Combining individuating and context-general cues in lie detection

David Peebles (d.peebles@hud.ac.uk)
Department of Psychology, University of Huddersfield,
Queensgate, Huddersfield, HD1 3DH, UK

Chris N. H. Street (c.street@keele.ac.uk)
School of Psychology, Keele University, Keele,
Staffordshire, ST5 5BG, UK

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 a judgement of either truth or lie. One account, the Adaptive Lie Detector theory (ALIED: Street, Bischof, Vadillo, & Kingstone, 2016) is sufficiently well defined that its assumptions may be implemented in a computational model. In this paper we describe our attempt to ground ALIED in the representations and mechanisms of the ACT-R cognitive architecture and then test the model by comparing it to human data from an experiment conducted by Street et al. (2016). The model provides a close fit to the human data and a plausible mechanistic account of how specific and general information are integrated in the formation of truth-lie judgements.

Keywords: ALIED; ACT-R; Lie Detection

The Adaptive Lie Detector theory

The Adaptive Lie Detector theory (ALIED Street, 2015) argues that people make reasonable use of the information available to them to adaptively reach a lie or truth judgement. According to ALIED, 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 a statement is a lie or true (e.g., Pinocchio's nose growing), this has the heavier weight in the judgement. But when individuating information has low diagnosticity, ALIED claims that people do not simply guess at random, but rather that their context-general knowledge (e.g., "most people tell the truth in this setting") has the heavier weight in the judgement, thereby allowing for a satisficing judgement in the absence of more diagnostic individuating information. This account is considered adaptive insofar as it reflects an informed decision-making process.

Individuating information varies in its degree of diagnosticity however. If a particular cue has been learnt to be highly diagnostic, people use this information to attain high accuracy (Blair, Levine, & Shaw, 2010; G. D. Bond, Malloy, Arias, Nunn, & Thompson, 2005; Levine & McCornack, 2014). But individuating cues to deception are typically either unavailable (Luke, 2019) or if they are available, have low diagnosticity (DePaulo et al., 2003; Hartwig & Bond, 2011; Sporer & Schwandt, 2006). In such cases, people will try to bring relevant general knowledge to bear on the issue, for example their beliefs about the prevalence of lying in certain situations. As the diagnosticity of individuating cues is perceived to reduce, ALIED argues that context-general knowledge will have a heavier weighting in the judgement process to make an informed—but overgeneralised—judgement.

Because individuating cues typically have low diagnosticity, and because context-general information is informed by the fact that people predominantly tell the truth (DePaulo, Kashy, Kirkendol, Wyer, & Epstein, 1996; Halevy, Shalvi, & Verschuere, 2014), in most situations it is rational to make a truth judgement, leading to the frequently observed tendency to judge others' statements as truths (known as the "truth bias": C. F. Bond & DePaulo, 2006).

By contrast, in situations where lying is more prevalent or the belief that people lie is widespread (i.e., context-general information suggests that people typically lie), ALIED claims that the bias is to judge statements as being untruthful, for which there is evidence (e.g., G. D. Bond et al., 2005; Masip, Alonso, Garrido, & Herrero, 2009). ALIED argues therefore that the observed truth bias often found in studies is not a cognitive disposition, but rather an adaptive judgement in the absence of specific (individuating) information.

Testing the ALIED theory

Street et al. (2016) conducted an experiment to investigate the interactive effect of individuating and context-general cues on lie detection. Participants were required 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 took 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; Nosofsky, 1986), consisting of a training phase followed by a test phase. Crucially however, this experiment differed from typical designs by introducing additional context-general information prior to the test phase to measure its effect on the learned associations with individuating cues.

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 their 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 based on these cues and received feedback on their responses. They learnt about four cues in this way: voice pitch, facial expression, number of silent periods in sentences, and number of self-references such as 'I' and 'me'.

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 the same lie-truth responses to the presentation of individual cues 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, calculated as the number of truth judgements divided by the total number of judgements made) for each cue diagnosticity. These are shown for both the easy and hard conditions in Figure 1.

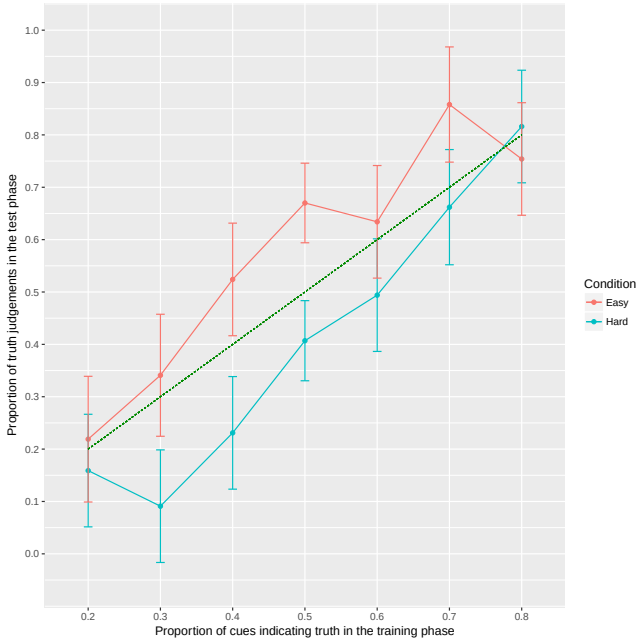


Figure 1: Proportion of truth judgements on trials for each diagnosticity level of truth cue in the training phase, split by easy and hard game conditions, from the experiment reported in Street et al. (2016). The dotted line indicates where the proportion of judgements matches the frequency of the cue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

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 than in the hard condition. The analysis also revealed a significant interaction between cue diagnosticity and context in that the effect of context-general information is greater as the individuating cue diagnosticity decreases, in line with ALIED's prediction. These results provide a quantitative demonstration of how people's judgements arise 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.

Human performance in the experiment can be accounted for by a Bayesian model (Street et al., 2016) but questions remain concerning how the two types of knowledge are learned and cognitively represented and what the cognitive mechanisms of interaction that produce the observed adaptive behaviour may be. Like all other proposed explanations of lie-truth judgement formation, ALIED is silent on these matters, and yet invariably these core cognitive mechanisms are fundamental to the judgement process. To address this shortcoming, we describe below a cognitive process model to account for the human performance revealed in this experiment. Developed within the ACT-R theory (Anderson, 2007), the model explains performance in terms of the basic learning and retrieval mechanisms of declarative memory, and provides an algorithmic level account of lie-truth judgement formation consistent with the assumptions of ALIED.

The ACT-R cognitive architecture

ACT-R is a well-established theory of the core components of the human cognitive system, including declarative and procedural memory, cognitive control, and learning, and how they are integrated to produce intelligent behaviour. Space limitations preclude a detailed description of ACT-R here but a comprehensive account of the cognitive architecture can be found in Anderson (2007).

In summary, ACT-R consists of a set of modules that acquire information from the environment, process information, and execute motor actions to achieve goals. ACT-R's declarative memory consists of a network of symbolic knowledge chunks while its procedural memory comprises a set of "if…then" production rules. Cognition proceeds via a symbolic pattern matching process that selects production rules whose conditions match the state of the system and tasks are performed through the successive actions of productions.

ACT-R augments the symbolic knowledge level with a subsymbolic level of numerical computations that instantiate learning mechanisms governing production rule selection and declarative retrieval. As the current model's performance is determined primarily by the declarative memory learning and retrieval mechanisms these are now described in more detail.

Chunks in declarative memory each have a level of activation, a value related to the recency and frequency of their use. The learning and forgetting of chunks is achieved through the gradual adjustment of their activations over time which affects the probability and timing of their subsequent retrieval. When chunks are created in declarative memory, they have an initial level of activation which decays over time, reducing the probability of their subsequent retrieval. Memory re-

trieval in ACT-R occurs when a production rule contains a retrieval request to the declarative memory module containing one or more cues, in the case of the current experiment, the cues being the individuating features and context-general information. The activation of a chunk i, A_i, is defined as

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

where B_i is the *base-level* activation of chunk i which reflects the history of the chunk's use, 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.

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 defined as

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{2}$$

where n is the number of presentations of chunk i, t_j is the time since the j^{th} presentation, and d is the parameter determining the activation decay rate. The retrieval probability of each chunk i, P_i is a function of its activation, A_i and is defined as

$$P_i = \frac{1}{1 + e^{\frac{-(A_i - \tau)}{s}}} \tag{3}$$

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 matching a retrieval request, the probability of chunk i being selected is a function of its activation relative to the activations of the others. If no chunk has an activation greater than the retrieval threshold then none will be retrieved and a retrieval failure will be signalled.

To capture aspects of human memory behaviour such as retrieval errors, ACT-R has a *partial matching* mechanism that takes the similarity between the probe and memory chunks into account. With partial matching, the activations of chunks of the same type as the probe are modified in proportion to their similarity to the probe according to

$$P_i = \sum_k P M_{ji} \tag{4}$$

where P_i, the partial matching value of chunk 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. The result is that when a chunk does not completely match a retrieval request, its activation is reduced in proportion to the number of mismatching slots, reducing its retrieval probability. This mechanism is critical to capturing the effect of context-general information in the experiment.

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. ACT-R is implemented as a software system and incorporates models of vision and motor control which can be connected to—-and interact with—external task and simulation environments. The model we developed runs in a simulation of the entire experiment and includes perceptual processes of reading and encoding cues and feedback from the computer screen and manual processes required to enter responses via a keyboard[1].

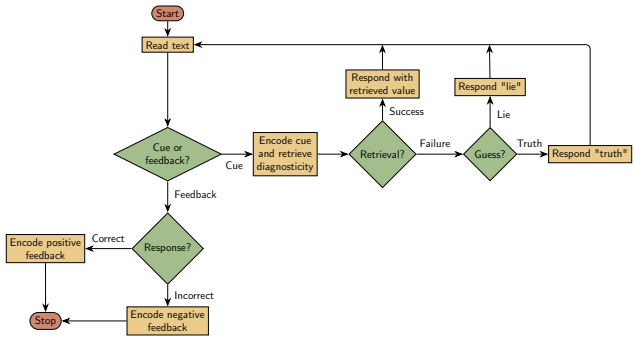
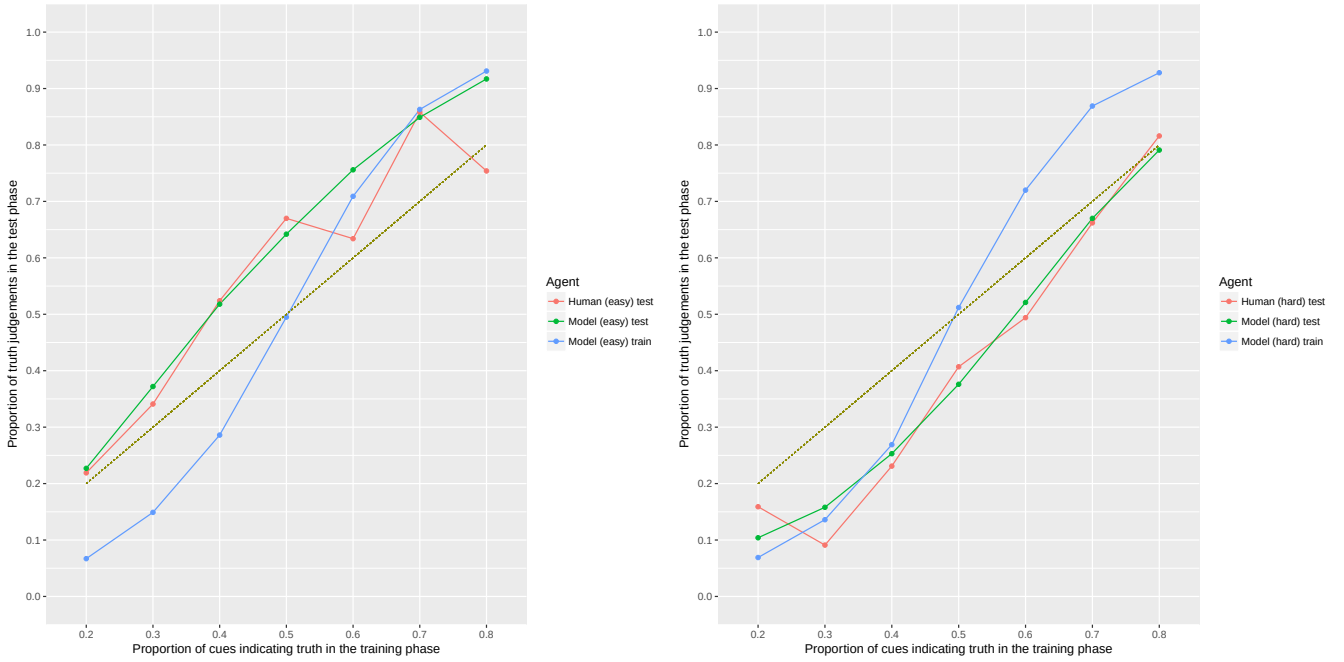


Figure 2: Control flow of the ACT-R model carrying out a trial of the experiment. Rectangles correspond to actions carried out by production rules, diamonds represent junctions in the experiment or model's behaviour.

The control flow of the model executing a single trial (in either the training or test phase) of the experiment is illustrated in Figure 2. 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 as it was explicitly reinforced during the experiment) that easy games are associated with truth telling and hard games with lying.

On each training trial, the model reads one of four cues ('voice', 'face', 'silence' and 'self') on the computer screen and uses it to probe declarative memory for a chunk representing the cue. If the retrieval is successful, the model is able to access associated information about whether the cue is indicative of telling the truth or not and 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 screen which the

[1]The ACT-R model of the experiment is available on GitHub: https://github.com/djpeebles/act-r-lie-detection-model



(a) Easy condition

(b) Hard condition

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

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 training trials are complete, the model is provided with the additional context-general information regarding the difficulty of the game ("easy" or "hard") and then completes the 80 test trials. The test phase is almost identical to the training phase in that the model is provided with the same cues (but this time only 20 of each in random order) and must respond as before. As in the original experiment, no feedback was provided in the test phase.

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 that affect the chunk activation calculation and subsequent retrieval probability were adjusted: the retrieval threshold parameter (τ in Equation 3) was set to 0.4, the activation noise parameter (s in Equation 3 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 4) was set to 0.65. These values are within the typical ranges for ACT-R models and the same parameter values were used for both experiment conditions.

The mean proportion of truth judgements as a function of cue diagnosticity in the training phase from the ACT-R model and human participants are compared for the easy and hard conditions in Figures 3a and 3b 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 conditions was very close, R^2 (easy) $= 0.92$, $RMSD$ (easy) $= 0.08$, R^2 (hard) $= 0.98$, $RMSD$ (hard) $= 0.04$.

Explaining the model's performance

Training phase

To recap, the model's performance in this task depends primarily on the declarative memory chunks created during the training phase which represent the learned associations between each cue and truth and lie responses. As the model proceeds through the training phase, eight chunks in total are created (two for each cue) that represent, in the form of activation, the model's evolving beliefs regarding the strength of association between each cue and the truth and lie responses. Chunk activations are updated through base-level learning during training and determine the likelihood of truth and lie responses to each new cue presentation.

The graphs in Figure 3 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 (when the truth-diagnostic cue > 0.5) or under-estimated (when the lie-diagnostic cue, < 0.5) the proportion of truthful statements as the cue diagnosticity increased (i.e., as the proportion of truth cues moved further away from 0.5 in either direction).

The sigmoid shape of the curve is due to a number of non-linearities in ACT-R's equations, for example in the choice probability function (Equation 3), differences in activation between the competing chunks, and the retrieval threshold and activation noise parameters. This pattern is consistent with data from human probability learning tasks with feedback which shows that people maximise their responses during learning rather than simply matching the environmental probabilities (e.g., Barron & Erev, 2003; Shanks, Tunney, & McCarthy, 2002).

Test phase

When the training phase is completed, 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[2]). Once retrieved, this response bias then becomes an element of the goal which is used as an additional cue in subsequent memory retrieval requests for individuating cues.

The effect of this additional cue is to change the dynamics of the retrieval process and this is where ACT-R's 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 4) to 0, effectively eliminating the possibility of retrieval errors during training. Enabling partial matching in the test phase however ensures that all eight cue chunks in declarative memory enter into the retrieval process and adds the value of the partial matching computation (Equation 4) to the chunk activation calculation (Equation 1).

When a retrieval request containing both individuating and context-general cues is made, each chunk's mismatch penalty is computed according to its dissimilarity to the probe on the two elements (cue name and context-general response bias) and used to revise its activation. The context-general cue decreases the activations of chunks that don't match it, thereby biasing the retrieval of individuating cues in favour of matching chunks. This increases and decreases the probability of a "truth" response in the easy and hard conditions respectively.

The outcome of this process is that, although the chunk with the highest activation is still the one retrieved, the winning chunk may not be an exact match to the elements specified in the retrieval request. This mechanism successfully captures how additional context-general knowledge affects memory retrieval in the test phase and provides a plausible explanation for the response shifts in the two conditions.

The graphs in Figure 3 reveal that the effect of additional context-general information differs across the individuating cue diagnosticity levels, with 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. To explain why this happens, consider again the easy condition in Figure 3a where the context-general information suggests that most people will tell the truth. When the cue is present on only 20% of 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, due to partial matching, the activation of the "lie" individuating cue chunks is reduced, 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.

Discussion

Lie detection research consistently shows that people are truth-biased (for a meta-analysis see C. F. Bond & DePaulo, 2006). In contrast, there are relatively few empirical demonstrations that people can be lie biased, leading some to consider truth bias as arising from a default form of processing and lie bias as something that may result from an additional trigger or further processing (cf. Gilbert, Krull, & Malone, 1990; Levine, 2014).

This assumption is not part of the ALIED theory however, which instead argues that the observed patterns of behaviour are a result of the relative weighting of individuating and context general information during the decision-making process, and that therefore truth and lie biases both arise from the same processing operations.

Theories of lie detection are typically relatively loosely specified, high-level, verbal descriptions, and while ALIED has a credible Bayesian account of how lie-truth judgements are made, it has no cognitively plausible account of the processes that achieve this. The computational model presented here addresses this shortcoming by providing a mechanistic account of how the integration of individuating and context-general information can occur in terms of the storage and subsequent retrieval of declarative chunks. In doing so, the model brings to bear the many well established cognitive principles, mechanisms and constraints of the ACT-R cognitive architecture 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.

The model supports the ALIED theory by showing how the effect of context-general information on people's responses is related to their existing beliefs about the prevalence of the behaviour; context-general information has a greater effect

[2]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.

on the decision as the diagnosticity of the individuating cue reduces. According to the model, if you have learned that an individuating cue is weakly diagnostic of a particular behaviour, then receiving context-general information that contradicts the association will have a great effect on your subsequent judgements. In contrast, if you have learned that an individuating cue is strongly diagnostic of a particular behaviour and you subsequently receive context-general information that further supports that knowledge, then it will have only a very small effect on your judgements because the association is already very strong.

This model also supports ALIED's position on one of the key debates in judgement and decision-making research: whether people employ the compensatory strategy of integrating multiple cues when making judgements or instead adopt the non-compensatory approach of ignoring most cues and using only one (Gigerenzer & Todd, 1999; Newell & Shanks, 2003). By demonstrating how individuating and context-general information can be integrated, the model substantiates ALIED's compensatory approach that uses all available information when forming a judgement.

In attempting to ground an existing higher-level computational (i.e., purely Bayesian) account of lie-truth judgement formation in a formal, algorithmic and representational information processing theory, we believe that this endeavour exemplifies a core aim of cognitive science to develop explanations of cognitive phenomena at multiple, mutually constraining, levels of description (Marr, 1982; Cooper & Peebles, 2015) and that the rigour and precision brought about by formal modelling can only benefit theories of lie detection.

References

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

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, 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.

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.

Estes, W. K. (1972). Research and theory on the learning of probabilities. *Journal of the American Statistical Association*, *67*(337), 81–102.

Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.

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.

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, 7). Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, *137*, 643–59.

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.

Luke, T. J. (2019). Lessons from Pinocchio: Cues to deception may be highly exaggerated. *Perspectives on Psychological Science*, *14*(4), 646–671.

Marr, D. (1982). *Vision*. W.H. Freeman.

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.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207–238.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 53–65.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of experimental psychology: General*, *115*(1), 39.

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.

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.