No evidence for unconscious lie detection: A significant difference does not imply accurate classification.

Volker H. Franz,[1*] Ulrike von Luxburg[2]

[1]Department of Psychology, University of Hamburg,

von Melle Park 5, 20146 Hamburg, Germany

[2]Department of Computer Science, University of Hamburg,

Vogt–Koelln–Str. 30, 22527 Hamburg, Germany

[*]To whom correspondence should be addressed; E-mail: volker.franz@uni-hamburg.de.

The R–code of all analyses can be found at https://osf.io/7825t

ten Brinke, Stimson, and Carney (2014; tBSC14) reported that unconscious processes detect liars better than conscious processes, the latter being typically close to chance (~54% correct; Bond and DePaulo, 2006). tBSC14 concluded that "although humans cannot consciously discriminate liars from truth tellers, they do have a sense, on some less-conscious level, of when someone is lying" (p. 1103) and argued that "accurate unconscious assessments are made inaccurate either by consolidation with or correction by conscious biases and incorrect decision rules" (p. 1104). In short: tBSC14 suggested that humans unconsciously know quite well whether somebody is lying and that it is conscious deliberations that render these accurate unconscious assessments inaccurate.

Such conclusions could potentially have far reaching practical consequences. For example, based on these conclusions, we could advise jurors and eye-witnesses at court to trust mainly their intuition and avoid conscious deliberations. However, this is a dangerous road to travel. There are well-documented cases in which eye-witnesses erred in their intuitive judgment and only conscious deliberation led to the truth (Loftus, 2003). Therefore, before concluding that "accurate lie detection is, indeed, a capacity of the human mind, potentially directing survival- and reproduction-enhancing behavior from below introspective access" (tBCS14 p. 1104), we should make sure that there is strong scientific evidence. While the plausibility of tBSC14's data has already been challenged (Levine and Bond, 2014; but see ten Brinke and Carney, 2014), we show that tBSC14's statistical reasoning is flawed and that a more appropriate analysis of their data does not provide evidence for accurate unconscious lie detection.

In tBSC14[1], participants watched 12 videos of interrogations, 6 showing a liar, 6 a truth-teller (participants were not told who was liar/truth-teller). Then participants performed two tasks. In the direct (conscious) task they saw pictures of the suspects, classified them as liars or truth–tellers and performed poorly (percent-correct: PC = 49.6% chance-level: 50%). In the indirect (unconscious) task, the pictures of the suspects ("primes") were masked, such that they could not be perceived consciously. Participants sorted well visible words ("targets") like "deceitful" or "honest" into the categories "lie" or "truth". Participants were significantly faster if prime and target were congruent (e.g., the word "deceitful" preceded by a picture of a liar) than if they were incongruent. Based on this significant congruency effect, tBSC14 concluded that there are "accurate unconscious assessments" (p. 1104) of liars vs. truth–tellers in the indirect (unconscious) task; better than the chance–level performance in the direct (conscious) task.

However, this conclusion is flawed. The test for a significant congruency effect is only concerned with the question whether a 'true' difference in reaction times (RTs) exists in the population, no matter how big. We can only conclude from this effect that *some* classification

[1] We concentrate on Exp. 2 of tBSC14, as only this experiment presented unconscious stimuli. Exp. 1 investigated whether consciously well-visible pictures of the suspects had indirect effects on another task. Nevertheless, our critique also applies to Exp. 1, because there tBSC14 also infer good classification from a significant RT-difference for these indirect effects, while our analysis --- using the method described under (i) in our main text --- results in PC = 51.1% (SD = 3.71%), which is again clearly below the 54% described as "detection incompetence" by tBSC14 (p. 1098).

of the suspects has happened, but we do not know *how accurate* this was and whether it was more accurate than in the direct task. To make the claim that the RTs are evidence for good unconscious classification, tBSC14 would have needed to show that the RTs can be used to classify whether the suspects were truth-tellers or liars. Only then is it possible to compare this 'indirect classification accuracy' to the accuracy in the direct task.

How can such an indirect classification be performed? Due to the experimental design, classifying suspects as truth-tellers vs. liars is equivalent to classifying trials as congruent vs. incongruent (e.g., if a trial is classified as congruent and the well visible target was "deceitful", then the suspect is classified as liar). Because tBSC14 argued that the congruency effect is evidence for accurate unconscious classification (fast RTs in congruent trials, slow RTs in incongruent trials), all we need to do is find an appropriate threshold t and classify all trials with RTs smaller than t as congruent and trials with RTs larger than t as incongruent.

We performed this classification on the tBSC14 data with different choices of thresholds: (i) Under the assumption that RTs follow normal or lognormal distributions (Ulrich & Miller, 1993), the threshold that leads to the best expected accuracy in a design with equal number of congruent and incongruent trials is the median of the RTs (e.g., MacKay, 2003; p. 190). Therefore, we classified the trials of each participant using her/his median RT as threshold, computed the accuracy over all trials of the participant, and averaged the accuracies across participants. This results in an average accuracy of PC = 50.6% (SD = 2.65%). (ii) Avoiding assumptions about the RT distributions, we selected a threshold according to the standard procedures of machine learning (Shalev-Shwartz, Ben-David, 2014): We randomly split the trials of each participant into equal-sized training and test sets (other split sizes led to similar results). On the training set, we determined the threshold that leads to the best accuracy and used it to classify the test set. We repeated this procedure ten times with different random splits of the data for each participant. This leads to an average accuracy of PC = 49.5% (SD = 2.60%) (iii) To construct an overly optimistic upper bound---that is, the highest accuracy that possibly can be achieved on the given data---we evaluated the accuracy of all possible thresholds over all trials of each participant, determined the best result, and averaged the obtained accuracies across participants. This results in an accuracy of PC = 53.7% (SD = 1.99%), meaning that for these data, no possible classifier exists with an accuracy larger than 54% — the very number that was interpreted as "detection incompetence" by tBSC14 (p. 1098). In short: the classification accuracy in the indirect task is just as poor as in the direct task and can for all practical purposes be considered as being at chance. There is no evidence for accurate unconscious assessments.

Although this result seems clear and consistent, one might ask whether it is fair to assess indirect classification performance with RTs from single trials. One might argue that it would be better to average the RTs of multiple trials, thereby reducing measurement error and possibly improving accuracy. It is known from machine learning that such procedures can under certain conditions improve accuracy (e.g., Bühlmann, 2004). We tested this idea in two variants: (iv) For each participant we averaged all trials related to each suspect, classified these averages using the median across all suspects as threshold, and averaged the accuracies over all participants. This resulted in PC = 51.9% (SD = 16.30%) (v) We averaged the RTs related to each suspect across all trials and all participants, thereby including all available information for the classification. This resulted in PC = 50.0%. In short, even if we combine

RTs from multiple trials and multiple participants, there is no evidence for better accuracy in the indirect task than in the direct task.

To understand why a significant difference does not indicate accurate classification, consider an intuitive example. Suppose we tried to *classify* individual adults as female versus male based on their weights. The weight distributions for the two genders overlap a lot, so we would perform poorly[2]. On the other hand, if we performed a standard *significance test*, we would form two groups of same-gender adults and compared their mean weights. If the groups are large enough, we can easily obtain a significant difference. This shows how a significant difference can coexist with essentially chance-level classification accuracy and that we cannot infer good classification accuracy of the individual adults from the fact that the group-means are significantly different. Good classification requires more, namely a clear separation of the female/male weight distributions at the level of the individual adults (or the RT-distributions at the level of the individual suspects in tBSC14). For tBSC14 we have shown that --- no matter how we aggregate the individual trials --- the RT-distributions of liars vs. truth-tellers always overlap so heavily that no good classification can be obtained.

To conclude, the data of tBSC14 do not provide any evidence for accurate unconscious lie detection. A significant difference in the indirect task does not indicate accurate unconscious classification. For a meaningful comparison of the classification accuracies in the indirect and

[2] We thank an anonymous reviewer for suggesting this example. The average male/female weight difference is app. 13 kg; SD in each group app. 33 kg (McDowell, et al., 2008). This results in an effect size of d=13/33=0.4 and a classification accuracy of PC = 58%. This is small, but still well above the accuracy we found for tBSC14.

direct tasks, they have to both measure just that: classification accuracy. We thank tBSC14 for making their data available (Eich 2014). This allows for a rapid self-correction of science without going through lengthy replication attempts first --- which easily can take years, if they are successful at all (Ioannidis, 2012).

Acknowledgments

We thank Gilles Blanchard and Frank Rösler for comments. UvL was supported by DFG / grant LU1718/1-1 / Research Unit 1735 “Structural Inference in Statistics: Adaptation and Efficiency”.

Author contributions

VHF discovered the methodological flaws and reanalyzed the data of tBSC14 in R, UvL verified all arguments and re-implemented the analyses independently in Matlab. Both wrote the paper.

References

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

Bühlmann, P. (2004). Bagging, boosting and ensemble methods. In: Handbook of Computational Statistics: Concepts and Methods (eds. Gentle, J., Härdle, W. and Mori, Y.), pp. 877-907. Springer.

Eich, E. (2014). Business not as usual. *Psychological Science*, *25*(1), 3–6.

Ioannidis, J. P. A. (2012). Why science is not necessarily self–correcting. *Perspectives on Psychological Science*, *7*, 645–654.

Levine, T. R., & Bond, C. F. (2014). Direct and indirect measures of lie detection tell the same story: A reply to ten Brinke, Stimson, and Carney (2014). *Psychological Science*. (published online June 25, 2014)

Loftus, E. (2003). Science and society — Our changeable memories: Legal and practical implications. *Nature Reviews Neuroscience*, *4*, 231–234.

MacKay, D. J. C. (2003). *Information theory, inference & learning algorithms.* New York, NY, USA: Cambridge University Press.

McDowell, M. A., National Center for Health Statistics (US), et al. (2008). *Anthropometric reference data for children and adults: United States, 2003–2006.* US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Shalev-Shwartz, S. & Ben-David, S. (2014) Understanding machine learning: from theory to algorithms. Cambridge University Press.

ten Brinke, L., Stimson, D., & Carney, D. R. (2014). Some evidence for unconscious lie detection. *Psychological Science*, *25*(5), 1098–1105. (published online Mar 21, 2014)

ten Brinke, L., & Carney, D. R. (2014). Wanted: Direct comparisons of unconscious and conscious lie detection. *Psychological Science*. (published online Aug 15, 2014)

Ulrich, R., & Miller, J. (1993). Information–processing models generating lognormally distributed reaction–times. *Journal of Mathematical Psychology*, *37*, 513–525.