Acta Psychologica 207 (2020) 103085

Contents lists available at ScienceDirect

Acta Psychologica

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

Investigating the effect of trustworthiness on instruction-based reflexivity

Mathias Van der Biest*,[1], Emiel Cracco*,[1], David Wisniewski, Marcel Brass, Carlos González-García

Ghent University, Department of Experimental Psychology, Faculty of Psychology and Educational Sciences, Henri Dunantlaan 2, B-9000 Gent, Belgium

ARTICLE INFO

Keywords:
Trustworthiness
Instruction-based reflexivity
Social cognition
Instructions
Cognitive control

ABSTRACT

Unlike other species, humans are capable of rapidly learning new behavior from a single instruction. While previous research focused on the cognitive processes underlying the rapid, automatic implementation of instructions, the fundamentally social nature of instruction following has remained largely unexplored. Here, we investigated whether instructor trustworthiness modulates instruction implementation using both explicit and reflexive measures. In a first preregistered study, we validated a new paradigm to manipulate the perceived trustworthiness of two different virtual characters and showed that such a manipulation reliably induced implicit associations between the virtual characters and trustworthiness attributes. Moreover, we show that trustworthy instructors are followed more frequently and faster. In two additional preregistered experiments, we tested if trustworthiness towards the instructor influenced the cognitive processes underlying instruction implementation. While we show that verbally conveyed instructions led to automatic instruction implementation, this effect was not modulated by the trustworthiness of the instructor. Thus, we succeeded to design and validate a novel trustworthiness manipulation (Experiment 1) and to create a social variant of the instruction-based reflexivity paradigm (Experiments 2 and 3). However, this instruction-based reflexivity effect was not modulated by the instructors' trustworthiness.

1. Introduction

The human capacity to learn new behaviors, rules, and actions based on instructions is a unique skill that separates human cognition from that of other species (Cole, Laurent, & Stocco, 2013). This rapid transformation of the declarative content of the instruction into a meaningful action-oriented format, capable of driving the instructed behavior, occurs already after a single presentation, even before its first execution. In contrast, it takes non-human primates months to learn simple, repetitive match-to-sample tasks or set-shifting tasks (Nakahara, 2002; Verrico et al., 2011). Instruction following is thus a key human skill we use daily, such as when constructing furniture from a manual, using new technology, or following directions. As such, instruction following and implementation is considered a central pillar of human collaborative behavior and crucial for human cultural evolution (Heyes, 2018).

Recently, cognitive neuroscientists have started to characterize an intriguing form of instruction following, namely rapid instructed task learning (Brass, Liefooghe, Braem, & De Houwer, 2017; Cole et al., 2013; Cole, Bagic, Kass, & Schneider, 2010; Meiran, Pereg, Kessler, Cole, & Braver, 2015). This type of instruction following reflects the fast (i.e., single trial) encoding and flexible implementation of novel instructions that leads to a phenomenon called "instruction-based reflexivity" (IBR) (e.g. Liefooghe, De Houwer, & Wenke, 2013; Liefooghe, Wenke, & De Houwer, 2012; Meiran et al., 2015). IBR refers to the reflexive activation of instructed actions, irrespective of task relevance and action familiarity (Liefooghe et al., 2013). For example, behavioral studies demonstrated with different experimental paradigms that instructions can interfere with irrelevant ongoing task behavior (Liefooghe, Demanet, & Vandierendonck, 2010; Meiran, Liefooghe, & De Houwer, 2017). Crucially, however, a necessary condition to observe IBR is that the instructions are transformed from a declarative (i.e., semantic) to a procedural (i.e., action-oriented) format, by forming the intention to implement the instructions (Liefooghe et al., 2013; Wenke, Gaschler, Nattkemper, & Frensch, 2009). This dissociation between declarative and procedural representations of instructions is also supported by brain imaging (Formica, González-García, Senoussi, & Brass, 2020; González-García, Arco, Palenciano, Ramírez, & Ruz, 2017; Muhle-Karbe, Duncan, De Baene, Mitchell, & Brass, 2016).

Even though the IBR is a reflexive effect, it can still be modulated. For

* Corresponding authors.
E-mail addresses: mathias.vanderbiest@ugent.be (M. Van der Biest), emiel.cracco@ugent.be (E. Cracco).
[1] Mathias Van der Biest and Emiel Cracco contributed equally.

https://doi.org/10.1016/j.actpsy.2020.103085
Received 10 January 2020; Received in revised form 27 March 2020; Accepted 30 April 2020

example, increasing workload diminishes the IBR effect (Meiran & Cohen-Kdoshay, 2012), while decreasing the response deadline and hence modulating the difficulty-context results in an increased IBR effect instead (Liefooghe et al., 2013). Likewise, increasing the likelihood of implementing the instruction (i.e., the frequency of prospective use) enhances the IBR effect (Whitehead & Egner, 2018). This demonstrates that automatic instruction implementation is affected by other cognitive processes. However, such an approach is agnostic regarding the fundamentally social nature of instruction-following. That is, we mostly receive instructions either directly from someone, for example, a police officer redirecting traffic, a pilot getting instructions from the traffic tower, a teacher in front of a class, or indirectly, from tutorial videos, an instruction manual, or even helpdesk bots. This social dimension of instruction following has important implications, as it would suggest that IBR may be sensitive not only to cognitive but also to social variables. In line with this view, it has been demonstrated before that the social traits of our interaction partners can modulate our behavior. For example, trustworthiness has been shown to influence helping behavior (Wang, Wang, Han, Liu, & Zhang, 2018) and credibility (McGinnies & Ward, 1980). Similar social variables have further been shown to affect high-level cognitive (Baarendse, Counotte, O'Donnell, & Vanderschuren, 2013) as well as motor functions (Cracco, Genschow, Radkova, & Brass, 2018). Concerning instruction following, Hale, Payne, Taylor, Paoletti, and De C Hamilton (2018) recently demonstrated the influence of instructor trustworthiness on decision-making processes and *explicit instruction* following in a virtual reality maze study. They manipulated the trustworthiness of two virtual characters during an interview. Afterwards, participants escaped from a virtual maze and could ask direction advice from one of the two virtual character. Participants approached the trustworthy virtual character a significantly more often and followed their advice more often compared to the advice of the untrustworthy virtual character. This evidence suggests that social variables affect the explicit implementation of instructions. Crucially, however, whether social variables such as instructor's trustworthiness also modulate the reflexive effect of instruction implementation remains unknown.

Given the extensive evidence for IBR and the fact that this effect can be modulated in principle, here we investigate whether social variables affect the automatic effects of instruction implementation, treating instructions as inherently social. More specifically, we investigated how the automatic IBR is modulated by the trustworthiness of the instructor. In a first experiment, we developed a novel paradigm to experimentally manipulate perceived trustworthiness (The Door Game) and validated the procedure using both implicit (i.e. Implicit Association Test) and explicit measurements (i.e. the percentage of advice following). In two additional experiments, we tested the influence of trustworthiness, as manipulated by The Door Game, on instruction following, as measured using the IBR paradigm (Liefooghe et al., 2012). In order to do so, we developed a 'social' version of the IBR paradigm where a virtual character instructs participants to carry out specific S-R mappings. We hypothesized a reduced IBR for untrustworthy compared to trustworthy instructors.

2. Experiment 1

2.1. Method

2.1.1. Participants

Forty-nine first-year psychology students (36 female, M_{age} = 19.65 years, SD_{age} = 2.26, all naïve to the purpose of the experiment) at Ghent University participated in the experiment in return for one course credit and monetary reimbursement up to 2.30 euro. The experiment was conducted in accordance with the local institutional ethics committee, and all participants gave written informed consent. This study was preregistered (https://aspredicted.org/blind.php?x=f8wr7a).

2.1.2. Apparatus

The experiment was programmed in Psychopy (Peirce et al., 2019), and all the instructions were presented on a black background with a white font. Participants were tested on a 15-inch. Dell computer monitor with corresponding Sennheiser 215 headphones.

2.1.3. Trust manipulation

To manipulate the trustworthiness of the instructor, a new social manipulation was designed based on a simple game. Participants received advice from one of two different human-like digital virtual characters to choose one out of three doors (e.g., "I would pick the red door"). The verbal instructions were recorded by two native Dutch speakers and were synchronized with the lip movements of two different virtual characters, created with the software CrazyTalk. As a result, we obtained four instructors, based on two voices and two virtual characters, named Lulu, Soni, Paola, and Cati. For each participant, two different virtual characters with different voices were randomly selected. After hearing the advice, participants had to select one out of three doors (i.e., Red, Blue, Green) by pressing the corresponding numeric keyboard button (i.e., "1", "2", "3"). Participants were instructed that they could freely choose to follow the advice or not and had up to 5 s to decide. One door led to winning 0.10 Eurocent, another to losing 0.10 Eurocent, and a third door did not lead to any reward or punishment. Participants received feedback by means of a green circle (winning 0.10), red circle (losing 0.10), or white circle (no reward nor punishment). Moreover, the amount of money earned was updated and displayed on a corner of the screen during the entire game.

One of the virtual characters consistently advised picking the door leading to monetary reward (i.e., trustworthy) while the other virtual character (i.e., untrustworthy) gave good, bad, or neutral advice in 33% of the trials each (see Fig. 1). Crucially, the participant was not informed about this difference and learned this trustworthiness distinction within the 36 trials of The Door Game.

2.1.4. Implicit association test

To measure trustworthiness, we tested if the trustworthy and untrustworthy virtual characters were implicitly associated with trustworthy and untrustworthy attributes after The Door Game, using an Implicit Association Test (Greenwald, Nosek, & Banaji, 2003). This paradigm consisted of five blocks, in which the participant had to respond with left-right keyboard responses (.i.e., "I or E" keys) to categorize different stimuli (e.g. either virtual character, trustworthiness attributes, or both). In the first block, the virtual characters were associated with a specific response (e.g. "*if you see a picture of Paola press I*", "*if you see a picture of Cati press E*"). In the second block, trustworthy and untrustworthy attributes were associated with a specific response (e.g. "*if you see a word associated with trustworthiness, press I*", "*if you see a word associated with untrustworthiness, press E*"). These associations were practiced in blocks one and two, where only one stimulus type was presented (i.e., virtual character in block 1, and (un)trustworthy attributes in block 2). In the third block, both virtual characters and attributes were presented intermixed, and categorized according to the stimulus-response mappings learned in the previous blocks.

Following, was a fourth block in which the response mapping for the trustworthiness attributes was reversed and like block two only one stimulus type (i.e., trustworthiness attributes) was presented (e.g. "Trustworthy press E", "Untrustworthy press I"). The last block was identical to the third block, but with the response, mapping practiced in the fourth block.

2.1.5. Procedure

The overall structure of the experiment was presented to participants prior to participating. First, participants played The Door Game for 36 trials. On 50% of the trials, the trustworthy virtual character gave her advice for a door, whereas the untrustworthy virtual character gave advice in the remaining trials. The trustworthiness of the virtual

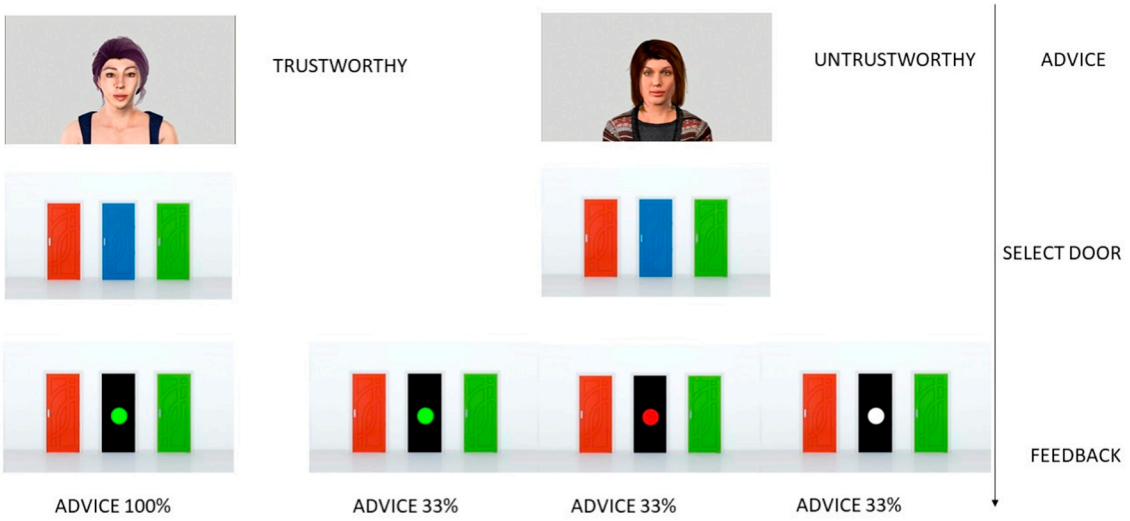

Fig. 1. The design of The Door Game. The arrow depicts the timeline. During the advice phase, either a trustworthy or an untrustworthy virtual character gives their preference for a door, in this example "I would pick the blue door". Next, the participant had to select a door by pressing a corresponding keyboard key (e.g. "1" = red, "2" = blue, "3" = green). When the virtual character is trustworthy, unbeknownst to the participant, following the advice always leads to an optimal monetary reward (green circle). When the virtual character is untrustworthy, following the advice leads to a monetary reward, a punishment (red circle), or no reward nor punishment (white circle) in 33% of the trials each. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

character and voices were counterbalanced across participants but remained constant within one participant. Trial order was randomized. After selecting the preferred door, participants received feedback by means of a green circle (+10 cent) if the correct door was selected, or a red (−10 cent) or white (0) circle, if an incorrect door was selected, staying on display for 1 s. The total amount of money that had already been earned was always shown in the upper right corner. All participants started with one Euro.

Following this trustworthiness manipulation, the existence of an association between the virtual character and trustworthiness was tested implicitly with the Implicit Association Test (Greenwald et al., 2003). The first, second, and fourth block consisted of 40 trials, in which the participant had to respond to either virtual characters (i.e., block 1) or (un)trustworthy words (i.e., block 2 and 4), with the corresponding response (e.g. "I" or "E"). The third and the last block consisted of 60 trials, with intermixed trustworthiness attributes and virtual characters. Participants were instructed to respond as fast as possible. If an error was made, a red fixation cross was presented beneath the stimuli, and the next trial would start after 150 ms when pressing the correct response.

2.1.6. Design

The independent variable for this within-subject design was instructor trustworthiness (e.g., trustworthy or untrustworthy). The explicit dependent variables were following rate, reaction times, and money earned during The Door Game. The dependent variable of the Implicit Association Test was "D1", which represents the strength of the association between the avatars and the attributes so that a positive D1 reflects an association between the trustworthy virtual character and trustworthiness attributes and vice versa, while a negative D1 reflects an association between the trustworthy virtual character and untrustworthy attributes (and vice versa) (Greenwald et al., 2003).

2.2. Analyses

All analyses were conducted in Rstudio (R Core Team, 2017), in combination with JASP (JASP Team, 2018).

2.2.1. The Door Game

To examine if the trustworthiness manipulation worked, we investigated if participants followed the advice of the trustworthy virtual character more compared to the untrustworthy virtual character by comparing the proportion of advice following for the trustworthy and untrustworthy virtual character using a paired *t*-test. In addition to the preregistered analysis, a paired *t*-test was conducted to compare the reaction times of door selection following the advice of the trustworthy and untrustworthy instructors.

2.2.2. Implicit association test

Prior to the analysis, trials in which participants were slower than 10,000 ms were excluded. Furthermore, we excluded participants who made over 40% of mistakes. To analyze the implicit association between the trustworthiness of the virtual character and words associated with (un)trustworthiness a D1 score was calculated by taking the difference in reaction times for congruent and incongruent blocks during the Implicit Association Task divided by the standard deviation of the reaction times across conditions. This was calculated according to the standard guidelines of Greenwald et al. (2003). A two-sided one-sample *t*-test to compare D1 scores to zero was conducted. Values significantly above zero refer to faster responding on blocks where response mappings are shared between trustworthy virtual characters and positive trust-related words, and between untrustworthy virtual characters and negative trust-related words. Negative values reflect faster responding when the mappings are shared between trustworthy virtual characters and negative trust-related words, and between untrustworthy virtual characters and positive trust-related words. Furthermore, a Pearson correlation between the amount of earned money and the implicit bias was calculated.

2.3. Results

2.3.1. The Door Game

A two-sided paired *t*-test revealed that participants followed the trustworthy virtual character (M_{follow} = 0.83, SD_{follow} = 0.22, SE_{follow} = 0.03) significantly more often than the untrustworthy virtual character (M_{follow} = 0.33, SD_{follow} = 0.10, SE_{follow} = 0.01), $t(48) = 13.58$, $p < .001$, $d = 1.94$, and responded significantly faster after trustworthy (M_{RT} = 1106 ms, SD_{RT} = 667, SE_{RT} = 95) compared to untrustworthy virtual character advice (M_{RT} = 1444 ms, SD_{RT} = 880, SE_{RT} = 126) trials, $t(48) = -4.39$, $p < .001$, $d = -0.63$.

2.3.2. Implicit association test

A two-sided one-sample *t*-test to compare the D1 score (M_{D1} = 0.31, SD_{D1} = 0.32, SE_{D1} = 0.05) to zero showed that the D1 scores were significantly larger than zero, $t(48) = 6.75$, $p < .001$, $d = 0.96$. Moreover, we observed a positive correlation between the amount of earned money during the game and D1, $r = 0.36$, $p < .01$.

2.4. Discussion Experiment 1

The results of the first study demonstrate the effectiveness of The Door Game to manipulate the trustworthiness of virtual characters acting as instructors, as shown both on explicit (i.e. advice following) and implicit measures (D1 scores of the IAT). Participants were capable to learn the trustworthiness of a virtual character without prior information within 36 trials, as they significantly followed the explicit advice of the trustworthy virtual character more and responded faster to her advice, compared to the untrustworthy virtual character In addition to these measures, the result of the Implicit Association Test suggested that participants implicitly associated trustworthy virtual characters with trust-related attributes and untrustworthy virtual characters with untrust-related ones. Moreover, the amount of earned money was positively correlated with this implicit bias, suggesting that participants with stronger implicit associations between the virtual characters and trust adapted their behavior more strongly to the virtual characters' trustworthiness during The Door Game. Overall, Experiment 1 thus validates a new and easy task to manipulate a complex social factor such as trustworthiness (Ashraf, Bohnet, & Piankov, 2006). In the second experiment, we investigated the influence of trustworthiness on instruction implementation.

3. Experiment 2

3.1. Method

3.1.1. Participants

A group of 119 first-year psychology students (105, female, M_{age} = 19.43 years, SD_{age} = 4.35, all naïve to the purpose of the experiment) at Ghent University participated in the experiment in return for one course credit and a performance-based reward (up to 3.30 euro). An a priori power analysis indicated that in order to detect a small effect (d = 0.3) reliably (α = 0.05, power = 0.90) using within-subject manipulations, we needed a sample size of $N = 119$. With this sample, all effect sizes with d ≥ 0.18 would be significant at α = 0.05 (Lakens, Scheel, & Isager, 2018). Twenty-one participants were excluded from the analyses based on preregistered exclusion criteria (https://aspredicted.org/blind.php?x=ss8fx3), namely excessive errors ($N = 5$), unresponsiveness to the trustworthiness manipulation ($N = 14$), or both ($N = 2$). Responsiveness to the manipulation was measured with two Likert scales asking participants to rate the trustworthiness of both characters 1 (i.e., highly untrustworthy) to 5 (i.e., Highly trustworthy). Based on these responses, a trustworthiness index was calculated. This reflects the difference between the ratings for the trustworthy and untrustworthy virtual character. When this difference was negative or zero the participant was excluded. Note that the trustworthiness index was calculated due to the ambiguity of the response on the preregistered open questions (see Appendix A). The experiment was conducted in accordance with the local institutional ethics committee, and all participants gave written informed consent.

3.1.2. The door game

The materials of The Door Game were identical to the first experiment. However, now The Door Game was played for six blocks. The first block included 36 trials, whereas the remaining blocks had 12 trials each.

3.1.3. IBR

The general design and procedure of the IBR paradigm were adapted from Braem, Liefooghe, De Houwer, Brass, and Abrahamse (2017) and consisted of two tasks, an inducer task, and a diagnostic task. For the former, the participant was presented a rule consisting of two S-R mappings (e.g. "*press left if the word is NEWS, press right if the word is BIKE*"). This rule must be retained until the participant was presented one of the two words (e.g. 'NEWS') in a colored print (e.g. 'green'). Between the inducer screen and target, participants were

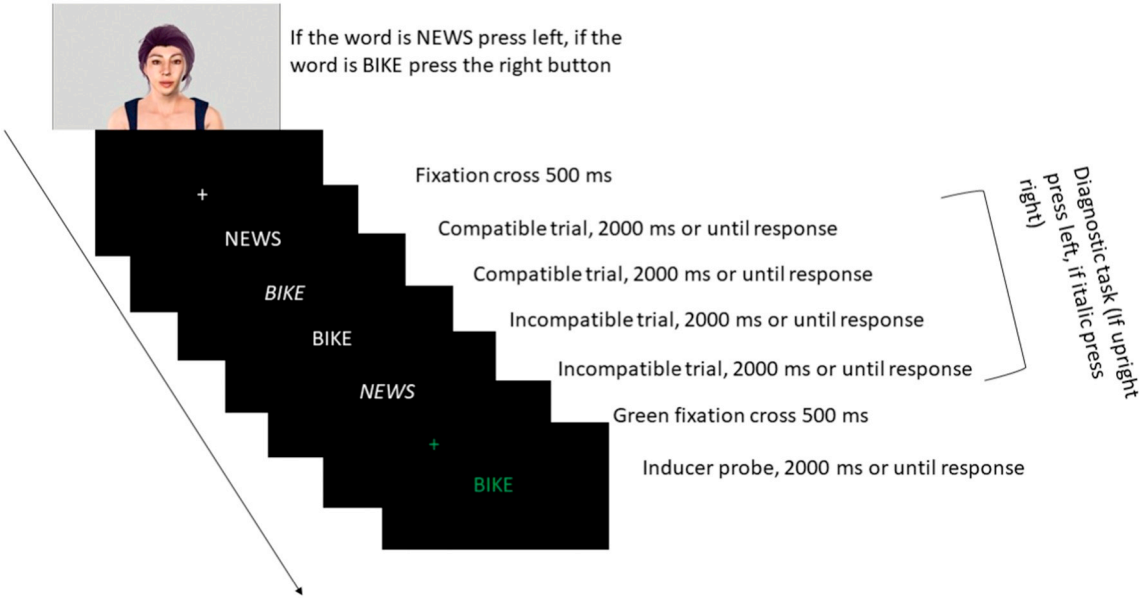

Fig. 2. Example of the Instruction based reflexivity task. The black arrow depicts the timeline. First, the trustworthy or the untrustworthy virtual character gives an S-R instruction (e.g. "*if the word is NEWS press Left If the word is BIKE press the right*"). Following is a white fixation cross for 500 ms, which is the start of the diagnostic runs. For this example, the diagnostic run consists of four trials, but this could also be 8, 12, or 16 trials. On each trial, one of the two S-R mappings were presented in italics (i.e., press right) or upright (i.e., press left). These trials can either compatible (e.g. first two trials) or incompatible (e.g. last two trials) with the instructed S-R mapping. The green fixation cross indicates the end of the diagnostic runs and prepares the participant for the inducer probe. Which is one of the two S-R mappings printed in green. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

shown a diagnostic task. During diagnostic trials, participants had to respond to a different dimension of the instructed stimuli, namely font (e.g. '*press left when the word is in italics, press right when the word is printed upright*'), independent of the previously instructed inducer task. The instructions for the diagnostic task remained identical over the experiment. Such task configuration lead to situations in which the required response by the diagnostic task overlapped with the instructed inducer response, and in which the correct response for the diagnostic runs were either compatible (e.g." Italics, NEWS"), or incompatible (e.g. "Upright, NEWS") (see Fig. 2).

A list of 43 rules consisting of two Dutch words with a similar word frequency and a length of four letters was constructed (e.g. "*NEWS-BIKE*"), and participants were divided according to their participant number in two groups with opposite S-R mappings (e.g. "*If the word is NEWS press Left, if the word is BIKE press right*", or "*If the word is BIKE press left, if the word is NEWS press right*").The left-right keyboard response configuration responses ("D" and "K" on a QWERTY keyboard) were identical during diagnostic and inducer runs. Furthermore, the same pair of stimuli were presented for both tasks (see Supplementary Table 1 Virtual Avatar for all stimuli for experiments 1, 2, 3).

The instructions of the IBR were presented by means of the same virtual characters as in The Door Game. Note that for each participant the virtual character and voice configuration were identical in The Door Game and the IBR task. These spoken S-R mappings were randomly assigned to five blocks in groups of eight S-R mappings and one practice block which always consisted of the same two rules (i.e., "*if VOICE press left, if END press right*", "*if ADVICE press left, if LADY press right*", or vice versa). Within each block, the S-R mappings were randomly paired with a specific number of diagnostic tasks runs (two runs with four diagnostic trials, two runs with eight diagnostic trials, two runs with 12 diagnostic trials, two runs with 16 diagnostic trials). These randomized lengths of runs made the appearance of the inducer task less predictable (Meiran et al., 2015).

3.1.4. Design

A 2 × 2 within-subject design with trustworthiness (e.g. trustworthy or untrustworthy) and compatibility (e.g. compatible or incompatible) as independent variables was used. The crucial dependent variables were reaction times and error rates on the diagnostic trials during the IBR task. Additionally, we decided to compute the inverse efficiency scores (i.e., RT/1-Error Rates) (IES), a compound variable that corrects for speed-accuracy trade-offs, as sometimes these are a more powerful measurement in reaction times experiments as the IBR (Vandierendonck, 2017).

3.1.5. Materials

The experiment was programmed in Psychopy (Peirce et al., 2019), and all stimuli of the diagnostic task were presented in white font on a black background. The inducers probes were printed in green. Further, participants were tested on a 15-inch dell computer monitor with corresponding Sennheiser 215 headphones.

3.1.6. Procedure

Participants were tested in groups of a maximum of five. Prior to their participation, participants were informed about the overall structure of the experiment (i.e., The Door Game and IBR paradigm). Overall, the experiment consisted of 12 alternating blocks of The Door Game and the IBR (i.e. one practice block and five experimental blocks). The experiment started with a block of The Door Game, consisting of 36 trials (50% trustworthy, 50% untrustworthy), followed by two practice trials of the IBR task, a new block of The Door Game, consisting of 12 trials (50% trustworthy, 50% untrustworthy), was presented. This was then followed by an alternating sequence of IBR and The Door Game blocks. Between each block, the instructions of either the IBR or The Door Game were repeated, and participants had the chance to take a break for as long as needed. The instructions for the diagnostic trials and the response mappings of the inducer task were explained on the screen before the practice IBR block and repeated before each IBR block. Each run in the IBR blocks started with a video of a virtual character (50% trustworthy, 50% untrustworthy), verbally instructing new S-R mappings for the inducer task. Note that these S-R mappings were only instructed by means of videos and were not visually presented.

After the video, a fixation cross was presented for 500 milliseconds, and the diagnostic runs started. Each diagnostic trial was presented until a response was made or after 2000 ms had passed. When an incorrect response was given, participants received a red square as feedback. Following a full diagnostic run, a green fixation cross was presented for 500 ms, and the inducer probe was presented. Identical to the diagnostic trials, the probe inducer was presented for 2000 ms or until a response was made. Note, that no feedback was presented after the inducer task. After 500 ms a new video and corresponding diagnostic and inducer trials started.

When the experimental phase was completed, participants filled in two short questionnaires on the computer, and one written questionnaire. The first questionnaire was the short Right-Wing Authoritarianism scale, which consisted of 14 Dutch items and was based on Altemeyer (1998). This questionnaire was included only for exploratory purposes and results are not reported. Next, participants were shown the two virtual characters and asked to rate their trustworthiness on a scale from 1 (e.g. "*very untrustworthy*") to 5 (e.g. "*very trustworthy*"). Finally, participants answered two questions about The Door Game (see Appendix A).

3.2. Analyses

All analyses were conducted in Rstudio (R Core Team, 2017), in combination with JASP (JASP Team, 2018).

3.2.1. Preprocessing

Prior to analysis, a trustworthiness index was calculated (i.e., trustworthiness rating for the trustworthy virtual character minus trustworthiness rating for the untrustworthy virtual character). If this score was zero or negative, the participant was excluded from further analyses. Additionally, the preregistered exclusion criteria were applied to the dataset. Note that the preregistration was somewhat ambiguous with respect to the exclusion criteria. To clarify, we planned on using the same exclusion criteria as those described in (Braem et al., 2017). All participants that responded incorrectly on ≥40% of the diagnostic or inducer trials were excluded. At the trial level, all diagnostic trials following an error (9%) or with a response time faster than 200 ms (< 0.001%), and all diagnostic trials with an inaccurate probe inducer response (19%), were excluded. For the inducer task, all trials with a response time faster than 200 ms (< 0.001%) were excluded. In addition, all trials in the practice block were excluded from the analyses.

3.2.2. The Door Game

The analyses for The Door Game were identical to the first experiment.

3.2.3. IBR

To investigate the instruction-based reflexivity, three within-subject repeated measures ANOVA models (type III) were constructed for reaction times of trials with a correct response, error rates, and IES in diagnostic trials with compatibility (compatible or incompatible) and type of instructor (trustworthy or untrustworthy) as a within-subject factor. To examine the influence of trustworthiness on the inducer task, a paired sample test, comparing the reaction times of the trials with a correct response, error rates and IES following instructions from the trustworthy or untrustworthy virtual character was conducted. Additionally, a Pearson correlation was calculated to investigate the association between the amount of money earned during The Door

Game and the interaction effect of the reaction times on the diagnostic runs. This to explore if and in which direction the trustworthiness inductions of The Door Game, of which the amount of money is a proxy, is associated with the diagnostic interactions scores.

3.3. Results

3.3.1. The Door Game

A paired sample *t*-test showed that participants significantly followed the trustworthy virtual character (M_{follow} = 0.93, SD_{follow} = 0.11, SE_{follow} = 0.01) more than the untrustworthy virtual character (M_{follow} = 0.33, SD_{follow} = 0.07, SE_{follow} < 0.001), $t(97) = 48.30$, $p < .001$, $d = 4.88$. Additionally, participants were significantly faster to select a door when the virtual character was trustworthy (M_{RT} = 730 ms, SD_{RT} = 541, SE_{RT} = 55), compared with untrustworthy (M_{RT} = 908 ms, SD_{RT} = 561, SE_{RT} = 57), as analyzed with a paired-sample *t*-test, $t(97) = -7.80$, $p < .001$, $d = -0.79$.

3.3.2. IBR - diagnostic trials

A repeated measure ANOVA on reaction times showed a significant main effect of compatibility, $F(1,97) = 55.56$, $p < .001$, $MSE = 0.001$, $\eta_p^2 = 0.36$, but no main effect of trustworthiness, $F(1,97) = 1.87$, $p = .17$, $MSE = 0.001$, $\eta_p^2 = 0.02$, nor an interaction effect, $F(1,97) = 1.28$, $p = .26$, $MSE < 0.001$, $\eta_p^2 = 0.01$. The same significant main effect of compatibility was found for the error rates, $F(1,97) = 53.71$, $p < .001$, $MSE = 0.004$, $\eta_p^2 = 0.36$, as well as the same non-significant main effect of trustworthiness, $F(1,97) = 0.05$, $p = .83$, $MSE = 0.002$, $\eta_p^2 = 0.00$, and a non-significant interaction effect $F(1,97) = 1.42$, $p = .24$, $MSE = 0.002$, $\eta_p^2 = 0.01$. Likewise, analyses of the IES demonstrated a significant main effect of compatibility $F(1,97) = 72.43$, $p < .001$, $MSE = 0.006$, $\eta_p^2 = 0.43$, a non-significant main effect of trustworthiness $F(1,97) = 1.51$, $p = .22$, $MSE = 0.003$, $\eta_p^2 = 0.02$, and non-significant interaction effect, $F(1,97) = 2.56$, $p = .11$, $MSE = 0.003$, $\eta_p^2 = 0.03$ (see Fig. 3).

To investigate the influence of earned money on the reaction times, a Pearson correlation was calculated between the amount of money earned and the reaction times interaction effect. Prior, the difference scores (i.e., $(RT_{trustworthy\_compatible} - RT_{trustworthy\_incompatible}) - (RT_{untrustworthy\_compatible} - RT_{untrustworthy\_incompatible})$), which reflect this interaction effect were calculated. This resulted in a non-significant correlation, $r = -0.06$, $p = .57$.

3.3.3. IBR - inducer task

A paired sample *t*-test revealed no significant difference between the reaction times following instructions from trustworthy (M_{RT} = 820, SD_{RT} = 199, SE_{RT} = 20) and untrustworthy (M_{RT} = 830, SD_{RT} = 206, SE_{RT} = 21) instructors, $t(97) = -0.91$, $p = .36$, $d = -0.09$. Similar, no significant difference was found between error rates for trustworthy (M_{ER} = 0.19, SD_{ER} = 0.14, SE_{ER} = 0.01) and untrustworthy (M_{ER} = 0.18, SD_{ER} = 0.13, SE_{ER} = 0.01) instructors, $t(97) = 0.18$, $p = .86$, $d = 0.02$, and for the IES for trustworthy (M_{IES} = 1.06, SD_{IES} = 0.42, SE_{IES} = 0.04) and untrustworthy (M_{IES} = 1.06, SD_{IES} = 0.41, SE_{IES} = 0.04) instructors, $t(97) = -0.08$, $p = .93$, $d = 0.00$.

3.4. Discussion Experiment 2

The results of the second experiment replicate the effectiveness of The Door Game in manipulating the trustworthiness of the virtual characters That is, participants followed the untrustworthy virtual character significantly less and more slowly than the trustworthy virtual character, and 79% of the participants indicated that the trustworthy virtual character was highly trustworthy (i.e., *"5"* on a Likert scale) and the untrustworthy virtual character highly untrustworthy (i.e., *"1"*).

The results of the diagnostic task of the IBR paradigm revealed significant main effects of compatibility for reaction times, error rates, and IES, replicating previous studies on the reflexive effects of instructions (e.g. Liefooghe et al., 2012; Meiran et al., 2015) and extending these by showing that an IBR of similar size can be obtained via verbally instructing participants. However, in contrast with our predictions, there was no interaction between this compatibility effect and trustworthiness, nor a main effect of trustworthiness. Similarly, there was no main effect of trustworthiness on the inducer task.

The goal of the third experiment was to replicate the results of the second study, provide clear evidence that IBR can be induced with verbally conveyed instructions, and that The Door Game is a valid manipulation of trustworthiness. Most importantly, we wanted to

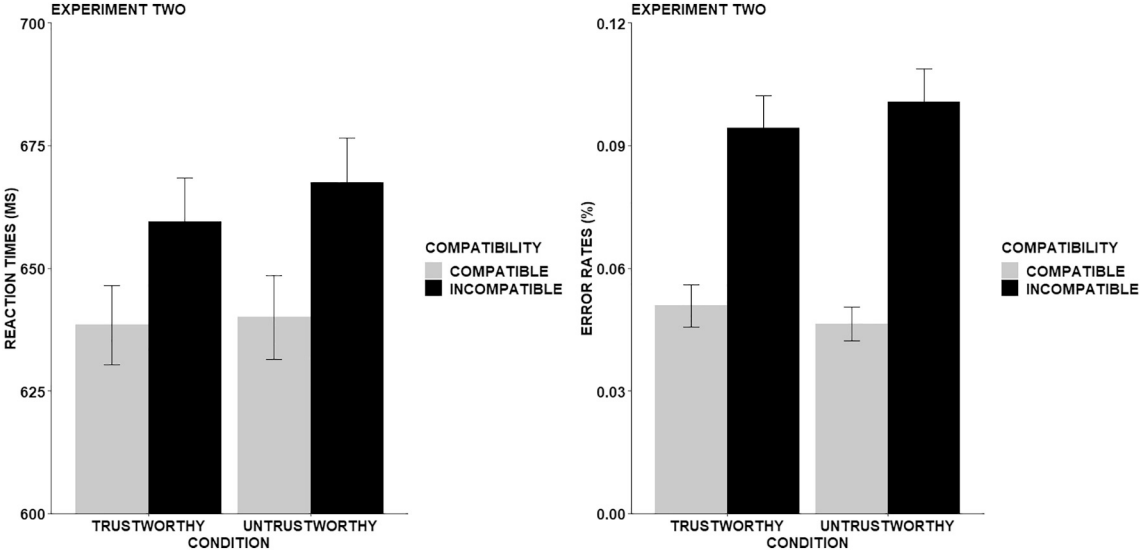

Fig. 3. Mean error rates and reaction times for the second experiment. All error bars are ± standard error of the mean. Both reaction times and error rates are depicted separately for instructor trustworthiness (i.e., trustworthy and untrustworthy) and compatibility (i.e., compatible and incompatible). We found a significant main effect of compatibility, as participants were faster and made fewer errors during the compatible diagnostic trials, irrespectively of the trustworthiness of the instructor.

provide further evidence for the absence of an interaction. Finally, a memory task was added to the general design of the second experiment to explore whether participants memorized S-R mappings instructed by the untrustworthy virtual character more compared to the trustworthy virtual character.

4. Experiment 3

4.1. Method

4.1.1. Participants

All one hundred and fifty-five participants (136 female and 1 non-binary, M_{age} = 18.86 years, SD_{age} = 2.92, all naïve to the purpose of the experiment) were first-year psychology students at Ghent University that participated in return for a monetary performance reward and a course credit. This large sample size, as well as the exclusion criteria and analysis plan, were preregistered (https://aspredicted.org/yw6w4.pdf). For the analyses, 40 participants were excluded, due to too many mistakes on the diagnostic or the inducer task (N = 14), a failure to manipulate the trustworthiness as measured with the Likert scale questionnaire (N = 21), or both (N = 5). Note, that the same exclusion criteria as in Experiment two are applied.

4.1.2. Task and procedure

The apparatus, materials, and design were identical to the second experiment, but an additional memory task was given at the end of the experiment. Participants were presented the 42 S-R rules they executed during the IBR inducer task and 42 newly constructed S-R rules, which consisted of the same words as the executed inducer task rules but in a different order and with a possible switched left or right position. Participants had to indicate if they had executed this S-R mapping (e.g., “press C”) during the IBR task or not (e.g., “press M”). There was no time limit during this task, and prior to each rule, a fixation cross was presented for 200 ms. The order of the new and old S-R mappings was randomized.

4.2. Analyses

The analyses and preprocessing were identical to the second experiment and (Braem et al., 2017). All diagnostic runs followed by an inaccurate probe inducer response (15%) and all diagnostic trials following an error (8%) or with a response time faster than 200 ms (< 0.001%) were excluded from the analyses of the diagnostic trials. For the inducer task, all trials with a response time faster than 200 ms (< 0.001%) were excluded. For exploratory purposes, the analyses of the diagnostic trials were repeated with block as an additional factor and a Pearson correlation between the interaction effect and the amount of money earned during The Door Game calculated for the dependent variable reaction times and IES. To investigate the performance on the memory task, a two-tailed paired sample t-test was conducted. Moreover, to increase statistical power, the data of the second and third experiments were also combined in pooled analyses. Finally, Bayesian analyses were conducted on the diagnostic runs in the pooled analyses, for which the default settings of JASP (JASP Team, 2018) were used.

4.3. Results

4.3.1. The Door Game

Participants significantly followed the advice of the untrustworthy virtual character (M_{follow} = 0.32, SD_{follow} = 0.06, SE_{follow} < 0.00) less often than that of the trustworthy virtual character (M_{follow} = 0.94, SD_{follow} = 0.12, SE_{follow} = 0.01) as shown by a paired samples t-test, t(114) = 50.82, p < .001, d = 4.74. In the same vein, a paired samples t-test showed that participants were significantly faster to select a door following the advice of the trustworthy virtual character (M_{RT} = 688 ms, SD_{RT} = 309, SE_{RT} = 29) compared to the untrustworthy virtual character (M_{RT} = 849 ms, SD_{RT} = 435, SE_{RT} = 41), t(114) = −5.93, p < .001, d = −0.55.

4.3.2. IBR - diagnostic trials

To examine the main effects of trustworthiness, compatibility and the interaction effect, for the reaction times of the trials with a correct response, error rates, and IES, three within-subject repeated measures ANOVA models were constructed. For the reaction times, only a significant main effect of compatibility was found, F(1,114) = 53.62, p < .001, MSE < 0.001, η_p^2 = 0.32, while both the main effect of trustworthiness, F(1,114) = 1.11, p = .29, MSE = 0.001, η_p^2 = 0.00, and the interaction effect F(1,114) = 0.08, p = .78, MSE < 0.001, η_p^2 = 0.00, remained non-significant. Likewise, the analysis of the error rates, demonstrated the absence of a main effect of trustworthiness, F(1,114) = 1.17, p = .28, MSE = 0.001, η_p^2 = 0.01, and interaction effect F(1,114) = 0.51, p = .48, MSE = 0.002, η_p^2 = 0.00, but a significant main effect of compatibility F(1,114) = 20.98, p < .001, MSE = 0.005, η_p^2 = 0.16, was found. For the IES, the analysis showed a significant main effect of compatibility, F(1,114) = 25.30, p < .001, MSE = 0.01, η_p^2 = 0.18, and trustworthiness, F(1,114) = 4.57, p = .03, MSE = 0.003, η_p^2 = 0.04, but a non-significant interaction effect, F(1,114) = 0.19, p = .66, MSE = 0.003, η_p^2 = 0.00 (see Fig. 3). The main effect of trustworthiness indicates that the IES was larger for the trustworthy than for the untrustworthy virtual character (see Supplementary Table 1).

In order to investigate the three-way interaction between trustworthiness, compatibility and block, three additional within-subject repeated measures ANOVAs were constructed. Mauchly's test indicated a violation of the sphericity assumption for reaction times *Mauchly's* W = 0.743, p < .001, *Greenhouse-Geisser* ε = 0.865, error rates *Mauchly's* W = 0.747, p < .001, *Greenhouse-Geisser* ε = 0.872, and IES *Mauchly's* W = 0.156, p < .001, *Greenhouse Geisser* ε = 0.483. Therefore, the p-values were corrected with the Greenhouse-Geisser correction. Analy and exploratory reasons, the same correlation between the amount of earned money and the interaction effect for reaction times was calculated for the IES. This yielded a non-significant correlation between the amount of money earned and the interaction effect of the IES, r = −0.03, p = .75.

4.3.3. IBR - inducer task

A paired sample t-test showed no significant difference between the reaction times of the trials with a correct response on trustworthy (M_{RT} = 860 ms, SD_{RT} = 187, SE_{RT} = 17) compared to untrustworthy (M_{RT} = 861, SD_{RT} = 188, SE_{RT} = 18) instructors, t(114) = −0.08, p = .94, d < − 0.01. In similar vein, no significant difference between the error rates for trustworthy (M_{ER} = 0.14, SD_{ER} = 0.12, SE_{ER} = 0.01) and untrustworthy instructors (M_{ER} = 0.16, SD_{ER} = 0.13, SE_{ER} = 0.01), t(114) = −1.05, p = .29, d = −0.10, and for IES scores for trustworthy (M_{IES} = 1.03, SD_{IES} = 0.30, SE_{IES} = 0.03) and untrustworthy (M_{IES} = 1.08, SD_{IES} = 0.46, SE_{IES} = 0.04) instructors was found, t(114) = −1.24, p = .22, d = −0.12.

4.3.4. Memory

A paired sample t-test demonstrated that the accuracies were significantly larger for the instructions that were given by the untrustworthy virtual character in the IBR experiment (M = 0.48, SD = 0.16, SE = 0.02) compared to instructions that were given by the trustworthy virtual character (M = 0.45, SD = 0.17, SE = 0.01), t(114) = 1.63, p = .05, d = 0.15. However, it is important to note, that both accuracy rates were beneath chance level, although, this was only significant, as tested with a one sample t-test for the trustworthy t(114) = −3.23, p < .001, d = −0.30, and only marginally significant for the untrustworthy virtual character, t(114) = −1.52, p = .07, d = −0.14.

4.3.5. Pooled results

For exploratory purposes, and to increase statistical power, we also combined the data of the second and the third experiment ($N = 213$, $M_{age} = 19.15$, $SD_{age} = 3.64$) and repeated the aforementioned analyses.

4.3.6. IBR - diagnostic trials

Analyses with three repeated measure ANOVAs showed that there was a main effect of compatibility for reaction times $F(1,212) = 108.68$, $p < .001$, $MSE < 0.001$, $\eta_p^2 = 0.34$, error rates, $F(1,212) = 66.75$, $p < .001$, $MSE = 0.004$, $\eta_p^2 = 0.24$, and IES, $F(1,212) = 78.98$, $p < .001$, $MSE = 0.009$, $\eta_p^2 = 0.27$. In contrast, there was no evidence for a main effect of trustworthiness for reaction times $F(1,212) = 0.02$, $p = .89$, MSE = 0.001, $\eta_p^2 = 0.00$, error rates $F(1,212) = 0.37$, $p = .54$, $MSE = 0.001$, $\eta_p^2 = 0.00$, or IES $F(1,212) = 0.71$, $p = .40$, $MSE = 0.003$, $\eta_p^2 = 0.00$, nor for an interaction effect, $F(1,212) = 0.96$, $p < .33$, $MSE < 0.001$, $\eta_p^2 = 0.00$; $F(1,212) = 0.13$, $p = .72$, $MSE = 0.001$, $\eta_p^2 = 0.00$;$F(1,212) = 0.64$, $p = .42$, $MSE = 0.003$, $\eta_p^2 = 0.00$ (see Fig. 4).

Equivalent Bayesian analyses of the reaction times revealed strong evidence in favor of the null effect of trustworthiness $BF_{01} = 12.5$, and an interaction, $BF_{01} = 6.67$, while we found strong evidence for the main effect of compatibility, $BF_{10} > 150$. In a similar vein, Bayesian analysis for the error rates, demonstrated evidence for a non-significant main effect of trustworthiness, $BF_{01} = 12.5$ or interaction effect, $BF_{01} = 9.09$, however, clear evidence for a main effect of compatibility was found, $BF_{10} > 150$. Furthermore, there was clear evidence for the main effect of compatibility for IES, $BF_{10} > 150$, while there was no evidence for the main effect of trustworthiness, $BF_{01} = 11.11$, nor interaction effect, $BF_{10} = 7.69$.

4.3.7. IBR - inducer task

A paired sample *t*-test comparing the reaction times on trustworthy ($M_{RT} = 841$ ms, $SD_{RT} = 193$, $SE_{RT} = 13$) compared to untrustworthy ($M_{RT} = 847$ ms, $SD_{RT} = 197$, $SE_{RT} = 13$) showed a non-significant difference, $t(212) = -0.69$, $p = .49$, $d = -0.05$, similar results were found for the trustworthy error rates ($M_{ER} = 0.16$, $SD_{ER} = 0.13$, $SE_{ER} < 0.01$) and untrustworthy error rates ($M_{ER} = 0.17$, $SD_{ER} = 0.13$, $SE_{ER} < 0.01$), $t(212) = -0.70$, $p = .48$, $d = -0.05$, and for trustworthy IES ($M_{IES} = 1.05$, $SD_{IES} = 0.36$, $SE_{IES} = 0.02$) and untrustworthy IES scores ($M_{IES} = 1.07$, $SD_{IES} = 0.43$, $SE_{IES} = 0.03$), $t(212) = -1.07$, $p = .29$, $d = -0.07$.

4.4. Discussion Experiment 3

The third experiment successfully replicated the results of the second experiment. Participants responded slower, made more mistakes, and showed larger IES on incompatible diagnostic trials compared to compatible ones. However, there was no main effect of trustworthiness nor a significant interaction between trustworthiness and compatibility for the reaction times and error rates on the diagnostic runs nor on the inducer trials. Moreover, the analysis showed a significant main effect of trustworthiness on IES in the diagnostic runs, suggesting that when correcting for errors, participants responded slower on the diagnostic runs when instructed by the trustworthy compared to the untrustworthy virtual character. However, this effect was not confirmed by the pooled and Bayesian analyses. Instead, the analyses showed a significant main effect of compatibility and no significant main effect of trustworthiness nor an interaction effect for any of the dependent variables.

5. General discussion

Here, we investigated the effect of trustworthiness on automatic instruction implementation. In a first study, we developed and validated a new trustworthiness manipulation, using both explicit (i.e. the percentage of direct instructor advice following) and implicit measures (i.e. D1 Implicit Association Test). Results revealed that participants followed the advice of the trustworthy virtual character more and implicitly associated this virtual character more with trustworthiness than the untrustworthy virtual character. The results of the second and third experiments replicated the classic reflexive IBR effect (Liefooghe et al., 2012; Meiran et al., 2015) using verbal instead of written instructions. However, this effect was not modulated by the trustworthiness of the instructor, and these results were replicated in a third study. Additionally, the results of The Door Game of the second and the third experiment confirmed the validity of our trustworthiness manipulation.

By using verbal instructions presented by virtual characters, we created a novel 'social' version of the IBR paradigm that is ecologically more valid, allowing for contextual (e.g. social) manipulations. Experiment two and three provide more evidence for the automatic power of new instructions on ongoing behavior (Liefooghe et al., 2012; Meiran et al., 2015), even without printed S-R mappings but with instructions given by virtual characters with differing social

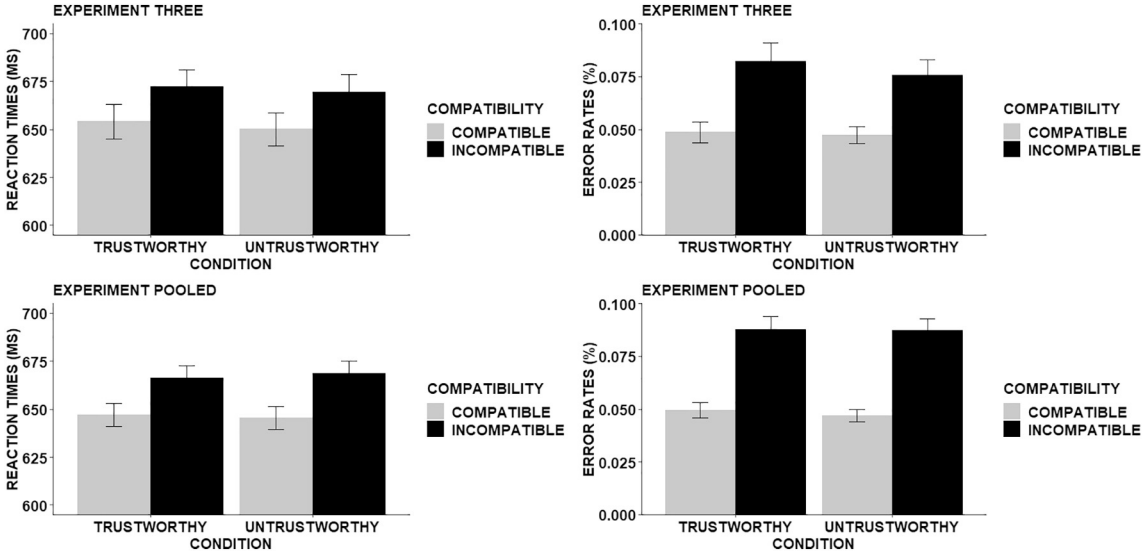

Fig. 4. Mean error rates and reaction times for the second experiment and pooled analysis. Reaction times and error rates are depicted according to the compatibility and instructor trustworthiness. A significant main effect of compatibility was found irrespectively of the interaction with the trustworthiness of the instructor. All error bars are ± standard error of the mean.

characteristics (i.e., trustworthiness). However, in contrast to our predictions, with the current experimental paradigm, the trustworthiness of the instructor did not modulate this automatic effect. While there was a clear effect of advisor trustworthiness on explicit responding behavior in The Door Game, hence demonstrating the influence of trust on ongoing behavior (Hale et al., 2018; van't Wout & Sanfey, 2008), this effect did not transfer to reflexive behavior, as measured with the diagnostic trials.

There are at least three possible explanations for the absence of an effect of trustworthiness on reflexive measures of instruction following. First, it could be that the trust associations established during The Door Game did not transfer to the IBR. Indeed, in the IBR task, participants must always follow the instruction, irrespective of the instructor. As a result, it could be argued that all instructors were trustworthy during these blocks. However, previous research did find such transfer effects of social variables to unrelated behavior as animacy on motor priming (Liepelt & Brass, 2010), the influence of trustworthiness on affective evaluation (Aguado, Román, Fernández-Cahill, Diéguez-Risco, & Romero-Ferreiro, 2011), and the influence of pro- and antisocial primes on automatic imitation of socially (in)appropriate gestures (e.g. Cracco et al., 2018). Furthermore, traditional trustworthiness manipulations such as the investment game or the prison dilemma show transfer and learning effects over paradigms and tasks (e.g. Collins, Juvina, & Gluck, 2016; Hale et al., 2018; Juvina, Saleem, Martin, Gonzalez, & Lebiere, 2013). Therefore, future research will be needed to fully rule out a possible lack of transfer due to the design of the IBR task itself.

Second, it is possible that the instructor trustworthiness established in The Door Game does transfer to the IBR task but does not modulate the IBR effect. While the IBR has repeatedly been shown to be modulated by cognitive variables effects, such as working memory load or the intention to implement (Meiran & Cohen-Kdoshay, 2012; Muhle-Karbe et al., 2016), and has found to be correlated with intelligence (Meiran, Pereg, Givon, Danieli, & Shahar, 2016), this reflexivity effect might be insensitive to contextual manipulations, such as the aforementioned social context. It is important to note, however, that the IBR effect reflects only one aspect of instruction following, namely instruction implementation, or the formation of an action-oriented representation. It is possible that social variables modulate instruction following only in an earlier stage, prior to the formation of the action-oriented format (i.e., *when the instruction is still in its declarative format*) or when transforming the declarative format into an action-oriented format (Brass et al., 2017), and that this effect is filtered out when instructions are represented in action-oriented representation.

Finally, it is possible that instruction implementation is not modulated by social variables. However, this would contrast with recent prominent proposals. For example, Heyes (2018), argued that human adaptive behavior evolved not only through genetics but also through cultural evolution and that the latter is built on social metacognitive capacities such as instruction following. Similarly, in their theoretical framework of instruction following, De Houwer, Hughes, and Brass (2017) emphasized the crucial role of instructions in society, as without instructions an essential line of information communication would be lost. It would also contrast with empirical evidence showing that the credibility of the instructor modulates the strength and hence influence of the message (e.g., Vogel & Wänke, 2016), and with evidence showing that a broad variety of social variables can modulate decision-making processes (e.g. for a review see van den Bos, Jolles, & Homberg, 2013). In the same vein, it has repeatedly been shown that social context has a profound influence on explicit instruction following, as for example, when monitored by an observer, the probability of instruction following significantly increases (Donadeli & Strapasson, 2015), or increased direction following when instructed by a trustworthy compared to an untrustworthy virtual character (Hale et al., 2018). In line with these findings, are the results of the conducted memory task. Participants had a significantly better memory of the S-R rules when instructed by the untrustworthy compared to the trustworthy virtual avatar. This intriguing result is congruent to studies showing that untrustworthy informational sources require larger attentional resources and increased attention (see Farmer, Apps, & Tsakiris, 2016; Rule, Slepian, & Ambady, 2012; Vanneste, Verplaetse, Vanhiel, & Braeckman, 2007). However, it is important to be cautious when interpreting these results. Memory recall for both instructors was below chance. Future research is needed to further elaborate on this finding.

In summary, the current study succeeded in designing a novel paradigm to manipulate an abstract social concept, namely trustworthiness. Furthermore, we were able to design a 'social' version of the IBR paradigm in which a virtual character gives verbal instructions. This paradigm showed similar IBR effects compared with previous paradigms using written instructions. However, in contrast to our expectations, we were unable to demonstrate an effect of social variables on instruction implementation. There are different potential reasons why this might be the case and future research will have to explore different approaches to manipulate trust and target different phases of instruction processing, implementation, and following to provide a definitive answer on whether and at which level trustworthiness modulates instructions following.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.actpsy.2020.103085.

Open practices statement

Raw and preprocessed data, analyses scripts, JASP files, and virtual character stimuli for all the experiments can be found on https://osf.io/x5gyh/, as well as all preregistrations.

Funding

M.V.D.B. was supported by Special Research Fund of Ghent University BOF.GOA.2017.0002.03. C.G.G. was supported by the Special Research Fund of Ghent University BOF.GOA.2017.0002.03 and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 835767. DW was supported by the Research Foundation Flanders (FWO), the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 665501, and FWO grant FWO.KAN.2019.0023.01. E.C. is funded by the Research Foundations Flanders (FWO18/PDO/049).

Credit authorship contribution statement

Mathias Van der Biest: Software, Formal analysis, Investigation, Writing - original draft, Data curation, Visualization. **Emiel Cracco**: Conceptualization, Methodology, Resources, Project administration, Formal analysis, Visualization, Writing - review & editing. **David Wisniewski**: Writing - review & editing, Conceptualization, Methodology. **Marcel Brass**: Supervision, Writing - review & editing. **Carlos González-García**: Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing.

Declaration of competing interest

None.

Appendix A. Subject number

1) "Did you perceive a difference in the behavior of the two avatars during the game?"

2) "If so, at what point during the experiment did you realize this difference?"

References

van't Wout, M., & Sanfey, A. G. (2008). Friend or foe: The effect of implicit trustworthiness judgments in social decision-making. *Cognition, 108*(3), 796–803. https://doi.org/10.1016/j.cognition.2008.07.002.

Aguado, L., Román, F. J., Fernández-Cahill, M., Diéguez-Risco, T., & Romero-Ferreiro, V. (2011). Learning about faces: Effects of trustworthiness on affective evaluation. *The Spanish Journal of Psychology, 14*(2), 523–534. https://doi.org/10.5209/rev_SJOP.2011.v14.n2.1.

Altemeyer, B. (1998). The other "authoritarian personality". *Advances in experimental social psychology. Vol. 30. Advances in experimental social psychology* (pp. 47–92). . https://doi.org/10.1016/S0065-2601(08)60382-2.

Ashraf, N., Bohnet, I., & Piankov, N. (2006). Decomposing trust and trustworthiness. *Experimental Economics, 9*(3), 193–208. https://doi.org/10.1007/s10683-006-9122-4.

Baarendse, P. J. J., Counotte, D. S., O'Donnell, P., & Vanderschuren, L. J. M. J. (2013). Early social experience is critical for the development of cognitive control and dopamine modulation of prefrontal cortex function. *Neuropsychopharmacology, 38*(8), 1485–1494. https://doi.org/10.1038/npp.2013.47.

van den Bos, R., Jolles, J. W., & Homberg, J. R. (2013). Social modulation of decision-making: A cross-species review. *Frontiers in Human Neuroscience, 7*. https://doi.org/10.3389/fnhum.2013.00301.

Braem, S., Liefooghe, B., De Houwer, J., Brass, M., & Abrahamse, E. L. (2017). There are limits to the effects of task instructions: Making the automatic effects of task instructions context-specific takes practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(3), 394–403. https://doi.org/10.1037/xlm0000310.

Brass, M., Liefooghe, B., Braem, S., & De Houwer, J. (2017). Following new task instructions: Evidence for a dissociation between knowing and doing. *Neuroscience & Biobehavioral Reviews, 81*, 16–28. https://doi.org/10.1016/j.neubiorev.2017.02.012.

Cole, M. W., Bagic, A., Kass, R., & Schneider, W. (2010). Prefrontal dynamics underlying rapid instructed task learning reverse with practice. *Journal of Neuroscience, 30*(42), 14245–14254. https://doi.org/10.1523/JNEUROSCI.1662-10.2010.

Cole, M. W., Laurent, P., & Stocco, A. (2013). Rapid instructed task learning: A new window into the human brain's unique capacity for flexible cognitive control. *Cognitive, Affective, & Behavioral Neuroscience, 13*(1), 1–22. https://doi.org/10.3758/s13415-012-0125-7.

Collins, M. G., Juvina, I., & Gluck, K. A. (2016). Cognitive model of trust dynamics predicts human behavior within and between two games of strategic interaction with computerized confederate agents. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.000 49.

Cracco, E., Genschow, O., Radkova, I., & Brass, M. (2018). Automatic imitation of pro- and antisocial gestures: Is implicit social behavior censored? *Cognition, 170*, 179–189. https://doi.org/10.1016/j.cognition.2017.09.019.

De Houwer, J., Hughes, S., & Brass, M. (2017). Toward a unified framework for research on instructions and other messages: An introduction to the special issue on the power of instructions. *Neuroscience & Biobehavioral Reviews, 81*, 1–3. https://doi.org/10.1016/j.neubiorev.2017.04.020.

Donadeli, J. M., & Strapasson, B. A. (2015). Effects of monitoring and social reprimands on instruction-following in undergraduate students. *The Psychological Record, 65*(1), 177–188. https://doi.org/10.1007/s40732-014-0099-7.

Farmer, H., Apps, M., & Tsakiris, M. (2016). Reputation in an economic game modulates premotor cortex activity during action observation. *European Journal of Neuroscience, 44*(5), 2191–2201. https://doi.org/10.1111/ejn.13327.

Formica, S., González-García, C., Senoussi, M., & Brass, M. (2020). *NEURAL OSCILLATIONS DISSOCIATE BETWEEN MAINTENANCE AND PROCEDURALIZATION OF NOVEL INSTRUCTIONS* [preprint]. *Neuroscience*. https://doi.org/10.1101/2020.01.20.912162.

González-García, C., Arco, J. E., Palenciano, A. F., Ramírez, J., & Ruz, M. (2017). Encoding, preparation and implementation of novel complex verbal instructions. *NeuroImage, 148*, 264–273. https://doi.org/10.1016/j.neuroimage.2017.01.037.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of Personality and Social Psychology, 85*(2), 197–216. https://doi.org/10.1037/0022-3514.85.2.197.

Hale, J., Payne, M. E., Taylor, K. M., Paoletti, D., & De C Hamilton, A. F. (2018). The virtual maze: A behavioural tool for measuring trust. *Quarterly Journal of Experimental Psychology, 71*(4), 989–1008. https://doi.org/10.1080/17470218.2017.1307865.

Heyes, C. (2018). Enquire within: Cultural evolution and cognitive science. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1743), 20170051. https://doi.org/10.1098/rstb.2017.0051.

JASP Team (2018). *JASP (version 0.9)[computer software]*.

Juvina, I., Saleem, M., Martin, J. M., Gonzalez, C., & Lebiere, C. (2013). Reciprocal trust mediates deep transfer of learning between games of strategic interaction. *Organizational Behavior and Human Decision Processes, 120*(2), 206–215. https://doi.org/10.1016/j.obhdp.2012. 09.004.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963.

Liefooghe, B., De Houwer, J., & Wenke, D. (2013). Instruction-based response activation depends on task preparation. *Psychonomic Bulletin & Review, 20*(3), 481–487. https://doi.org/10.3758/s13423-013-0374-7.

Liefooghe, B., Demanet, J., & Vandierendonck, A. (2010). Persisting activation in voluntary task switching: It all depends on the instructions. *Psychonomic Bulletin & Review, 17*(3), 381–386. https://doi.org/10.3758/PBR.17.3.381.

Liefooghe, B., Wenke, D., & De Houwer, J. (2012). Instruction-based task-rule congruency effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 38*(5), 1325–1335. https://doi.org/10.1037/a0028148.

Liepelt, R., & Brass, M. (2010). Top-down modulation of motor priming by belief about Animacy. *Experimental Psychology, 57*(3), 221–227. https://doi.org/10.1027/1618-3169/a000028.

McGinnies, E., & Ward, C. D. (1980). Better liked than right: Trustworthiness and expertise as factors in credibility. *Personality and Social Psychology Bulletin, 6*(3), 467–472. https://doi.org/10.1177/014616728063023.

Meiran, N., & Cohen-Kdoshay, O. (2012). Working memory load but not multitasking eliminates the prepared reflex: Further evidence from the adapted flanker paradigm. *Acta Psychologica, 139*(2), 309–313. https://doi.org/10.1016/j.actpsy.2011.12.008.

Meiran, N., Liefooghe, B., & De Houwer, J. (2017). Powerful instructions: Automaticity without practice. *Current Directions in Psychological Science, 26*(6), 509–514. https://doi.org/10.1177/0963721417711638.

Meiran, N., Pereg, M., Givon, E., Danieli, G., & Shahar, N. (2016). The role of working memory in rapid instructed task learning and intention-based reflexivity: An individual differences examination. *Neuropsychologia, 90*, 180–189. https://doi.org/10.1016/j.neuropsychologia.2016.06.037.

Meiran, N., Pereg, M., Kessler, Y., Cole, M. W., & Braver, T. S. (2015). The power of instructions: Proactive configuration of stimulus–response translation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 768–786. https://doi.org/10.1037/xlm0000063.

Muhle-Karbe, P. S., Duncan, J., De Baene, W., Mitchell, D. J., & Brass, M. (2016). Neural coding for instruction-based task sets in human Frontoparietal and visual cortex. *Cerebral Cortex,* bhw032. https://doi.org/10.1093/cercor/bhw032.

Nakahara, K. (2002). Functional MRI of macaque monkeys performing a cognitive set-shifting task. *Science, 295*(5559), 1532–1536. https://doi.org/10.1126/science.1067653.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., ... Lindeløv, J. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*. https://doi.org/10.3758/s13428-018-01193-y.

R Core Team (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2016.

Rule, N. O., Slepian, M. L., & Ambady, N. (2012). A memory advantage for untrustworthy faces. *Cognition, 125*(2), 207–218. https://doi.org/10.1016/j.cognition.2012.06.01.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods, 49*(2), 653–673. https://doi.org/10.3758/s13428-016-0721-5.

Vanneste, S., Verplaetse, J., Vanhiel, A., & Braeckman, J. (2007). Attention bias toward noncooperative people. A dot probe classification study in cheating detection. *Evolution and Human Behavior, 28*(4), 272–276. https://doi.org/10.1016/j.evolhumbehav.2007.02.005.

Verrico, C. D., Liu, S., Asafu-Adjei, J. K., Sampson, A. R., Bradberry, C. W., & Lewis, D. A. (2011). Acquisition and baseline performance of working memory tasks by adolescent rhesus monkeys. *Brain Research, 1378*, 91–104. https://doi.org/10.1016/j.brainres.2010.12.081.

Vogel, T., & Wänke, M. (2016). *Attitudes and attitude change* (2nd ed.). NY: Routledge/Taylor & Francis Group.

Wang, X., Wang, N., Han, S., Liu, S., & Zhang, L. (2018). The influence of facial trustworthiness on helping behavior: The role of attachment type. *Acta Psychologica Sinica, 50*(11), 1292. https://doi.org/10.3724/SP.J.1041.2018.01292.

Wenke, D., Gaschler, R., Nattkemper, D., & Frensch, P. A. (2009). Strategic influences on implementing instructions for future actions. *Psychological Research Psychologische Forschung, 73*(4), 587–601. https://doi.org/10.1007/s00426-009-0239-x.

Whitehead, P. S., & Egner, T. (2018). Frequency of prospective use modulates instructed task-set interference. *Journal of Experimental Psychology: Human Perception and Performance, 44*(12), 1970–1980. https://doi.org/10.1037/xhp0000586.