, in the training phase, there were only four levels of cue diagnosticity indicative of honesty: 20 %, 35 %, 65 % and 80 %, identical to Study 2 above. Second, in the test phase participants saw each cue paired with another cue 20 times, creating six combination conditions: 20 %/35 %, 20 %/65 %, 20 %/80 %, 35 %/65 %, 35 %/80 % and 65 %/80 %. To create the test phase blocks the 120 trials were randomly ordered and then presented as a single batch with breaks interspersed after every 40th trial. In a test trial the two cues were presented simultaneously, one above the other on the screen.

4.1.3. Generalising the model

The model described here remains unchanged apart from modifications that allow two cues to be used for retrievals during the test phase. Like the previous models, the context-general response bias in this model is a slot value in retrieval requests during the test phase, influencing retrieval through the partial matching mechanism: cue-response chunks whose response slot matches the bias associated with the current context have higher activation, while mismatching chunks are penalised according to ACT-R's mismatch penalty equation.

Unlike the previous models however, in this model cues are not used with the response bias as a direct retrieval constraint, but remain as chunks in the imaginal buffer, influencing retrievals via ACT-R's *spreading activation* mechanism (see appendix for details). This different approach is required because using multiple cues as slot constraints makes retrieval requests overly specific and because spreading activation allows both cues to influence retrieval by increasing the retrieval likelihood of any chunk in memory that is related to the two individuating cues. This graded, probabilistic contribution better reflects how multiple sources of information influence memory retrieval without requiring exact structural matches, and remains consistent with the underlying claim that retrieval is probabilistically influenced by learned cue–response associations and context.

4.2. Results and discussion

4.2.1. Behavioural results

Fig. 8 shows the proportion of truth judgements for each diagnosticity combination for both the easy and hard conditions (that is, the context suggested most would tell the truth or that most would lie, respectively).

The core behavioural research question is whether people integrate multiple cues to make their judgement (a compensatory approach) or whether they ignore some cues and use only the more reliable ones (a non-compensatory approach). Consistent with a compensatory decision-making scheme, the 65 %–80 % cue pairing resulted in a higher proportion of truth judgements ($M = 0.693, SD = 0.250$) compared to the 35 %–80 % cue pairing ($M = 0.536, SD = 0.273$), $t(104) = 5.38, p < 0.001, d = 0.53, 95\%$ CI on $d[0.32, 0.73]$. Similarly, the 20 %–35 % cue pairing resulted in a lower proportion of truth judgements ($M = 0.284, SD = 0.244$) compared to the 20 %–65 % cue pairing ($M = 0.426, SD = 0.292$), $t(104) = 4.33$, $p < 0.001$, $d = 0.42$, $[0.22, 0.62]$.

A secondary analysis tested whether the context-general information had the greatest effect when the joint probability of honesty resulting from the individuating cues was nearer chance and the least effect when the joint probability of honesty or deception was above chance. This prediction arises from ALIED's claim that people show greater reliance on context-general information as the individuating cues become less diagnostic.

A reasonable prediction is that context would have the greater effect (i.e., the difference in PTJ between the context conditions would be greatest) when the individuating cue pair had a joint low diagnosticity (i.e., the 35 %–65 % cue pair). The difference in PTJ between context conditions was predicted to be smallest when the individuating cue pairs were highly diagnostic of a particular decision—i.e., either 20 %–35 % (both suggesting deception) or 65 %–80 % (both suggesting honesty). This integration of the two cues assumes a compensatory decision-making scheme, which is reasonable given the above results. But the same prediction is also consistent with decision-makers using a non-compensatory strategy who select only the most reliable individuating cue and only use context when the most diagnostic individuating cue available has low diagnosticity.

A 2 (context, between subjects: hard and easy) $\times 3$ (cue pair, within subjects: 20 %–35 %, 35 %–65 %, and 65 %–80 %) mixed ANOVA was conducted with the PTJ as the dependent variable. The key predicted interaction was non-significant, $F(1.81, 185.98) = 0.71$, $p = 0.477$, $\eta^2 = 0.007$. There was a linear main effect of cue pair such that the smallest observed PTJ was when the 20 %–35 % cue pairing was presented (estimated marginal mean, or $EMM = 0.286, SE = 0.024$) and the largest observed PTJ was when the 65 %–80 % cue pairing was

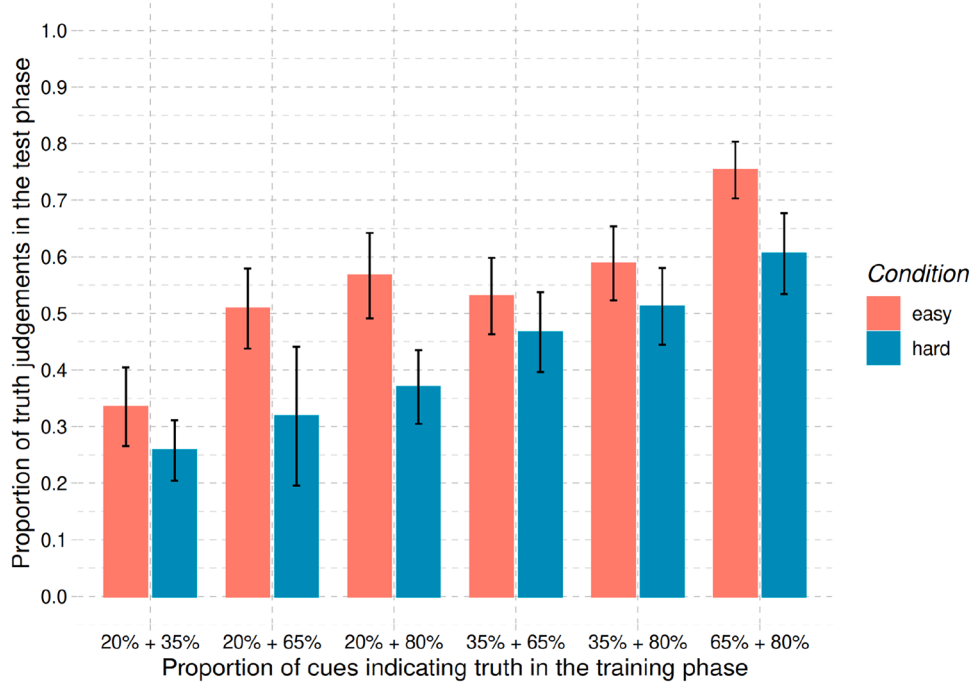


Fig. 8. PTJ in the test phase for each proportion of truth cue conditions in the training phase separated by easy and hard game conditions, Study 2. Error bars denote 95 % confidence intervals.

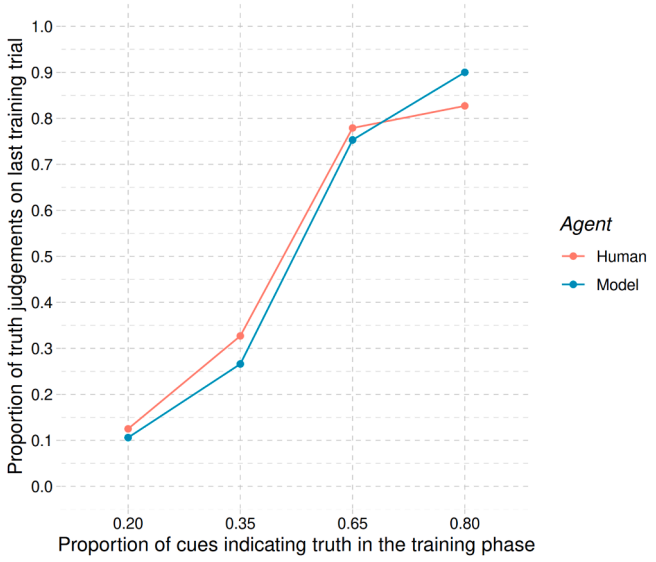


Fig. 9. Proportion of truth judgements for proportions of truth cue conditions on the last trial of the learning phase, humans and model, Study 3.

presented ($EMM = 0.696, SE = 0.024$), $F(1.81, 185.98) = 68.47$, $p < .001$, $\eta^2 = 0.399$. A main effect of context was also observed: participants made more truth judgements in the easy condition ($EMM = 0.544, SE = 0.021$) compared to the hard condition ($EMM = 0.453, SE = 0.020$), $F(1, 103) = 9.97$, $p = .002$, $\eta^2 = 0.088$. Thus, it would appear that there is evidence to reject the lack of an effect of context and to reject the lack of an effect of cue-pairing, but there was no evidence regarding whether context and cue pairing interacted.

To assess whether the interaction effect demonstrated evidence that shifts belief towards the null rather than a lack of sufficient information, a default Bayes factor was calculated using a Cauchy prior centred on zero and scaled over Cohen's d scaled at $r = 0.50$ to determine the probability of the evidence under a model of just the main effects of context and cue-pair compared to a more complex model that also included their interaction. The Bayes factor indicates that a shift in credibility towards the simpler model (without the interaction) is warranted by a factor of 8.36.

Street et al. (2016) developed two models of the judgement process: a compensatory decision-making interpretation of ALIED (i.e., where multiple cues are weighted and integrated into the judgement, in this case using Bayes theorem) and a non-compensatory decision-making interpretation of ALIED (i.e., where only the most diagnostic cue is used in the judgement and all other cues are ignored). Both models fit the data relatively well and it was not possible to distinguish these accounts. The behavioural data from the current study is difficult to align with a non-compensatory strategy, contributing to the compensatory versus non-compensatory decision-making debate in the domain of lie-truth judgement formation (see Pohl, 2011).

4.3. Model results

To fit the model to the human data, four parameters were adjusted. The first three (retrieval threshold, activation noise, and mismatch penalty) were described in relation to the previous model and were set to 0.4, 0.3, and 0.6 respectively. These values are very similar to the earlier model (0.4, 0.21, and 0.65 respectively). The fourth parameter relates to the additional spreading activation mechanism used for this model in the chunk activation equation (S in Eq. (A.1), see appendix). This parameter determines the strength of association between chunks in the retrieval request and the chunks in memory (i.e., how much weight the cues have in retrieving a chunk relative to the chunk's current base-level activation determined by its history of use). This parameter was set to 3.0. As with the previous model, 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 model was evaluated by running it 60 times for each experiment condition and the data for each condition averaged.

The PTJ on the last trial of the learning phase for the human and model responses are plotted in Fig. 9. The data show that people

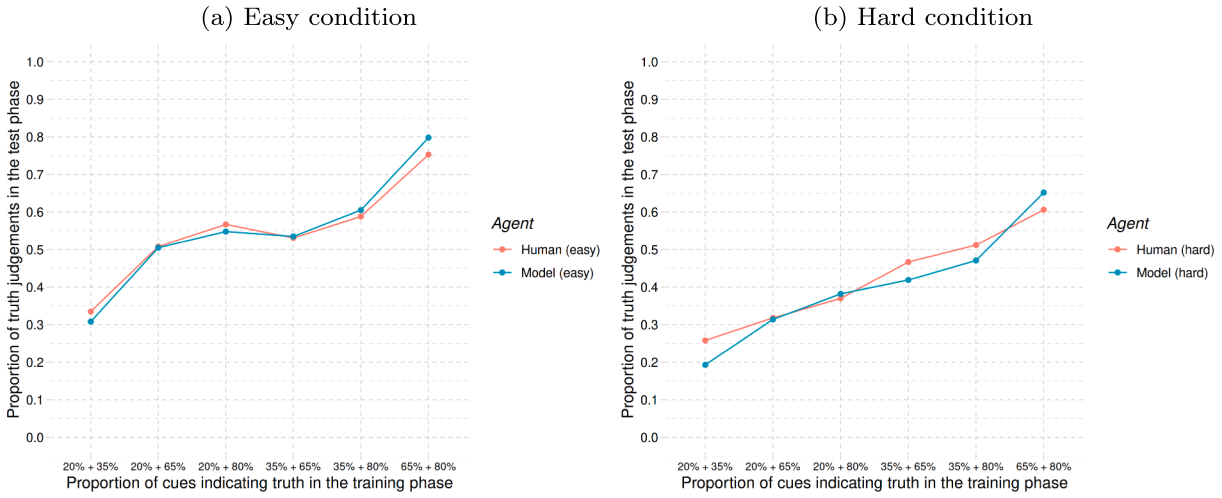


Fig. 10. Proportion of test phase truth judgements for each proportion of truth cue conditions in the training phase, human and model, Study 3, easy condition (a), hard condition (b).

are—as the model predicts—maximising their responses during the learning phase ($R^2 = 0.80$, $RMSD = 0.31$). This result replicates Study 2 and is in accord with previous evidence that people do not simply match environmental probabilities during probability learning tasks when provided with feedback (e.g., Newell et al., 2013).

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 Fig. 10(a) and (b) respectively. The fit of the model to the human data for both difficulty conditions was very close, $R^2(easy) = 0.99$, $RMSD(easy) = 0.03$, $R^2(hard) = 0.93$, $RMSD(hard) = 0.05$. By maintaining the model's core structure but making a small adaptation to generalise the model to handle multiple individuating cues, the model has been able to be applied to more complex tasks. Importantly, the model has been able to provide a close fit the human data and capture novel, non-linear behaviour to provide an explanation of lie-truth judgements. To date, there exists no theory of lie detection that can make predictions at this level of explanation.

4.3.1. *Explaining the model's performance during the test phase*

For the training phase of Study 3, the model is identical to that for Study 1 apart from a generalisation to allow encoding of multiple cues. As with the previous model, the training phase creates eight trial chunks in long-term memory, two for each cue—one associated with truth, the other associated with lie. After the training phase the model was provided with the experiment condition information, "easy" or "hard", and before starting the test trials the model retrieved from memory the context-general response bias associated with each ("truth" or "lie" respectively)[3]. Once retrieved, this response bias then became an element of the goal which was used as an additional probe in subsequent memory retrieval requests for individuating 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. Partial matching 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.

5. General discussion

We present the first cognitively plausible computational model of lie-truth judgements by implementing it within the ACT-R cognitive architecture. Our model provides a close fit to human performance data across three studies. It successfully accounts for how individuating cues (specific to a statement) interact with context-general information (base rates of honesty) to influence judgements. The model demonstrates that both truth and lie biases emerge from the same underlying memory mechanisms, challenging traditional views that these biases have different cognitive origins (cf. Gilbert et al., 1990; Levine, 2014). Our simulations closely aligned with human data, accurately predicted novel behavioural outcomes validated through empirical tests (Studies 2 and 3), and successfully generalised to scenarios involving multiple individuating cues.

Integrating ALIED with ACT-R bridges computational and algorithmic levels of explanation (Cooper & Peebles, 2015), providing a mechanistic account of how lie-truth judgements form. By grounding ALIED in established cognitive principles of memory encoding, activation, and retrieval, we move beyond descriptive accounts to explain precisely how people process deception cues. This implementation shows how ALIED's higher-level Bayesian description translates to specific memory operations that produce adaptive judgement patterns observed in humans. The rigorous constraints imposed by ACT-R provide valuable mechanistic insights into cognitive processes that previously remained unspecified within ALIED.

Our model generated the novel prediction that people maximise rather than match response probabilities during cue learning. Study 2 confirmed this prediction, showing that participants overestimate the diagnosticity of cues during training. The model also predicted the functional equivalence of truth and lie biases, supported by the symmetrical pattern of responses observed when context-general information was transformed between conditions. These successful

[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.

predictions highlight the predictive power of the model beyond simply fitting existing data, offering robust empirical support for our theoretical claims.

The model clarifies how people balance individuating and context-general information when forming judgements. When individuating cues have high diagnosticity, their memory activation strongly drives the judgement, minimising context effects. As cue diagnosticity decreases, context-general information exerts greater influence through partial matching in memory retrieval. This explains why context has stronger effects when individuating information is ambiguous, providing a cognitive mechanism for ALIED's core claims.

Study 3 addressed whether people integrate multiple cues (compensatory decision-making) or rely solely on the most diagnostic cue (non-compensatory decision-making). Results showed that participants' judgements were affected by both cues in a pair, even when one cue was more diagnostic than the other. The model captures this integration through spreading activation in memory, where multiple cues simultaneously influence retrieval probability. This addresses ongoing debates in the field, supports ALIED's position and aligns with broader research in judgement and decision-making, confirming that people generally utilise all available information adaptively to make informed judgements.

One response to this claim is that, rather than providing evidence for the integration of multiple cues within a trial, the human data may result from the averaging of single cue responses across trials. While Study 3 is consistent with compensatory integration of individuating cues, we cannot definitively rule out a non-compensatory process where the most diagnostic cue dominates on each trial and the aggregate pattern mimics integration. This distinction parallels a broader issue in memory modelling: ACT-R's fan effect model (e.g., Anderson, 1974) assumes graded averaging of evidence from multiple features, whereas other paradigms such as Extra-List Feature Effect (Mewhort & Johns, 2000) suggest a winner-takes-all strategy. Future experimental work could build on this contrast to more precisely test whether cue integration in lie-truth judgement is genuinely compensatory or reflects cue selection under uncertainty. Such work would not only refine models of deception judgment but also contribute to our understanding of cue weighting in memory-guided decision making more generally.

Our ACT-R implementation imposes rigorous cognitive constraints on the model, including realistic memory dynamics and processing limitations. The model operates on the same 50ms processing cycle as other models of human cognition, with all parameters within established ranges for ACT-R models. This computationally explicit representation of cognitive operations significantly advances theoretical explanations beyond heuristic or descriptive accounts, making the cognitive plausibility of ALIED explicit and testable.

It is possible to develop a model of this task that would rely on ACT-R's procedural mechanisms rather than declarative memory, and individual differences in strategy could give rise to greater reliance or emphasis on different memory systems (Yang et al., 2024). However, our declarative memory based model was guided by both theoretical and empirical considerations, specifically ALIED's assumption that performance is based on the retrieval and weighting of individuating and context-general information. In contrast, a procedural model would require abstracting cue-judgement associations into production utilities, which would obscure the underlying memory-based representational structure that ALIED presupposes. Moreover, the model's predictions about learning trajectories depend on base-level activation dynamics in declarative memory, effects that would be less naturally or transparently captured in a procedural, utility-based approach.

This work also supports the conception of lie-truth judgements as adaptive rather than error-prone. Critically, our model demonstrates that lie and truth biases can emerge from identical cognitive operations, challenging traditional perspectives which consider these biases as fundamentally distinct or default psychological states (Gilbert et al., 1990; Levine, 2014). If biases emerge from optimal use of available information, improving lie detection may require enhancing cue diagnosticity rather than attempting to overcome presumed cognitive biases. The model suggests that low accuracy in lie detection stems primarily from poor individuating cue diagnosticity in typical environments rather than faulty cognitive processing, aligning with ecological rationality perspectives.

While our model makes significant theoretical contributions, future research should address limitations related to ecological validity by exploring contexts closer to real-world lie detection scenarios. Future work should incorporate emotional and social interactive factors in decision-making and test the model with richer, more complex stimuli (Ask et al., 2020). The current model focuses on memory-based processes but could be expanded to include perceptual and attentional mechanisms that influence cue selection. Longitudinal studies could test how cue-response associations develop over time and across varied contexts, further validating the adaptive nature of lie-truth judgements in real-world settings. Finally, future experiments should explicitly test the robustness of compensatory cue integration in more complex, realistic settings with multiple competing cues and greater environmental uncertainty.

CRediT authorship contribution statement

David Peebles: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Conceptualization; **Chris N.H. Street:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization.

Data availability

A link to the model code has been made available in the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors would like to thank Dan Bothell for his invaluable advice in developing the model and Hanah Ahmed, Ibrahim Ali, Harry Sheard and Rachel West for their help in collecting the data for Studies 2 and 3.

Appendix A. Details of the modelling approach

The cognitive process model described in this paper provides a detailed, mechanistic account of adaptive decision making resulting from the interaction between individuating and context-general cues. The model is instantiated in the ACT-R theory of human cognition (Anderson, 2007) and so in this appendix we introduce the ACT-R cognitive architecture, detailing the mechanisms relevant to the model.

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 principles of declarative and procedural knowledge representation, cognitive control, and learning, derived from decades of experimental data and theoretical development, to create complete, integrated processing models of cognition. An important feature of ACT-R is that it is implemented

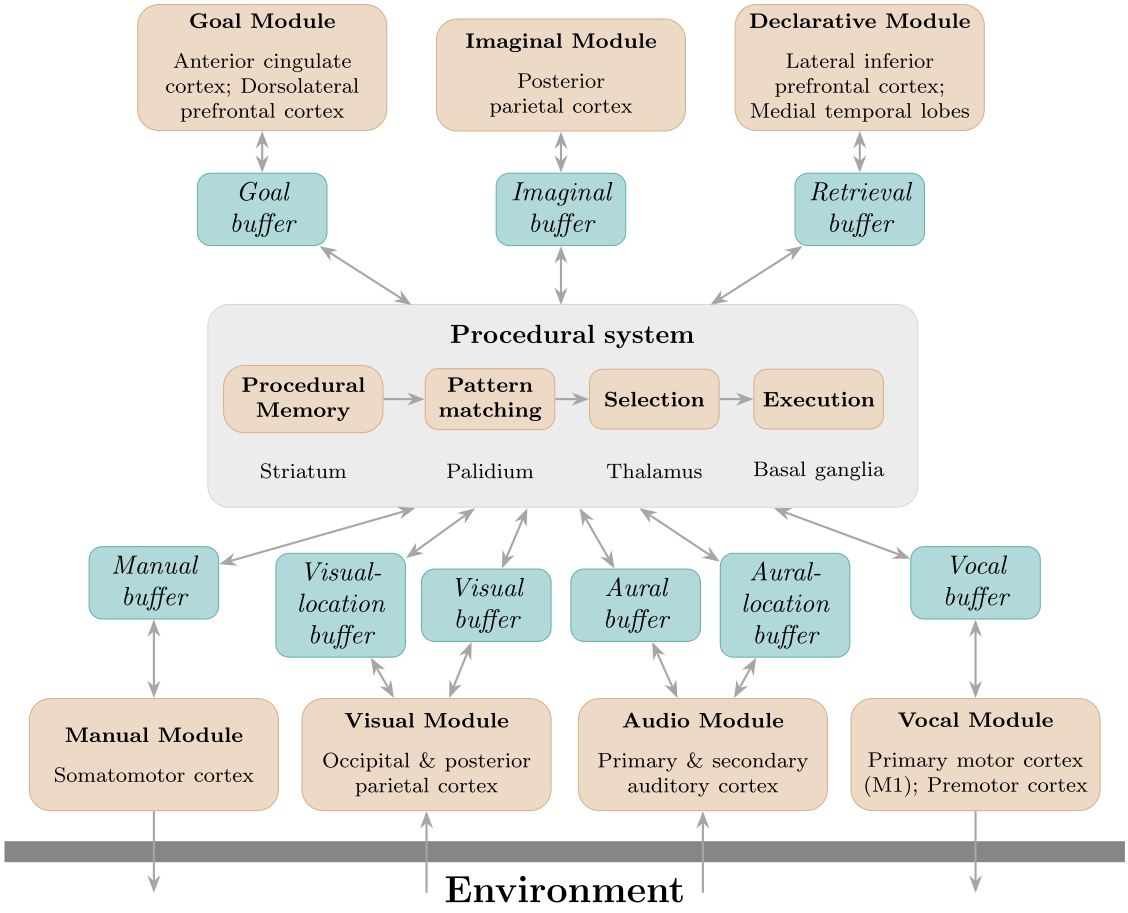


Fig. A.11. Functional components of the ACT-R cognitive architecture.

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.

ACT-R consists of a set of modules (shown in Fig. A.11). A module for declarative memory codes a network of knowledge 'chunks', where a chunk may be, for example, "liars avoid eye contact". The procedural memory module is represented by sets of production rules that encode knowledge about behaviour, e.g., when a behavioural cue is detected, retrieve from memory whether this cue is associated with deception or honesty. Other modules for vision and motor control are also specified.

Cognition is goal driven and proceeds in ACT-R through a set of if-then rules (called production rules) which are triggered by a match with information in the cognitive system. For example, in order to be executed, a particular if-then production rule may require that a word has been read from the computer screen. If this condition is true, then the rule "fires" and its actions will take effect (e.g., if a word has been seen on the screen, then look up the meaning of the word in memory). The result of this production rule will change the state of the cognitive system (i.e., the meaning of the word will now be available), and so different production rules may match the new state of the system. Tasks are performed through the successive sequential actions of production rules and the cycle in which buffers hold representations, patterns in these buffers are matched, a production fires, and the buffers are then updated for another cycle takes 50 ms (a value shared by several other influential cognitive architectures, including EPIC (Meyer & Kieras, 1997), Soar (Newell, 1990), and CAPS (Just & Carpenter, 1992).

There are two key values that affect the learning and retrieval of information in the system: chunk activation and production rule utility. Chunk activation determines the probability and speed of retrieving information or chunks from long-term memory. The utility of a production rule is used to resolve conflicts in production selection when two or more production rules match the state of the cognitive system.

Chunks

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.

Declarative memory

The three mechanisms underlying 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

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

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, S_{ji} is the strength of association between element j and chunk i, and ϵ is a noise component.

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

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

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

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

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.

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

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

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).

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

$$P_i = \sum_k P M_{ji} \tag{A.5}$$

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 subsequently being recalled. This is critical to implementing ALIED's context-general information use.

Spreading activation

Recall that in Eq. (A.1) the activation of a chunk i in declarative memory, A_i is affected, in part, by an *associative* component that reflects its relevance to the current context C (i.e., all the chunks j currently held in ACT-R's buffers). By default only chunks in the imaginal buffer influence a retrieval in this way.

During a retrieval request, if spreading activation is engaged, the chunk in the imaginal buffer spreads activation to each training trial chunk i in declarative memory based on the contents of its slots j (in this case the two cues). The amount of activation spread by a cue j to chunk i is based on the strength of association, S_{ji} and the attentional weighting given to element j, W_j. Associative strength is defined as

$$s_{ji} = \begin{cases} 0 & \text{If} j \not\equiv i \text{and} j \text{is not in a slot of chunk} i \\ S - \ln(\text{fan}_{ji}) & \text{If} j \equiv i \text{or chunk} j \text{is in a slot of chunk} i \end{cases} \tag{A.6}$$

where S is the maximum associative strength and fan_{ji} is a measure of how many chunks are associated with chunk j (Anderson, 1974).

References

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*(4), 451–474.

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.

Ask, K., Calderon, S., & Mac Giolla, E. (2020). Human lie-detection performance: Does random assignment versus self-selection of liars and truth-tellers matter? *Journal of Applied Research in Memory and Cognition*, *9*(1), 128–136.

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.

Bond, G. D., Malloy, D. M., Arias, E. A., Nunn, S. N., & Thompson, L. A. (2005). Lie-biased decision making in prison. *Communication Reports*, *18*(1–2), 9–19.

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

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.
Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, *67*(337), 81–102.
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, 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*, 643–659.
Hartwig, M., Granhag, P. A., Strömwall, L. A., & Andersson, L. O. (2004). Suspicious minds: Criminals' ability to detect deception. *Psychology, Crime & Law*, *10*(1), 83–95.
Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.
Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, *19*(1), 1–17.
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.
Levine, T. R., & Street, C. N. H. (2024). Lie–truth judgments: Adaptive lie detector account and truth-default theory compared and contrasted. *Communication Theory*, *34*(3), 143–153.
Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527.
Luce, R. D. (1959). Individual choice behavior: A Theoretical Analysis. New York: John Wiley.
Luke, T. J. (2019). Lessons from pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science*, *14*(4), 646–671.
Madsen, J. K., & Pilditch, T. D. (2018). A method for evaluating cognitively informed micro-targeted campaign strategies: An agent-based model proof of principle. *PloS one*, *13*(4), e0193909.
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*(3), 207–238.
Mewhort, D. J. K., & Johns, E. E. (2000). The extralist-feature effect: Evidence against item matching in short-term recognition memory. *Journal of Experimental Psychology: General*, *129*(2), 262–284.
Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part i. basic mechanisms. *Psychological Review*, *104*(1), 3–65.
Millar, M. G., & Millar, K. U. (1997). The effects of cognitive capacity and suspicion on truth bias. *Communication Research*, *24*(5), 556–570.
Montag, J. L. (2021). Limited evidence for probability matching as a strategy in probability learning tasks. In *Psychology of learning and motivation* (pp. 233–273). Elsevier (*vol. 74*).
Newell, A. (1990). Unified theories of cognition. Harvard University Press.
Newell, B. R., Koehler, D. J., James, G., Rakow, T., & Van Ravenzwaaij, D. (2013). Probability matching in risky choice: The interplay of feedback and strategy availability. *Memory & Cognition*, *41*, 329–338.
Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.
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.
Pohl, R. F. (2011). On the use of recognition in inferential decision making: An overview of the debate. *Judgment and Decision Making*, *6*(5), 423–438.
Reber, R., & Unkelbach, C. (2010). The epistemic status of processing fluency as source for judgments of truth. *Review of Philosophy and Psychology*, *1*(4), 563–581.
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.
Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, *27*(2), 125–140.
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.
Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.
Yang, Y. C., Sibert, C., & Stocco, A. (2024). Reliance on episodic vs. procedural systems in decision-making depends on individual differences in their relative neural efficiency. *Computational Brain & Behavior*, *7*(3), 420–436.