



Veritaps

Truth estimation from mobile interaction

Mottelson, Aske; Knibbe, Jarrod; Hornbæk, Kasper

Published in:

CHI 2018 - Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems

DOI:

[10.1145/3173574.3174135](https://doi.org/10.1145/3173574.3174135)

Publication date:

2018

Document version

Publisher's PDF, also known as Version of record

Document license:

[CC BY](#)

Citation for published version (APA):

Mottelson, A., Knibbe, J., & Hornbæk, K. (2018). Veritaps: Truth estimation from mobile interaction. In *CHI 2018 - Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems: Engage with CHI* [561] Association for Computing Machinery. <https://doi.org/10.1145/3173574.3174135>

Veritaps: Truth Estimation from Mobile Interaction

Aske Mottelson, Jarrod Knibbe, Kasper Hornbæk

Department of Computer Science

University of Copenhagen

DK-2300 Copenhagen, Denmark

{amot, jarrod, kash}@di.ku.dk

ABSTRACT

We introduce the concept of Veritaps: a communication layer to help users identify truths and lies in mobile input. Existing lie detection research typically uses features not suitable for the breadth of mobile interaction. We explore the feasibility of detecting lies across all mobile touch interaction using sensor data from commodity smartphones. We report on three studies in which we collect discrete, truth-labelled mobile input using swipes and taps. The studies demonstrate the potential of using mobile interaction as a truth estimator by employing features such as touch pressure and the inter-tap details of number entry, for example. In our final study, we report an F_1 -score of .98 for classifying truths and .57 for lies. Finally we sketch three potential future scenarios of using lie detection in mobile applications; as a security measure during online log-in, a trust layer during online sale negotiations, and a tool for exploring self-deception.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation: (e.g., HCI)

Author Keywords

Lie detection; Polygraph; Dishonesty; Deception; Mobile Input; Smartphones



INTRODUCTION

We frequently lie, whether to advance our own aims or to protect others [13]. Consequently, we are also subject to many lies. Though this provides ample opportunity for practice, humans are only slightly better than chance at detecting lies and exhibit a positive bias in assessing the truth [3]. This deficiency has led to a century-long interest in lie detection. Visual, vocal, and physiological features of communication have all been explored [25], but, to date, natural language processing leads the way in identifying lies in digital communication. Through linguistic, psychological, and personal features, research has demonstrated success in classifying dishonest prose such as spam and deceptive reviews [18].

However, writing prose covers only a small part of our digital input. As we increasingly use our mobile devices for digital communication, our input also comes to include individual taps and swipes, such as button clicks, checkbox selection, and number entry. This leaves much digital activity open for deceptive behaviour with our approximately chance-level truth assessments. To this end, we explore a content-agnostic approach to mobile lie detection, ignoring the content of the input (i.e., input text), and enabling lie detection across a much wider spectrum of input.

To enable content agnostic lie detection, we draw on research demonstrating that a variety of information is hidden in the details of mobile input, such as stress [14], boredom [19], and affective states [16]. We explore whether dishonesty and deception are similarly hidden. Research suggests the presence of physiological responses to lying, such as increased hand/finger activity [25]. We hypothesize that these responses, although subtle, can be identified through smartphone sensors.

We test this across three crowdsourced smartphone studies. In Study I, we verify that lying on a smartphone exhibits similar behavioural cues to lying in conversation, and that this can support the separation of honest and dishonest responses. In this study, following the paradigm of Williams et al. [26], participants are instructed to tell the truth or lie, and the cues are identified through response time. Visible trends in other sensor data, such as input speed, motivate a second study using a more natural, spontaneous lying paradigm. The results from Study II show that acceleration, rotation, and inter-key-press duration can drive lie classification with an F_1 -score of .77. Finally, we validate this result with an additional study, where we explore our identified features from Study II with a dice paradigm. In Study III we show 98% precision, and 97% recall for truths ($F_1 = .98$), and 65% precision and 59% recall for lies ($F_1 = .57$).

Following the studies, we sketch the concept of Veritaps, an additional layer of communication to assist mobile device users in their own lie detection accuracy. Veritaps enables users to automatically share a belief state indicator alongside their input. With high accuracy, Veritaps can label truthful input . We can also label inconclusive taps and swipes , informing the user that they should use caution or seek further information in assessing this input. We illustrate the opportunities of Veritaps across a range of example scenarios, including (i) automated lie analysis when completing online



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI 2018, April 21–26, 2018, Montréal, QC, Canada

Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-5620-6/18/04...\$15.00

DOI: <https://doi.org/10.1145/3173574.3174135>

forms, (ii) increased richness of trust in mobile messaging, and (iii) as a prompt to prevent self-deception.

We present the following contributions:

1. An exploration of lie detection across mobile devices, regardless of the input content.
2. Results from three studies, showing dishonesty affects user interaction with mobile devices.
3. Convincing classification rates of lies in mobile entry, potentially improving a user's ability to judge the veracity of others' mobile input.
4. Veritaps: a concept that allows users to share their belief states with other users and applications.

RELATED WORK

Our work explores lie detection in mobile input. Specifically, we are interested in classifying lying through sensor data, rather than actual user input, in order to make lie detection available for a broader range of mobile input types.

Classifying Behaviors from Mobile Sensors

Research shows that complex cognitive and affective phenomena can be inferred using commodity sensors. The linearity of swiping, for example, correlates with emotions during gameplay [6]. Similarly, speed, acceleration, and precision in touch input are indicative of affective states [16]. Mobile activity can also provide insight into a user's thinking, where app activity, battery level, and time of day are strong correlates of boredom [19].

Based on the idea of using mobile sensor data to support real-time inferences about human cognition, we explore indicators of lying in mobile sensor data.

Lie Detection

Deceptive behavior carries a range of verbal and nonverbal cues, and research has explored various strategies for using such cues to uncover deception. Among the most famous of these strategies is the polygraph. Polygraphs examine the subject's heart rate, galvanic skin response, respiration, and blood pressure as physiological markers of deception. It is widely accepted, however, that the interpretation of physiological responses and, thus, polygraph results, is 'a complex clinical task' [20]. The debate continues regarding the accuracy and applicability of polygraph testing. For example, a large body of research assessing the validity of polygraph techniques uses 'mock crime' scenarios, which inherently lack the consequences of real crime scenarios, and thus call into question the validity of their results [4].

Other work has provided evidence on verbal, visual, and vocal cues to deception (e.g., [25]). Zuckerman et al. [27], for example, suggested that lying is a more cognitively complex task than telling the truth, requiring liars to formulate internally and externally consistent events. These greater cognitive challenges result in greater response latency, more hesitations, increased pupil dilations, and fewer heartbeats.

More recently, research has shown that lies include more complex imagery, longer words, and a greater number of pauses

than truths [1, 11, 25]. This has led to automatic lie detection in text. Mihalcea et al. [15], for example, reported 71% accuracy in lie detection across three text corpuses. Ott et al. [18] used linguistic features (such as average word length or misspelling rate), psychological features (such as social or emotional clues), and personal features (such as references to money or religion), to classify spam and deceptive reviews.

Lying has also become a subject of exploration in crowdsourcing studies. Gino et al. [7] asked participants to report the outcome of random events (such as dice rolling or coin tossing). They identify lying across all of the input based on the deviation from the expected mean, offering an insight into lying across an entire study.

Opportunities for Lie Detection in Smartphones

While current research points towards physiological- and content-based lie detection, a common and robust strategy to lie detection has yet to be derived. We look for a commodity, content-agnostic approach to lie detection, that can be used to identify deception in basic mobile input; taps and swipes. We hypothesize that the bodily influences of deception can be measured using sensors available in consumer smartphones, making commodity lie detecting feasible.

STUDY I: SIMPLE LIES

Research shows that lying takes longer than telling the truth [24]. A common explanation is that the construction of a lie forces additional cognitive load compared to telling the truth, and thus causes longer response times.

Study I had two goals: (i) to establish whether lying through touch interaction on mobile devices produces results that are consistent with verbal responses in a laboratory, and (ii) to demonstrate the feasibility of separating honest and dishonest activity using mobile interaction data. We ran a mobile crowdsourced study of an experimental paradigm originally developed Williams et al. [26]. The participants were asked to either lie or tell the truth about the color of the screen, using common mobile UI elements. This paradigm offers an experimental procedure for studying both instructed and voluntary lies, while maintaining an even distribution of lies and truths. This provides a simple method for initially investigating differences in interaction patterns between telling lies and truths using mobile devices.

Task

The experiment progressed as a series of random-ordered trials, each beginning with an objective: TRUTH, LIE, or CHOICE. Directed trials (where participants were told to LIE or tell the TRUTH) presented a continue button, and the CHOICE trials had two buttons prompting the user to choose between lying or telling the truth (see Figures 1a and 1b).

Upon establishing the objective, participants were presented with a screen with a red or blue background. The participant's objective was written as a visual reminder at the top of the screen. The UI controls (button or slider) appeared at the bottom of the screen (see Figure 1d). The order of UI controls was randomised. Participants then had to activate the correct UI control according to (a) the color of the background, (b)

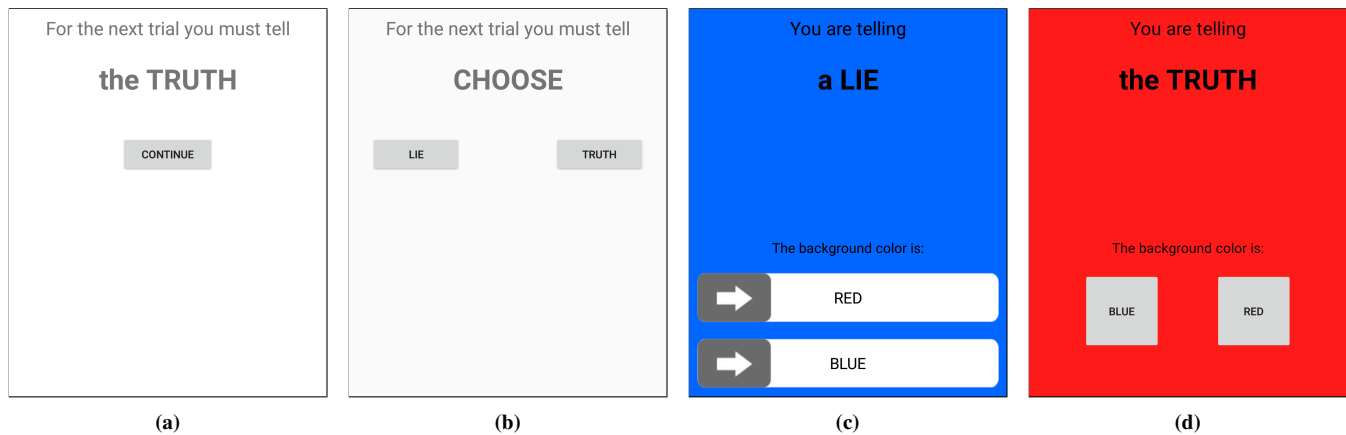


Figure 1: Example screens from the experimental application used for Study I. The figures show (a) a directed trial pre-screen, (b) a choice trial pre-screen, (c) a trial with sliders - where the participant should slide the ‘RED’ slider, in this case, and (d) a trial with buttons - where the participant should select the ‘RED’ button, in this case.

the text on the UI control, and (c) the trial’s objective (see Figure 1c). Trials were separated with a white screen for 1s. Participants were instructed to respond as quickly and accurately as possible. Participants were asked to lie and tell the truth half of the time each in the choice condition.

The task was similar to the original study [26], with the exceptions that: (i) instead of a lab-based study, participants were recruited online and completed the experiment on their own phones, (ii) vocal responses were replaced with selections using buttons or sliders, and (iii) the colors were changed to be visible for color blind (red and blue, instead of red and green).

Design

The study used a $2 \times 2 \times 2$ within-subjects design. The independent variables were honesty of response (lie vs. truth), type of instruction (directed vs. choice), and UI (button vs. slider). The dependent variable was response time. Each participant did a total of 192 trials, with 64 from the directed to lie condition, 64 from the directed to tell the truth condition, and 64 from the choice condition. In half of the trials participants responded by tapping a button, and the other half by dragging a slider. The order of trials was randomized. The study took 15 minutes on average.

Participants

We recruited 100 participants from Mechanical Turk, aged 19-59 ($M = 31$), 33 females. Participants installed our experimental application on their own Android smart phones (Android version ≥ 6.0), and followed onscreen instructions. Participants were reimbursed with \$2.00 USD.

To ensure only qualified participation, we (i) required 90% HIT approval, (ii) had participants pass a qualification test about the task before starting the HIT, (iii) stored a unique device ID to avoid multiple participations, and (iv) ensured that app and MTurk HIT participation count matched.

Data

Of the 100 participants, ten never lied and one never told the truth in the choice trials, and were therefore removed. The

remaining 89 participants performed a total of 16,671 trials. We removed (i) the first 10 trials per participant as warm-up rounds, (ii) 460 trials (2.9%) that lasted more than 4 seconds, and (iii) 623 incorrectly answered trials (3.9%). The analysis is made on a resulting data set comprising 14,788 trials.

Results

Lying took longer than telling the truth, both when answering with a button and a slider, and when being told whether to lie or when given the option to choose (see Figure 2).

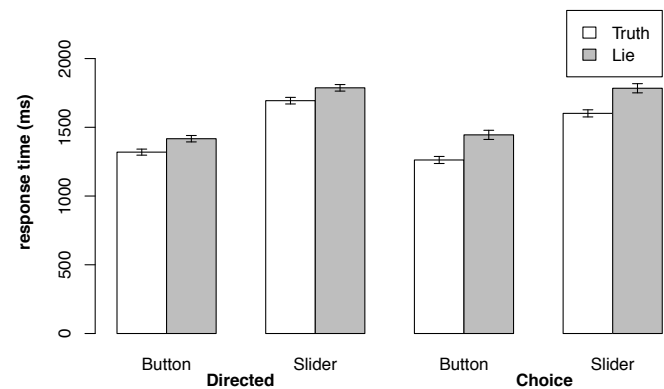


Figure 2: Response times for telling truths and lies using two different UIs both for the directed and choice conditions. Error bars show 95% confidence intervals. It took on average longer to tell a lie for both UIs.

Except for directed trials with the slider, participants took significantly longer when lying (on .5, see Table 1). The effect was larger when participants chose whether to lie or not.

Summary

The results show that constructing a lie is a cognitively harder task than simply telling the truth, reflected by the increased response time when participants were asked to lie about the background color of a mobile UI. This corroborates the findings of Williams et al. [26]. While the data sourced do not allow for an effective binary discrimination of truth and lies per

Condition	UI	F	df	p		Cohen's d
Directed	Button	5.05	176	.026	*	0.34
Directed	Slider	3.03	176	.084		0.26
Choice	Button	8.94	176	.003	**	0.45
Choice	Slider	7.00	176	.009	**	0.40

Table 1: Results from an ANOVA comparing truths and lies. Lying caused significantly longer responses for close to all conditions, with the largest effect for choice trials.

entry, the results imply the feasibility of separating honest and dishonest activity using mobile interaction data, specifically timing in this case. We further investigate if this difference can be observed for other parameters in Study II-III.

Although not statistically significant, we observed that slider interactions were performed faster (by 4.4%) when telling the truth; $F(1, 175) = 2.16, p = 0.14$. Although mean response times pertaining to honest and dishonest behaviour were distinguishable in this study, we were keen to explore whether additional features become more prominent with (a) spontaneous lying, and (b) a more natural distribution of truths and lies (i.e., [23]).

STUDY II: ULTIMATUM GAME

We ran a second study to analyze natural deceptive behavior. We employed a mobile version of the Ultimatum game, a commonly studied task in behavioral economics. In this variant the participants are offered an incentive to lie.

In the Ultimatum Game, the first participant (the proposer) receives a sum of money and proposes a division of the money between themselves and the second participant (the responder). The responder then either accepts the division, giving both participants the proposed funds, or rejects it altogether resulting in no payout for any of the participants. In the variant developed by Besancenot et al. [2], which we use, the proposer is given the opportunity to lie about the amount of allocated funds. Therefore, for each trial, the proposer to the responder (i) declares the amount that was allocated and (ii) proposes a division. This provides a monetary incentive for the participant to understate the provided funds, enabling the study of naturally occurring dishonest behavior.

Participants

We recruited 41 participants from the USA from Mechanical Turk, aged 22–63 ($M = 33$); 18 females, 36 right-handed. Participants were told that they were taking part in an economics experiment. Participants installed our experimental application on their own Android smart phones (Android version ≥ 6.0), and followed onscreen instructions. Participants were reimbursed \$1.00 USD, in addition to the money collected throughout the experiment, which ranged from \$1.74–\$4.52 ($M = \$3.35$). The experiment took at most 10 minutes. We employed the same qualification standards as for Study I.

Design

Each participant did 10 trials of proposals, excluding a warm-up round. The independent variable was funds allocated (25–99¢). The dependent variables were declared allocated funds

and the proposed division of money. Additionally, throughout the trials, the mobile application collected data related to interaction with the UI using touch, pressure, accelerometer, and gyro sensors.

All participants had the role of the proposer. Participants were paired with an AI in the responder role, presented as the human worker *Mary* with a fictional worker ID. Mary would simulate human latency when responding to proposals, and would accept or reject proposals based on the available heuristics and basic economic and moral behavior: greed was punished while fair divisions were rewarded.

The AI was implemented as nine simple steps that would accept offers deemed favorable, or refuse offers that were either directly too low ($< 25¢$), or too unfair ($3P < F$), where P is the proposal, and F the declared funds. The AI would also reject offers when they repeatedly showed lower declared funds than expected from a random sample. If all steps passed, a 75% chance of acceptance was returned, to introduce some degree of unpredictable behavior.

Mary did not know whether the participant was in fact honest or dishonest, but instead reasoned based on the distribution of declared allocations from all trials. Mary accepted 86% of all proposals made (very similar to human behaviour observed in other of the Ultimatum game studies [17]).

Procedure

Upon installing and opening the experimental application, participants were informed that they were playing the proposer and were paired with our AI (under the guise of another crowdworker). For each of the 11 rounds, an amount of US cents between 25 and 99 were allocated to the participant. The participant would then, using num-pads, first state the amount of allocated funds (about which they could lie), and then propose a division (see Figure 3a).

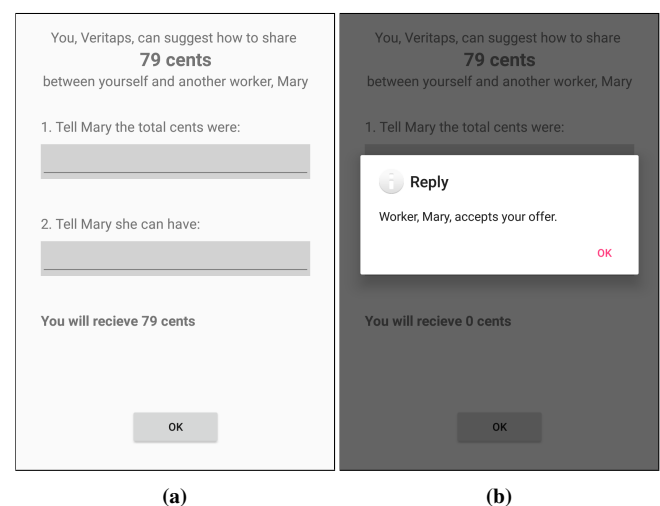


Figure 3: The experimental application used in Study II. Figures show (a) the screen where participants declare the allocated funds and propose a division, and (b) a positive response from the AI, Mary, acting as another human worker.

Shortly hereafter, the participant would receive a notification of whether the responder (Mary) had accepted the division (see Figure 3b). Participants collected money throughout the trials, and were paid according to their final score to create a monetary incentive to lie.

Data

An entry was defined as the window of time between when the proposal screen would appear (see Figure 3a), and until the participant hit *OK*. The resulting data set comprised 41 participants and 410 entries.

Participants lied about the available funds on average 35.1% of time; this was most prominent when the allocated funds were high. Seventeen participants never understated the available funds (59% lied at least once). Three participants understated at every entry. Participants discounted the actual endowment by 17.4% on average. The crowdsourced participants appear more loyal than laboratory participants (Besancenot et al. found that on average 88.5% of the proposers discount the actual endowment by 20.5% [2]); in this study we observe that 41% of the participants never lied at all, consistent with some feedback we received, such as:

It seemed fair to me to split the money evenly. I don't believe in dishonesty so I did not want to lie

– Crowdfworker

Classification

We built a binary truth/lie classifier based on the data obtained. We defined a lie as an entry where the declared funds were lower than the allocated.

Choice of Classifier

We tried a range of classification algorithms, including ensemble methods. An SVM with a radial basis function kernel provided the most promising classification accuracy. Hyper parameters were selected using grid search. The classifier was developed in Python using the ML library Scikit-learn.

Feature Generation

Features were chosen based on previous work in classification of human factors using mobile devices (e.g., [16, 19]), such as speed, precision, rotation, and acceleration (sampled at 50 Hz). We also included features from empirical observations of deception (e.g., [25]), such as immediacy and response length.

Feature Selection

We clustered our features in related groups (see Table 2), and handpicked the effective predictors for truth classification. The feature groups *acceleration* and *num-pad* presented the most viable features for classifying truths and lies, and were thus shown in our final classifier (i.e., manual feature selection).

Performance

We measure how well our predictor works, by reporting the average binary F_1 -score obtained over a randomized 5-fold cross validation. The F_1 -score can be interpreted as a weighted average of the precision and recall, where an F_1 -score reaches

Feature Group	Features	Description
Timing	immediacy response	t before first event entry duration
Finger size	touch area	finger contact size
Num-pad	key dynamics hold-time tap precision	see [5] button hold-down time distance to target center
Button clicks	hold-time click area backspaces	button hold-down time quadrant activated number of deletions
Done-button	taps precision hold-time pressure click area	number of times distance to target center button hold-down time screen pressure quadrant activated
Acceleration	x -, y -, and z	a for all axes
Rotation	α -, β -, and γ	ω around all axes
Signal Magnitude	$\sqrt{x^2 + y^2 + z^2}$	for both a and ω

Table 2: Feature groups and specific features for each group.

its best value at 1 and worst score at 0, and is defined as:

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

where precision and recall relate to true positives (TP), false positives (FP), and false negatives (FN) as:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Using a randomized 5-fold cross validation we obtain precisions of 81% and 66% for truths and lies respectively. The rates for recall are 88% and 52% for truths and lies respectively. This yields an average F_1 -score of .77; .81 for truths and .66 for lies. These performances are well over both chance level (.50), the baseline (.65), and human performance [3].

How Lies and Truths Differ

Next we report on how interaction with the mobile UI differed between honest and dishonest entries, in particular features that varied with the honesty of the interaction. We inspect the distribution of features using density plots: blue areas represent the prevalence of honest entries; red areas represent dishonest entries (those with deflated declared funds). Note that a single feature seldom alone is enough to support classification. Instead combinations of features make up the decision, which is not clear from a single feature's distribution.

Acceleration

We observed that a low mean acceleration was most frequent among honest entries. This suggests that honest entries resulted in less hand movement by the device-holding hand (their non-dominant hand). This was true both on the x -axis, and the z -axis (see Figure 4). This follows findings from an existing study of non-phone deceit [25], which showed that dishonesty causes increased hand/finger activity.

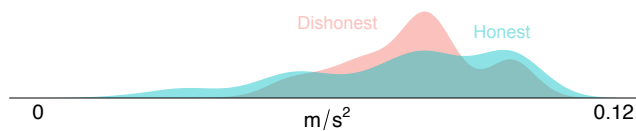


Figure 4: Mean acceleration on the z -axis during an entry. Entries towards the low spectrum are predominately honest.

Num-pad

For each entry an amount of cents between 25 and 99 was allocated, requiring participants to input a two-digit number in the declared input field using a num-pad. We observe that the duration between the first key event and the second key event is higher for dishonest entries (see Figure 5). This suggests that participants decide whether to lie, and by how much, per individual digit, rather than per input.

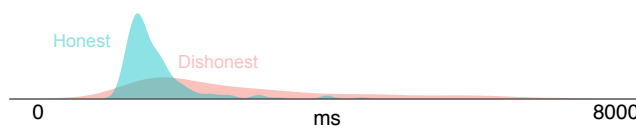


Figure 5: Duration between first and second num-pad key event. Truthful entries show shorter durations between the first two num-key presses.

Our num-pad dialog implementation could be dismissed by tapping outside of the num-pad area (instead of clicking ‘OK’). Additionally, if, after having entered a number, the participants decided to correct their entry, additional ‘OK’ taps could be performed. The more taps on the ‘OK’ button in the num-pad, the more likely an entry was to be honest (see Figure 6); we almost exclusively observe dialog dismissal amongst dishonest entries, and we almost only find honest entries for high number of taps on ‘OK’¹.

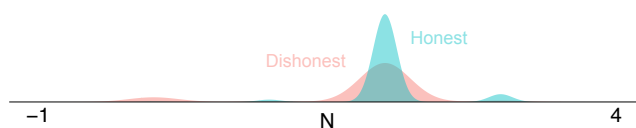


Figure 6: Total number of taps on the ‘OK’ button in the num-pad. Honest entries tend to contain more taps on ‘OK’; almost only dishonest entries closed the dialog without confirming ‘OK’; almost only honest entries reopened the dialogue and pressed ‘OK’ again.

Summary

Study II shows that the way people interact with their mobile UI can change with the level of honesty of the action. Specifically, movement of the phone (acceleration) and num-pad interactions varied. This increases our confidence in the feasibility of using sensor data to estimate the veracity of input. We built a classifier based on smartphone sensor data and achieved an average F_1 -score of .77. This classification accuracy shows that mobile sensor data can be a promising path towards lie

¹This may also suggest that honest users lied initially, before correcting their input to the truth. Dishonest users may show reluctance to ‘confirm’ their lie, and thus avoid pressing ‘OK’. Further research is needed to verify this behaviour.

detection. To validate these results, and to assess whether the results generalize to other settings, we ran a third study.

STUDY III: YATZY GAME

Both Study I and Study II showed that we can observe differences in interaction data between lies and truths using mobile UIs. In Study I, participants were instructed to lie and response time was the only distinguishing feature. In Study II, participants were made aware that they could lie without punishment, resulting in a higher proportion of lies than expected in everyday interaction [23]. From this study, a wider spectrum of mobile input became valuable features for classification.

In order to validate the classification results from study II, we ran a third study. This study still facilitated spontaneous lying, but made no reference to dishonesty in its description. The study required participants to play a dice-based game on a mobile device, inspired by a widely used experimental task in dishonesty research. The task supported spontaneous lying, and allowed for automatic labeling of discrete trials as either honest or dishonest. The participants were rewarded based on their reported score, thereby making lying profitable. We did not encourage participants to lie, and given that all participants passed an initial qualification test about the rules, we can assume that participants were aware of their wrongdoings. Overstating scores could provoke both moral dissonance and fear of not having the crowdwork approved (and thus not getting paid); we hypothesize that this manifests itself in the participant’s mobile interaction.

Task

A commonly used task in studying deceit and dishonest behavior requires participants to report on the outcome of randomized events such as rolling a die, or tossing a coin (see [12] for an overview). To encourage lying, participants are rewarded relative to the reported outcomes. The actual outcomes of the events are only known to the participants. This paradigm supports inferences about deceit across all reports (based on deviation from the expected mean) but the individual reports cannot be labeled as honest or dishonest. To support the training of a classifier, we used a dice rolling paradigm, but made changes to allow for labeling of discrete events. Additionally, we wished to collect data across a range of taps and swipes, so as to cover a wider spectrum of typical mobile input. The application required participants to *swipe* through lists, *tap* desired selections, and *tap* numbers on a num-pad.

We developed a mobile dice game, similar to the popular game Yatzy. The game consisted of 12 rounds of rolls with five dice. Each round required an initial roll, and two potential re-rolls of selected dice (see Figure 7a). Participants then chose from a list of possible combinations (such as sixes, or three-of-a-kind) and entered the score that a certain combination would yield (see Figure 7b). This was typically the sum of the dice. There was a total of 12 combinations; one for each round. Each combination could only be selected once. If the final dice of a round did not equate to a combination, then any combination could be selected and a score of 0 should be entered. The game recorded both the participants’ actual score and their

reported score. The participants were rewarded based on the sum of their reported scores, providing an incentive to lie:

- \$0.50 : below 150 points
- \$1.00 : between 150 and 200 points
- \$2.00 : more than 200 points

Participants were briefed about the rules and scoring system of the game. Prior to taking part, participants did a qualification test, to ensure that they understood the rules. A help text was available throughout the game for assistance. After completing the experiment, a debriefing screen explained the actual research agenda.

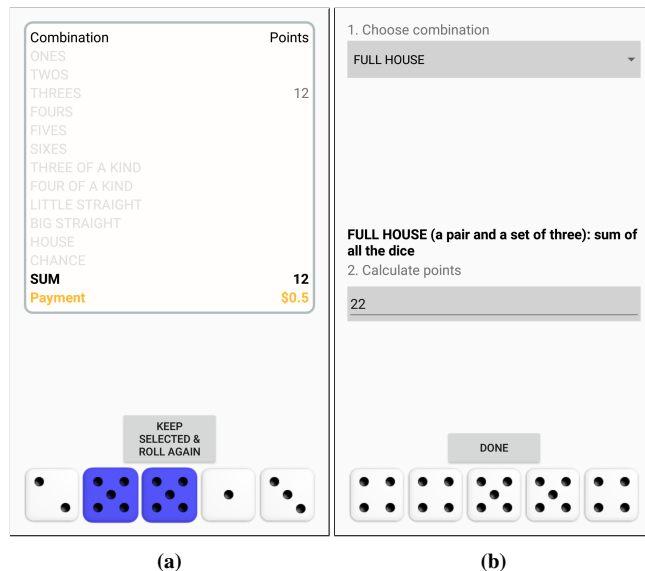


Figure 7: The experimental application used in Study III. Figures show (a) the home screen where participants roll and select dice (selected dice are blue), and see the score board, and (b) the entry screen where combinations and amount of points are entered. The entry screen appears after finishing three rolls and pressing ‘select combination’.

What Constitutes an Entry

In order to train our classifier, we labeled each entry as either a lie or a truth. When beginning a round, the participants were presented with the home screen (see Figure 7a). After rolling the dice the third time, and pressing *Select Combination*, they were presented with the entry screen (see Figure 7b). We define an entry as the time frame from when participants were presented with the entry screen, until and including they hit *Done*. During an entry, the user had to pick a dice combination from a list, and enter the amount of points that the combination and the dice roll amounted to. Swipes were recorded when scrolling the list of combinations; taps were recorded when entering the amount of points on a num-pad. IMU sensors recorded motion data throughout the entry.

We expected three possible outcomes of an entry in the game:

1. The participant reports their score accurately (Truth).
2. The participant purposefully inflates their score (Lie)
3. The participant unintentionally inflates the score (Truth - the participant does not intend to deceive)

In an attempt to differentiate (2) and (3), lies were defined as $score_{enter} - score_{real} > 4$. This was informed by the mean negative deviation from the real score (i.e., when participants under-reported their score, $M = -3.4$).

Participants

We recruited 51 participants from Mechanical Turk, aged 22-57 ($M = 31.5$), 20 females. Participants were told that they were reviewing a mobile game before its launch. Participants were paid according to their score, ranging from \$0.50 US to \$2.00 US, to incentivize lying. We employed the same qualification standards as for Study I.

Apparatus

We developed the application for Android version 6.0 and higher. To obtain comparable data between participants, we excluded tablets and other large-screen devices. A pilot study identified touch pressure level as a good predictor of truthful input, so for the final study we invited only participants who had phones with pressure sensors. This limited the phones to specific models from Google, LG, Motorola, HTC, and OnePlus. We also excluded mobile devices that could not report rotation or acceleration data.

Results

Fifty-one participants took part in the study, completing 561 unique entries, with 44 labeled as lies (8%); 31% of the participants lied at least once. The average lie provided the participant with 15.6 surplus points. Conversely, nine entries reported scores below the actual score, with an average shortfall of 3.4 points.

Our classification results show 98% precision, and 97% recall for truths ($F_1 = .98$), and 65% precision and 59% recall for lies ($F_1 = .57$).

Classification

The classifier was built using the same approach as Study II.

Data Cleaning

We removed participants whose entries indicated that they did not understand the rules, or deliberately rushed the game to optimize payment (amounting to four participants). No participant lied on every single entry.

We removed entries with entered points lower than the actual score (amounted to nine entries, mean shortfall -3.4 points). While they lack an intention to deceive, they could represent either miscalculations or lack of attention with the task. We remove them because correct classification is impossible. We also removed all first entries to account for participants learning to use the interface.

The final data set comprised 51 participants, and 561 entries. The lies covered 44 entries, amounting to 8%. We normalized features per participant (using L_2), and standardized the data set along all axes.

Feature Generation

We generated the same features as in Study II (see Table 2). Additionally, we computed features originating from interactions with the list of dice combinations as well as pressure data (see Table 3).

Feature Group	Features	Description
List	clicks on list	n clicks
Swipe	distance	$d(p_0, p_n)$
	duration	$t_n - t_0$
	length	$\sum d(p_i, p_{i+1})$
	linearity	r^2 , linear regression
	slope	linear regression
	speed	length / duration
	number	n swipes
Pressure	swipe pressure	screen pressure
	button pressure	screen pressure

Table 3: Additional features used for the classifier in Study III.

Feature Selection

Again we handpicked feature groups; *acceleration*, *pressure*, and *num-pad* presented the most viable feature groups for the classification task.

We used recursive feature reduction to eliminate specific bland features within each feature group. From the initial set of features, 11 remained:

- *Num-pad*: button precision (mean, min)
- *Pressure*: button pressure (mean, max, SD, pressure)
- *Pressure*: swipe pressure (mean)
- *Acceleration*: x -acceleration (mean, max, SD)
- *Acceleration*: z -acceleration (SD)

Both Study I and previous work explicitly consider timing as a key predictor of lying [21]. While promising in an administered setting, timing is not robust to the practicalities of day-to-day mobile device usage, where distractions can easily occur mid-input. For this reason, we did not use timing, or response length, as features in our classifier. Additionally, our focus for this study is on lie classification through physiological factors present in sensor data.

There is, however, a temporal dimension within the accelerometer data. Lies took, on average, longer to enter than truths, resulting in more accumulated acceleration data for dishonest input. The acceleration statistics that we computed go some way towards normalizing the effect of this increase in data. To reduce the effect further, we checked for entries longer than three standard deviations of the mean (there were none).

Performance

As Table 4 shows, we achieve high performance in classifying truths ($F_1 = .98$) and above-chance accuracy for lies ($F_1 = .57$). To clarify our results, we provide classification metrics for two other “classifiers”. *Coin-toss* demonstrates classification at random (i.e., tossing a coin), and *Naïve* reports truth for every input (i.e., the most common observation; ZeroR). We observe that our classifier performs well above the random and the naïve approach.

Classifier		Precision	Recall	F_1 -score
Veritaps	Truth	.98	.97	.98
	Lie	.65	.59	.57
	Avg	.96	.95	.95
Coin-toss	Truth	.92	.50	.65
	Lie	.08	.50	.14
	Avg	.85	.50	.61
Naïve	Truth	1.0	.50	.67
	Lie	0.0	0.0	0.0
	Avg	.92	.46	.62

Table 4: Performance metrics of classification results. To compare, we report the theoretical scores from a randomized/Coin-toss and a naïve/ZeroR classifier. The scores show the mean score from a 5-fold cross validation. Average is computed with respect to the skewed distribution of truths and lies.

How Lies and Truths Differ

A truth took on average 13.0s ($SD = 12.0$) to complete. A lie took on average 20.8s ($SD = 30.6$) to complete. Lies were most prominent in the beginning of the experiment; two thirds of all lies were made in the first half of the experiment.

To understand the fundamental differences between an average lie and an average truth, we pick a representative entry from both groups. The entries chosen are the two observations closest to the centroids of two k -means clusters. Here, we explain how the most influential features varied.

Num-pad

We analyzed a range of num-pad entry features, including key dynamics, precision, and hold-down time. Only precision proved to be an effective predictor - the truthful entry records taps with closer proximity to the button’s center.

Pressure

Most entries comprise two num-pad taps, excluding an additional tap on the done button. The truthful entry showed a higher average pressure, and also an increase in pressure between the taps. The lie showed a lower average pressure, and a decrease in pressure between taps.

Acceleration

Acceleration varies between lies and truthful entries, mainly on the x -axis. Specifically, the mean, max, and SD of x -axis acceleration contribute effective indication of truth in input. For these examples of entries, the mean x -acceleration is higher for the lie, which hints that the honest entry enforced a more steady hand during interaction, as in Study II.

Summary

Our results demonstrate an F_1 -score of .98 in classifying truths. We also achieve an F_1 -score of .57 in identifying lies. While promising, the recall rate of lies (59%) renders the technique impractical for binary lie-detection. This is in-line with other so-called lie detectors, such as the polygraph, that report indicators associated with lying for interpretation by a practitioner, rather than a binary classification [20].

Research has shown that we are only slightly better than chance at identifying lies when judging statements with an evenly distributed truth-value [3, 25]. Instead of acting as a standalone lie detector then, our results can provide cues to assist us in improving our lie detection accuracy (as in polygraph tests). Whereas, typically, we would need to assess the truth of any input information, the sensor data associated with mobile input allows us to identify only the subset of information that needs further consideration. We can pre-identify a large part of input as true. This can reduce the space of statements that require approximately chance-level lie analysis and make us significantly more accurate at lie-detection overall.

VERITAPS FOR MOBILE INPUT

We playfully propose to use the results as an additional layer of communication, *Veritaps*, helping a recipient determine the veracity of information from a sender. Veritaps marks both truthful input ✓, and questionable input ?. If the input is questionable, then the recipient can choose whether to request additional information from the sender. In this way, Veritaps can limit the space of interaction that requires further consideration and reduce veracity uncertainty in communication.

We sketch the use of Veritaps across three different styles of mobile interaction: data entry, inter-personal interaction, and personal reflection. Across these three domains, we provide concept use cases, based on the styles of interaction we explored in our studies, and highlight the potential benefits of increased accuracy in veracity judgment.

Mobile Data Entry

We perform many tasks on our mobile devices for which security is paramount, such as online banking. These tasks involve interactions that do not rely on prose, but instead focus on taps and swipes for navigating menu items and entering codes. This renders natural language processing techniques for lie detection inapplicable. Using Veritaps, however, can provide additional layers of security.

Veritaps could provide services with an additional layer of scrutiny for online forms. For example, by adding a mobile entry step to the submission process insurance companies could use Veritaps to flag submissions that need further attention or further supporting documents (see Figure 8a). Online marketplaces could in a similar way use Veritaps to flag suspicious classified advertisements (see Figure 8b).

Interpersonal Communication

We envision that Veritaps could also afford a layer of interpersonal communication, such as increasing confidence in conversations with strangers, or as a playful dimension between friends. For example, after engaging the seller of a car, Veritaps could assess the veracity of the chat messages, ensuring that information presented privately beyond the initial listing is also verified.

Personal Reflection

Self-deception is common and natural, and believed to relate closely to *ethical fading*; the decisions we make, justified by self-deception, that are ethically questionable [22]. Veritaps

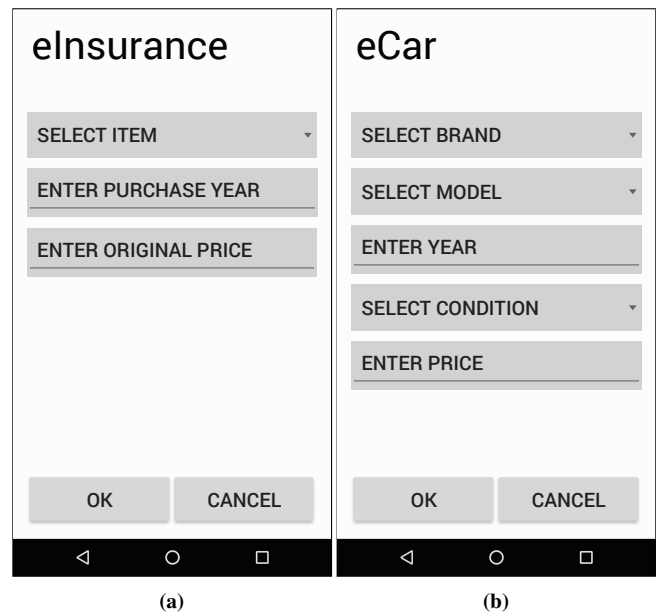


Figure 8: Example Veritaps applications: (a) Veritaps can be used to verify the veracity of an insurance claim, (b) Veritaps may verify the declared condition of a vehicle upon creating an online listing.

can provide prompts against self-deception. For example, you could install a Veritaps browser plugin that prompts you every time you exhibit deceptive behavior. The plugin could help make you aware of how often you are finding excuses for canceling on your trainer, or neglect your diet.

DISCUSSION

The results of the empirical studies, as well as the Veritaps concept, raise several discussion points. They concern the studies, the concept, and the ethical concerns of lie detection.

Our classification accuracies across a broader range of mobile input are relevant only for spontaneous lying. In our directed lying study (Study I), only response time provided a distinguishable feature. For this reason, we cannot speculate about Veritaps' accuracy for habituated lying. Classification accuracy would likely be low here, however, as we believe the spontaneity and guilt of lying creates the physiological features that empower our classification. We also assume that participants made their decision to lie during the entry step of the task and that we, therefore, capture this moment. Currently, we cannot separate the effects of this decision moment from the entry itself, and thus cannot be certain of the efficacy of one without the other. This needs further exploration.

Because pressure data contributed an important feature in our classification, we required participants to have phones with pressure sensors. This limited the applicable participants, as only recent Android phones have pressure sensors. As a result, we found little variation in the phones used in the study. This assisted our classification accuracy, as it reduced requirements for preprocessing of data. As smartphone chipsets are not standardized, a production setup of our proposed technique would optimally require a per-phone-model training process.

The lying we are able to classify has a number of characteristics that limit the generalizability of the findings. Three types of lying may be differentiated [25]: outright lies, exaggeration, and subtle lies. The experimental task in Study I and II examined outright lies, while Study III considered exaggeration. We do not know how our findings generalize to subtle lies. Also, interpersonal lies as present in Study II, might have caused participants to react quite differently compared to interaction without another human being. This could be an explanation for the difference in classification accuracies for Study II-III. Verifying this remains an avenue for future work.

As for the *Veritaps concept*, we have sketched simple example mock-up scenarios. There are practical challenges that arise with implementing Veritaps on smartphones, however.

Ideally, the smartphone would receive a steady stream of labelled ground truth input, to help train the classifier. In practice, however, this would lead to repeated training interruptions on the device and likely prevent adoption. The alternative would be for the phone to come pre-packaged with a trained model, however, without per-user training this could underperform.

Machine learning based estimators require considerable labelled data to perform well. We evaluated to what extent our proposed method would work without per-user calibration. To do this, we ran Leave-One-Subject-Out (LOSO) cross validation [9] using the classifiers from Study II-III. This caused an increase in performance for Study II, but a decrease for Study III. This shows us the more open-ended nature of the task in Study III works poor without per-user training, while other tasks are more suitable to use with pre-trained models.

Within our presented Veritaps scenarios, there remains an opportunity to learn the features that suggest truthful input and, therefore, trick the system. To reduce this risk, we propose that the user should not be shown their own Veritaps assessments, rather they should only be made available to the receiver of the information. In this way, it is important that both parties consent to engaging in a Truth-Verified interaction. Future work should attempt to implement these to study the users' reactions and adaptations based on feedback from our algorithm.

Further, the polygraph has been banned from use in courts in most justice systems because of negligible reliability. In the same way, we do not propose the use of Veritaps as any means of assessment of objective truth. The predictive insights provided by Veritaps should be used with caution, and we cannot recommend critical reliance on Veritaps in any system.

The Veritaps concept raises a number of *ethical questions*. First, lying is an important social lubricant. For instance, small lies play an important role in computer-mediated conversations (e.g., [10]). Also, many lies are simply ignored (the so-called ostrich effect [25]). Therefore, making lies explicit, as in some of the design concepts we discussed, threaten to undermine those functions and introduce mistrust into computer-mediated communication. We call for empirical studies of the Veritaps concept to understand how the availability of truth verification might impact the experience and outcomes of digital communication. Second, our algorithm, and even improved algorithms,

are likely to misclassify. This might challenge the basis of human conversation [8]; that the information we communicate is accurate and truthful. One reaction to this is to have both parties opt-in to having that basis challenged; this would work for several of the concepts we discussed.

Based on the feature analyses, we believe that user interfaces made up of standard UI elements as input fields and buttons are likely to perform best. Across Study I-III we found the details of simple user actions such as taps to carry more reliable information of the veracity of an action, than for instance sliding and scrolling gestures. Additionally, if the methods described in this paper were combined with content-based features, it could likely outperform the performances presented.

Our data do not suggest that smaller lies are harder to detect than bigger lies. For Study II, a binary distinction compared to a scale of deflated scores yielded the clearest division. In other words; an entry with a deflation of one cent, on average held more similar interactional properties with dishonest entries than honest entries.

CONCLUSION

We are frequently subject to lying, and to date lack means of classifying lies on mobile devices beyond written text and speech. This leaves a large space of interaction open to deception. We explore the feasibility of a content-agnostic, sensor-led approach to lie detection on smartphones that considers only taps and swipes. Through three studies we presented empirical evidence for the feasibility of commodity lie detection using mobile interaction.

First, we found significant differences in response times between lies and truths for simple mobile interactions.

Next, we reported on the individual interaction differences observed between lying and truth telling in a mobile version of the Ultimatum game that encouraged lying. The study showed that some features of mobile interaction varies with the honesty of an action. Specifically, properties of number entry were good indicators of deceit.

Last, we reported on a study where participants took part in a mobile dice game that incentivized lying. We trained a classifier on mobile sensor data that ignores the input data itself. We achieved 96% precision and 95% recall in truth detection, and 65% precision and 59% recall for lie detection. While promising, these results do not support reliable binary lie classification. Instead, we suggest their use a means of improving peoples' own near-chance level lie classification.

Based on the findings, we introduced Veritaps: an optional layer in mobile interaction, allowing users to share truth assessments of their input. We presented three potential use cases of Veritaps, across online form-filling, inter-personal communication, and personal reflection.

ACKNOWLEDGEMENTS

This work was supported by the European Research Council, grant no 648785.

REFERENCES

1. Luigi Anolli and Rita Ciceri. 1997. The Voice of Deception: Vocal Strategies of Naive and Able Liars. *Journal of Nonverbal Behavior* 21, 4 (Dec. 1997), 259–284. DOI: <http://dx.doi.org/10.1023/A:1024916214403>
2. Damien Besancenot, Delphine Dubart, and Radu Vranceanu. 2013. The value of lies in an ultimatum game with imperfect information. *Journal of Economic Behavior & Organization* 93 (2013), 239 – 247. DOI: <http://dx.doi.org/10.1016/j.jebo.2013.03.029>
3. C. F. Bond and B. M. DePaulo. 2006. Accuracy of deception judgments. *Pers Soc Psychol Rev* 10, 3 (2006), 214–234.
4. National Research Council. 2003. *The Polygraph and Lie Detection*. The National Academies Press. DOI: <http://dx.doi.org/10.17226/10420>
5. Clayton Epp, Michael Lippold, and Regan L. Mandryk. 2011. Identifying Emotional States Using Keystroke Dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 715–724. DOI: <http://dx.doi.org/10.1145/1978942.1979046>
6. Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. 2012. What Does Touch Tell Us About Emotions in Touchscreen-Based Gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)* 19, 4, Article 31 (Dec. 2012), 30 pages. DOI: <http://dx.doi.org/10.1145/2395131.2395138>
7. Francesca Gino and Scott S. Wiltermuth. 2014. Evil Genius? How Dishonesty Can Lead to Greater Creativity. *Psychological Science* 25, 4 (April 2014), 973–981. DOI: <http://dx.doi.org/10.1177/0956797614520714>
8. H. P. Grice. 1975. Logic and Conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, P. Cole and J. L. Morgan (Eds.). Academic Press, San Diego, CA, 41–58.
9. Nils Y. Hammerla and Thomas Plötz. 2015. Let's (Not) Stick Together: Pairwise Similarity Biases Cross-validation in Activity Recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1041–1051. DOI: <http://dx.doi.org/10.1145/2750858.2807551>
10. Jeff Hancock, Jeremy Birnholtz, Natalya Bazarova, Jamie Guillory, Josh Perlin, and Barrett Amos. 2009. Butler Lies: Awareness, Deception and Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 517–526. DOI: <http://dx.doi.org/10.1145/1518701.1518782>
11. Jeffrey T. Hancock, Lauren E. Curry, Saurabh Goorha, and Michael Woodworth. 2007. On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication. *Discourse Processes* 45, 1 (2007), 1–23. DOI: <http://dx.doi.org/10.1080/01638530701739181>
12. Benjamin E. Hilbig and Corinna M. Hessler. 2013. What lies beneath: How the distance between truth and lie drives dishonesty. *Journal of Experimental Social Psychology* 49, 2 (2013), 263 – 266. DOI: <http://dx.doi.org/10.1016/j.jesp.2012.11.010>
13. Timothy R. Levine, Rachel K. Kim, and Lauren M. Hamel. 2010. People Lie for a Reason: Three Experiments Documenting the Principle of Veracity. *Communication Research Reports* 27, 4 (2010), 271–285. DOI: <http://dx.doi.org/10.1080/08824096.2010.496334>
14. Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T. Chittaranjan, Andrew T. Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. StressSense: Detecting Stress in Unconstrained Acoustic Environments Using Smartphones. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. ACM, New York, NY, USA, 351–360. DOI: <http://dx.doi.org/10.1145/2370216.2370270>
15. Rada Mihalcea and Carlo Strapparava. 2009. The Lie Detector: Explorations in the Automatic Recognition of Deceptive Language. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers (ACLShort '09)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 309–312. <http://dl.acm.org/citation.cfm?id=1667583.1667679>
16. Aske Mottelson and Kasper Hornbæk. 2016. An Affect Detection Technique Using Mobile Commodity Sensors in the Wild. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. ACM, New York, NY, USA, 781–792. DOI: <http://dx.doi.org/10.1145/2971648.2971654>
17. Hessel Oosterbeek, Randolph Sloof, and Gijs van de Kuilen. 2004. Cultural Differences in Ultimatum Game Experiments: Evidence from a Meta-Analysis. *Experimental Economics* 7, 2 (01 Jun 2004), 171–188. DOI: <http://dx.doi.org/10.1023/B:EXEC.0000026978.14316.74>
18. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 309–319. <http://dl.acm.org/citation.cfm?id=2002472.2002512>
19. Martin Pielot, Tilman Dingler, Jose San Pedro, and Nuria Oliver. 2015. When Attention is Not Scarce - Detecting Boredom from Mobile Phone Usage. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 825–836. DOI: <http://dx.doi.org/10.1145/2750858.2804252>

20. Leonard Saxe, Denise Dougherty, and Theodore Cross. 1985. The validity of polygraph testing: Scientific analysis and public controversy. *American Psychologist* 40, 3 (1985), 355–366. DOI: <http://dx.doi.org/10.1037/0003-066X.40.3.355>
21. S. A. Spence, T. F. Farrow, A. E. Herford, I. D. Wilkinson, Y. Zheng, and P. W. Woodruff. 2001. Behavioural and functional anatomical correlates of deception in humans. *Neuroreport* 12, 13 (Sept. 2001), 2849–2853.
22. Ann E. Tenbrunsel and David M. Messick. 2004. Ethical Fading: The Role of Self-Deception in Unethical Behavior. *Social Justice Research* 17, 2 (June 2004), 223–236. DOI: <http://dx.doi.org/10.1023/B:SORE.0000027411.35832.53>
23. Lyn M. Van Swol, Deepak Malhotra, and Michael T. Braun. 2012. Deception and Its Detection: Effects of Monetary Incentives and Personal Relationship History. *Communication Research* 39, 2 (April 2012), 217–238. DOI: <http://dx.doi.org/10.1177/0093650210396868>
24. J. M. Vendemia, R. F. Buzan, and S. L. Simon-Dack. 2005. Reaction time of motor responses in two-stimulus paradigms involving deception and congruity with varying levels of difficulty. *Behav Neurol* 16, 1 (2005), 25–36.
25. Aldert Vrij. 2011. *Detecting Lies and Deceit: Pitfalls and Opportunities*. Wiley.
26. Emma J. Williams, Lewis A. Bott, John Patrick, and Michael B. Lewis. 2013. Telling Lies: The Irrepressible Truth? *PLOS ONE* 8, 4 (04 2013), 1–14. DOI: <http://dx.doi.org/10.1371/journal.pone.0060713>
27. Miron Zuckerman, Bella M. DePaulo, and Robert Rosenthal. 1981. Verbal and Nonverbal Communication of Deception. *Advances in Experimental Social Psychology* 14 (Jan. 1981), 1–59. DOI: [http://dx.doi.org/10.1016/S0065-2601\(08\)60369-X](http://dx.doi.org/10.1016/S0065-2601(08)60369-X)