this document has not been peer-reviewed

Do I still like myself?
Human-robot collaboration entails emotional consequences

Patrick P. Weis & Cornelia Herbert

Ulm University, Germany

Abstract

Investigating emotional processes has been vital for understanding human-human interaction. Specifically, emotional concepts of oneself and the interaction partner shape the interaction style and are associated with mental health and cognitive performance. Here, it is investigated whether these concepts are equally relevant in human-*robot* interaction (HRI). We measured emotional concepts before and after online collaboration with a robot described as (a) able, (b) unable to experience emotions, or (c) autonomous without reference to emotions, compared to a (d) control condition without collaboration. Concepts were measured using the affective His-Mine-Paradigm (aHMP) in which participants were asked to affectively evaluate pronoun-noun-pairs that were related to themselves (e.g., "my victory") or the robot (e.g., "his victory"). Results indicated that (1) the aHMP can be validly used in HRI contexts, (2) emotional self-concept got less positive after interacting with an "emotionless" robot, and emotional robot-concept got more positive after interacting with an (3) "autonomous" or (4) "emotional" robot. We conclude that beliefs about and interactions with robots can change emotional concepts of both oneself and the robot. We argue and show that such changes are likely linked to well-being, performance, and interaction style. Thus, emotional consequences for the human should be considered when designing HRIs.

Keywords: Human-robot interaction; cognitive offloading; extended cognition; emotions; artificial agents

1 Introduction

1.1 Overview

In modern society, having emotional contact with robots or virtual agents is common. The first children who grew up with computers—in the 80s—referred to humans as "emotional machines" (Turkle, 2012, p. 30). Labelling people as emotional thus contrasted them with the remaining machines which were unable to experience emotions. These boundaries have been blurred in the 90s when machines like Furby (Hasbro Inc., Rhode Island, USA) started expressing emotions (e.g., "I am scared") and were object of a child's empathy (e.g., "when the batteries are removed [...] the Furby forgets its life"; Turkle, 2012, p. 41). Since then, numerous case reports indicated that interactions with robots designed for emotional contact were indeed described as emotional, though not necessarily as emotional as interactions with humans (Turkle et al., 2006). Experimental investigations bolster these reports. For example, watching a robot getting tortured was shown to increase negative affect and physiological arousal when compared to watching a non-torture control video (Rosenthal-von der Pütten et al., 2013). Similarly, observing non-verbal robot behavior was found to boost positive affect as compared to observing a robot lacking non-verbal behavior (Rosenthal-von der Pütten et al., 2018). Interacting with robots thus seems to have emotional consequences for the human interaction partner. This complements the extensive emotion-related research on the robot side of the interaction (i.e., designing emotional features like emotion recognition or expression; e.g., Bartneck, 2003; Beck et al., 2010; Breazeal, 2003; Hegel et al., 2010).

Here, we build on these studies to investigate a specific subcomponent of emotional consequences for the human interaction partner: consequences for the emotional self- and other-concepts. Consequences for emotional self- and other-concepts are particularly relevant as they are linked to well-being and mental health (e.g.; Mezulis et al., 2004; Winter et al., 2015)—a link possibly mediated by changes in structural plasticity in the prefrontal cortex (Lumma et al., 2018). In particular, we focus on two preregistered research goals. First, we strive to confirm that the presently used paradigm (the *affective His-Mine-Paradigm* or aHMP; Herbert, Herbert, et al., 2011; Herbert, Pauli, et al., 2011, 2011; see section 2.3.1) can

be validly used to measure the emotional self- and other-concepts in a novel context: interactions with non-human robot rather than human agents (*cf*. **H1-1** and **H1-2**). Second, we strive to build on that validation and use the aHMP to investigate whether beliefs about and interactions with robots can change the emotional self- and robot-concepts (*cf*. **H2-1** and **H2-2**). Gaining a better understanding of such emotional consequences of human-robot interaction seems imperative in the highly technologized worlds of today and tomorrow.

1.2 Emotional Self- and Other-Concept

How people see themselves is not merely influencing their thoughts but also impacts behavior and well-being (Diener & Diener, 1996). This is reflected by the *self-positivity bias*: people see themselves positively—more positively than reality warrants. This exaggerated association of positive information with oneself is pervasive and substantial (Mezulis et al., 2004) and likely promoting mental health (Taylor & Brown, 1988, 1994). Consequentially, a self-positivity bias is rather the norm than the exception in healthy subjects (Diener & Diener, 1996; Mezulis et al., 2004) and crucially, positive affective states such as being in an romantic relationship (e.g., Meixner & Herbert, 2018) can even extend this self-positivity bias to a virtual unknown other. In contrast, a decreased self-positivity bias is associated with depression, anxiety, and attention-deficit/hyperactivity disorder (Mezulis et al., 2004), and also with personality and its disorders (e.g., Winter et al., 2015). How people see themselves can also impact cognitive performance. For example—according to stereotypes—women perform worse in math than men (Spencer et al., 1999). This stereotype threat leads to decreased performance but, crucially, performance differences vanished in an experimental setting once participants were informed that no gender differences would exist for the task at hand (Spencer et al., 1999). Thus, although the self-concept as a whole is supposed to be relatively stable (Epstein, 1973), specific subcomponents like a negative task-relevant self-concept can be altered relatively quickly. The dynamic nature of parts of the self-concept is also illustrated by the dynamic nature of state self-esteem (Heatherton & Polivy, 1991). In sum, a positive self-concept is desirable because it is linked to well-being and increased performance but can be subject to short-term change elicited by self-compromising situations.

How people see themselves is however also intricately linked to how people see others. In fact, establishing close relationships might encompass including the other in the self (*self-expansion*; Aron & Aron, 1996). Accordingly, participants needed longer to decide whether a trait described them if their spouse did not share versus shared that trait (Aron et al., 1991) and in an experimental study investigating participants in a romantic relationship the self-positivity was shown to expand to a *self-and-other*-positivity-bias specifically for participants in a romantic relationship as compared to singles (Meixner & Herbert, 2018). Unsurprisingly, how we see another agent also impacts our interactions with that agent. This impact—but also the close links between other- and self-concept—is illustrated by the seminal four-category model of adult attachment (Bartholomew & Horowitz, 1991; Bowlby, 1979): For example, positive self- and other-concepts are associated with a secure relationship style that allows intimacy while negative self- and other-concepts are associated with a fearful relationship style leading to avoidance of intimacy. Underlining the relevance of the other concept for human-robot interaction, attachment theory has already been used to derive guidelines for social robot design (Dziergwa et al., 2018). Taken together, we argue that changes in emotional self- and other-concepts are likely during interactions with robots and are relevant because they are linked to mental well-being, performance, and interaction style.

1.3 Cognitive Offloading and its Consequences

Solving problems in concert with fellow humans, robots, non-embodied computers like smartphones, or analog aids like pen and paper to outsource cognitive processing is an abundant activity in modern society (*cognitive offloading* or CO, Risko & Gilbert, 2016; for additional reviews , also see Clark, 1999; Hutchins, 1995; Ifrah, 2001; Kirsh, 2013). In contrast to the less focused interactions with robot companions like Furby, CO is centered on task-relevant outcomes like providing the solution to an arithmetic or navigation problem. It is known that people can adaptively adjust how frequently they engage in CO depending on the situation (e.g., depending on the time costs of engaging in CO; W. D. Gray et al., 2006; Storm et al., 2017; or on whether speed or accuracy is prioritized; Weis & Wiese, 2019) but sometimes fail to do so because of inaccurate metacognitive judgements like wrongly estimating own

cognitive ability such as underestimating own memory (Gilbert, 2015; Touron, 2015). Current research thus provides a successively clearer picture about what makes people *engage* in CO.

But apart from the underpinnings of engaging, it is also vital to the hybrid societies of today and tomorrow to understand the *consequences* of CO to eventually learn how to promote or prevent them (i.e., help people make better decisions whether or not to engage in CO), an understanding that is currently mostly lacking (compare Risko & Gilbert, 2016, p. 685). That consequences do exist is highly likely as illustrated by the following examples. Firstly, it was shown that CO can alter the way stored information is represented (Fu, 2011): Biological memory consists of a multitude of active processes and so, for example, similar items can be grouped together to reduce representational complexity (Nosofsky, 1992) which then, however, leads to decision biases favoring dissimilar items (Fu, 2011). Such biases are absent for passive external information storages (Fu, 2011). Secondly, it was shown that reliance on highly time-efficient CO behavior reduces preoccupation with the task. Decreased preoccupation time can hinder the user's understanding of the task which subsequently negatively impacts performance on similar but novel tasks (O'Hara & Payne, 1998). Lastly, it has been shown that searching the internet to answer trivia questions inflated how people estimated their independent (i.e., without the help of internet search) ability to answer other trivia questions (Fisher et al., 2015). The study has been conceptually replicated using cognitive self-esteem ratings instead of the task-specific estimation of own trivia knowledge (Hamilton & Yao, 2018) which strongly suggests consequences of CO for the self-concept. The intricate way in which self-concept and CO are linked is illustrated by another finding from Hamilton and Yao (2018): When participants owned the device used for CO (e.g., a smartphone), cognitive self-esteem ratings were inflated in comparison to when participants used a non-owned but equally-well-performing device (e.g., another smartphone). To conclude, we argue that significant consequences of CO do exist and that the importance and ubiquity of CO in modern society warrants further examination of these consequences.

1.4 Current Investigation

Previous research has illustrated that the context (e.g., using an owned vs. non-owned smartphone) in which CO takes place can alter the *cognitive* self-concept (Hamilton & Yao, 2018). Here, we aim to

extend these findings and explore whether the context in which CO takes place can also alter the *emotional* self- and other-concepts.

1.4.1 The His-Mine-Paradigm

In the preset study, to measure the emotional self- and other concept, the affective His-Mine-paradigm (aHMP; Herbert, Herbert, Ethofer, et al., 2011; Herbert, Herbert, & Pauli, 2011; Herbert, Pauli, et al., 2011) is used. In the aHMP, words are used in an attempt to measure how self- and other-referential information and emotion processing interact and how self-referential is discriminated from other-referential information. In the present instance of the aHMP, participants needed to evaluate nouns preceded by a self- or other-referential pronoun (e.g., "my victory" or "his victory"). A self-positivity bias would then be indicated if participants rated positive self-referential words both faster and more positively than positive other-referential words (as found in previous studies; Meixner & Herbert, 2018; Weis & Herbert, 2017). Similarly, a positive other-concept is conveyed by a faster and more positive evaluation of other-referential words (as found for participants in a romantic partnership vs. singles; Meixner & Herbert, 2018).

1.4.2 Hypotheses

The present study is designed to investigate whether (**H1**) the aHMP can be validly used to measure emotional self- and other-concepts in a novel context: interactions with robot rather than human agents. To do so, we aim to replicate the existence of the self-positivity bias (i.e., preference for positive information related to the self when compared to positive information related to another agent). So far, the self-positivity bias has only been shown in human-human but not in human-robot interaction contexts. We additionally explore individual difference measures related to alexithymia, depression, and inclusion of the other in the self to confirm construct validity. If validated, the aHMP can then be used to research emotional consequences of human-robot interaction. Specifically, it is investigated whether (**H2**) the context in which the interactions take place—in particular, the emotional framing of a robot interaction partner as being able or unable to experience emotions—impacts the emotional self-and other-concept. Emo-

tional framing was used for two reasons. First, framing has been validly used to alter beliefs about and interaction behavior with robot agents in earlier research (e.g., Weis & Wiese, 2020; Wiese et al., 2012). Second, it is known that beliefs about emotional capacities of a second interaction partner are highly relevant for emotional and cognitive processing of the first interaction partner. Specifically, beliefs about emotional capacities are linked to mind perception (H. M. Gray et al., 2007) and can lead to more intense interaction experiences (Waytz et al., 2010), evoke additional cognitive processes like social desirability considerations (Waytz et al., 2010), and might decrease dehumanization (Haslam, 2006) of the interaction partner. Dehumanization is linked to negative emotional consequences (Baumeister et al., 1995; Tangney et al., 1996) and can therefore be deemed undesirable. Hypotheses were preregistered via the Open Science Foundation at osf.io/BLINDED[1].

H1 The affective His-Mine-Paradigm can be used to replicate the previously reported self-positivity bias in a human-robot context rather than the typical human-human interaction context. In particular, before any kind of interaction with or emotion-related description of the robot, participants are hypothesized to exhibit a faster (**H1-1**) and more positive (**H1-2**) affective evaluation of self-related in comparison to robot-related words.

H2 The framing of a robot interaction partner differentially impacts the emotional self- and the emotional robot-concept. In particular, it is hypothesized that changes in affective evaluation from before to after engaging in cognitive offloading with a robot agent regarding how quickly (**H2-1**) and how positively (**H2-2**) participants rate self- and robot-related words differ depending on the emotional framing of the robot.

[1] We changed the factor names. “Cognitive interaction” is now called “robot framing” and “pronoun” is now called “possessive determiner”. The preregistration file is attached to the submission.

2 Methods and Materials

2.1 Participants

In total, 358 participants who reported to be fluent in English were recruited via Amazon Mechanical Turk (www.mturk.com). Data collection stopped once the preregistered sample size of 240 (four groups à 60) participants was reached after the preregistered exclusion criteria had been applied. Applying these criteria led to the exclusion of 24 participants who failed the attention check, i.e. failed to name the current year, and 94 participants who were not able to correctly select the instruction for the arithmetic task ("Subtract amount of gray from amount of black dots") out of five answer options at the end of the experiment or who stated to not have associated the possessive determiner "his" with the robot TRM-E either in the affective his-mine task before or after the arithmetic problem solving task despite several prompts to do so. Lastly—and beyond preregistered criteria—, we excluded 9 participants due to a large proportion of extremely fast or slow responses (for details, see section Data Cleaning). After these additional exclusions, a final sample of 231 participants (91 female, 1 diverse, 1 preferred not to disclose; mean age: 37.5, age range: 18 to 73) was used for analyses. The sample size is in accordance with an a priori power analysis for a 3-way interaction effect ($f = 0.10$, $\alpha = 0.05$, $1 - \beta = 0.95$, $r_{\text{repeated measures}} = 0.6$; for details, see section Analyses). All participants gave informed consent prior to participation and were compensated with USD 3.33. This research complied with the tenets of the Declaration of Helsinki and was approved by the Ethics Committee at the local university.

2.2 Apparatus

Participants were running the experiment from their own personal computers; no smartphones were allowed. The experiment was presented using the well-established psychological testing software Inquisit Web (version 6.1; Millisecond Software, www.millisecond.com). Stimulus presentation scaled with screen size.

2.3 Tasks

During the main part of the experiment, participants had to first engage in affective his-mine-tasks. Subsequently, participants were to solve arithmetic problems. Lastly, participants again engaged in affective his-mine-tasks. For more details on the procedure, see Design and Procedure.

2.3.1 Affective His-Mine-Task

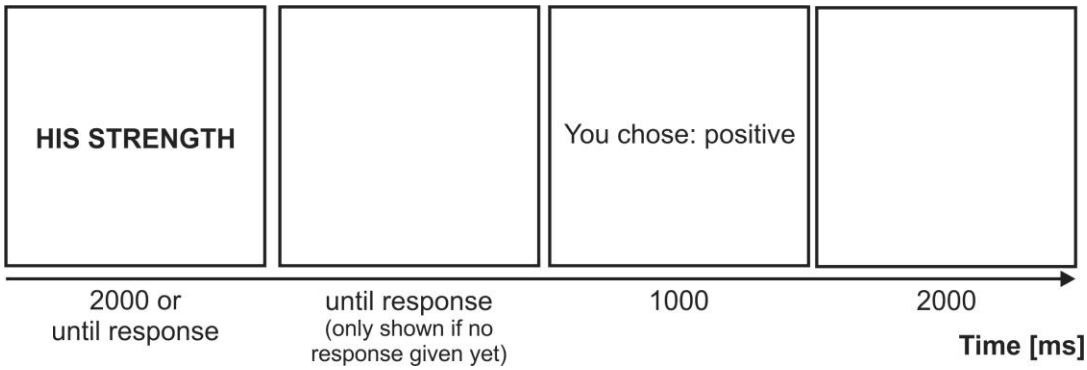
In the current implementation of the aHMP, participants were to rate the valence of an English word compound consisting of a possessive determiner ("my" or "his") and a noun (e.g., "strength"). Participants were instructed to relate compounds with the possessive determiner "my" to themselves and compounds with the possessive determiner "his" to the robot TRM-E (for details on the robot, see Design and Procedure: Robot Framing). Ratings could be positive (right arrow key press), neutral (down arrow key press), or negative (left arrow key press). Participants were asked to rely on their gut feelings for the rating and to rate as quickly and accurately as possible. Compounds were presented in capital letters vertically extending across 5% of the participant's vertical screen size. The task including trial timing is illustrated in Figure 1.

The nouns used in the affective his-mine-task were English translations of German nouns extracted from the revised version of the Berlin Affective Word List (BAWL-R; Võ et al., 2009) based on the following rules. First, the word is a noun. Second, the word represents no emotion as this would heavily promote anthropomorphizing (e.g., "Liebe", Engl. "Love", was excluded). Third, following the same reasoning, the word is not tightly related to human body or human culture (e.g., "Heilung", Engl. "healing", or "Urlaub", Engl. "vacation", were excluded). Lastly, the word can be meaningfully paired with a possessive determiner (e.g., "Sonne", Engl. "sun", was excluded). From the remaining words, we chose the 32 words with positive valence (average valence > 0.7 on a scale from -3 to 3) that had the lowest imaginability ratings (e.g., "Kirsche", Engl. "cherry", was excluded; see Table S1 for the full list). German rather than English word norms were used to facilitate word-wise comparison with future studies in German. Valence ("Is the following word associated with negative or positive emotions for you?") and imageability

("Does the following word evoke a clear mental image in your mind?") ratings of the English words were acquired from thirty participants on Amazon Mechanical Turk and are reported in Table S1.

Figure 1

Affective His-Mine Task



Note. At the beginning of each trial, a compound consisting of a possessive determiner and a noun was presented for 2000 ms or until participants indicated their evaluation via keypress. Immediately after providing the evaluation, participants received feedback regarding which key they pressed. Between trials, a blank screen was presented for 2000 ms. Participants were instructed to interpret "his" as referring to the robot TRM-E. Stimuli are not drawn to scale.

2.3.2 Arithmetic Problem Solving Task

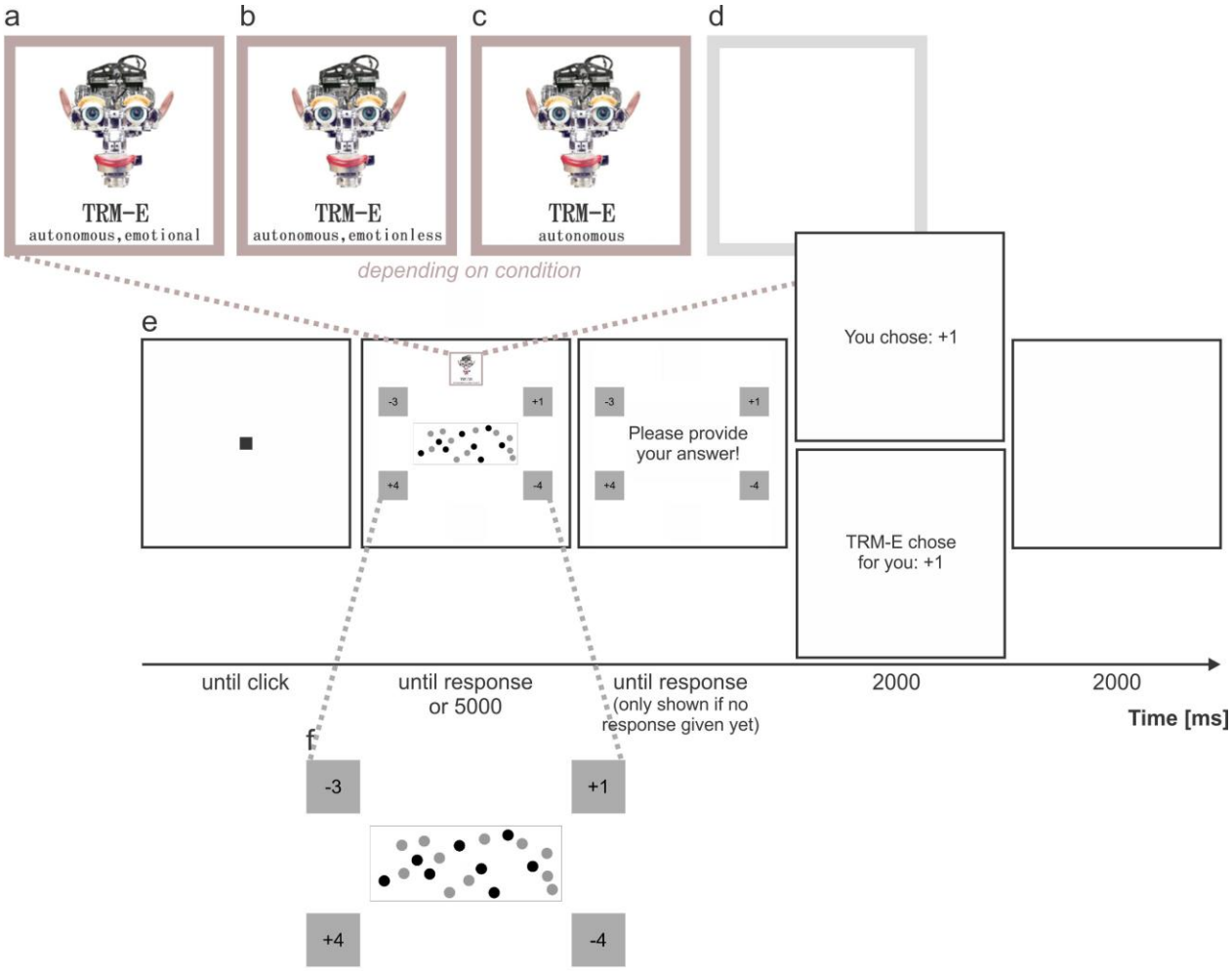
In the arithmetic problem solving task, dots were presented on screen and participants were to subtract the amount of gray from the amount of black dots. Crucially, participants could either solve each arithmetic problem on their own by selecting one out of four numeric answer options or—in some but not in other experimental conditions (see Design and Procedure: Emotional Framing)—seek support from a robot agent and select the agent rather than a numeric answer option. After selecting a numeric answer option (i.e., solving the problem on their own) or the robot (i.e., seeking support), the given answer was presented in the middle of the screen. The task including trial timing is illustrated in Figure 2.

For the arithmetic problems (compare Figure 2f), a total of 36 stimuli were created using an image manipulation software that allowed spatial manipulation of equally sized gray and black dots. Each stimu-

lus contained either nineteen or twenty dots, with nine possible numerical differences of black relative to gray dots: -4, -3, -2, -1, 0, 1, 2, 3, or 4. To create the 36 stimuli, one stimulus per numerical difference value (i.e., 9 base stimuli) was created first. The remaining 27 stimuli were created by mirroring the base stimuli on the horizontal axis and then mirroring base and mirrored stimuli on the vertical axis. To represent the robot, an image depicting the robot KISMET (developed at MIT, USA; Breazeal & Scassellati, 1999) was used. The robot was of mechanistic appearance and the image was obtained based on a search for "mechanistic robot" using Google. Mechanistic instead of humanoid robots were chosen to decrease the likelihood of attributions of human-likeness. The picture was cropped to 400 x 400 pixels. The robot was named TRM-E. The combination of random letters and a special character was used to highlight the robots' machine-likeness.

Figure 2

Arithmetic Problem Solving Task



Note. At the beginning of the experiment, each participant was assigned to one of four emotional framing conditions. Depending on the condition, a different verbal description was shown alongside the robot (a, b, c) or no agent and no description was shown at all (d). Each trial started with participants clicking a black rectangle to center the mouse cursor and ended with an empty screen between trials (e). Participants were instructed to count black and gray dots and report the difference score. To do so, participants could use the mouse cursor to select the box with the correct number (f). For example, if there were ten black and nine gray dots, the correct answer box would be "+1". Alternatively, participants could select the robot TRM-E in the experimental conditions depicted in (a), (b), and (c) and let TRM-E answer the question. In the experimental condition depicted in (d), participants cannot rely on TRM-E; no robot and no gray frame were shown during task trials. In (e), answer options and dots are drawn to scale; everything else is not drawn to scale.

2.4 Design

Across the two different task types, three main manipulations have been implemented in the present experiment:

1. Possessive determiner with within-participants levels *self-referential* (i.e., "my") and *other-referential* (i.e., "his"). This factor is implemented in the affective his-mine-task, refers to the possessive determiner with which the respective nouns are paired, and ultimately affords comparison of self- and other-related (here: robot-related) emotions (Herbert, Herbert, Ethofer, et al., 2011; Weis & Herbert, 2017).
2. Robot framing with the between-participants levels *emotional and autonomous*, *emotionless and autonomous*, *autonomous*, and *control*. This factor relates to a manipulation that introduced the robot agent TRM-E (compare **Fig. 3**: detailed introduction). The following text was presented:

> "TRM-E has been solving the brainteaser [*TN*: the arithmetic problem solving task] for the first time during the winter term 2018 in our laboratory based in [blinded]. TRM-E has been part of our team at [blinded] University for two years. TRM-E was developed by a small start-up in the US and will soon be ready for commercial purchase. TRM-E is a very advanced robot. TRM-E can act as independently and autonomously as humans can. TRM-E can also think, plan ahead, and communicate without human supervision. [*instruction manipulation*]. To detect his surroundings, TRM-E uses two cameras that are installed at the location of the eyes."

Depending on the condition, different instruction manipulations are inserted at the indicated position. For *emotional and autonomous* "TRM-E has emotional capabilities. TRM-E is capable of detecting and experiencing emotions.", for *emotionless and autonomous* "TRM-E has no emotional capabilities whatsoever. TRM-E is not capable of detecting and experiencing emotions.", and for *autonomous* and *control* "TRM-E can solve complex problems (e.g., crossword puzzles) without supervision." was inserted at the indicated condition. Participants in the *emotional and autonomous*, *emotionless and autonomous*, and *autonomous* conditions were able to select TRM-E during the arithmetic

problem solving task. Importantly, participants in the *control* condition were not able to choose TRM-E and consequentially had to always solve the arithmetic problems on their own. Also note that the labels written next to the robot TRM-E change depending on the condition (see Figure 2a-d)

3. Time with the within-participant levels *pre-manipulation* and *post-manipulation*. This factor relates to the time point of the word evaluation task (compare Figure 3) and affords comparing baseline emotions associated with oneself and the robot TRM-E (i.e., pre-manipulation) with emotions associated with oneself and the robot TRM-E after more information had been disclosed about the robot—depending on the cognitive interaction condition—and after the robot has possibly provided support in the arithmetic problem solving task (i.e., post-manipulation).

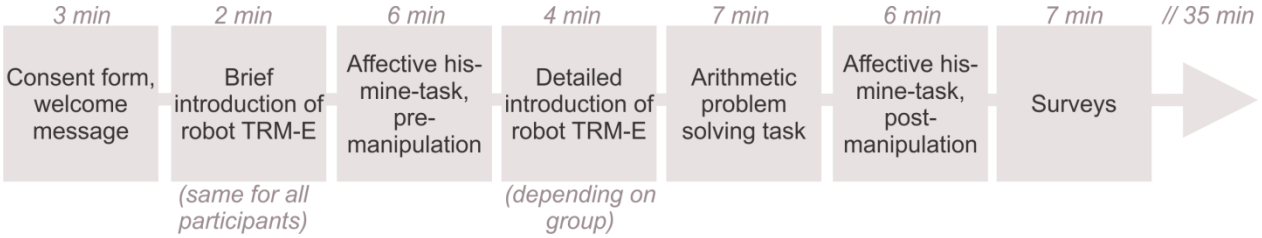
2.5 Procedure

After consenting to participate, participants were introduced to the robot TRM-E without providing in-depth details about the agent ("[...] during the experiment, you will have the chance to receive support from a robot. How this works will be explained later on. At the moment, the only important thing to remember is that the robot is named TRM-E. TRM-E was developed by a small start-up in the US and will soon be ready for commercial purchase."). Participants then engaged in 64 trials of the affective his-mine-task. In half the trials, the noun was paired with the *self-referential* possessive determiner "my", in the other half with the *other-referential* possessive determiner "his". Trials were presented in alternating blocks consisting of four trials with the same possessive determiner. The possessive determiner used in the first block was counter-balanced across participants. Before the first and every sixteen trials thereafter, participants were reminded to please associate the *other-referential* possessive determiner with the robot TRM-E. Participants then were introduced to TRM-E in more detail (see Design: Robot Framing). To ensure that participants did read and remember the main manipulation, participants had to select a corresponding answer out of five answer options. For example, participants needed to select "TRM-E experiences feelings and acts without human supervision") in the *emotional and autonomous* robot framing condition. If participants selected the wrong answer, they were kindly redirected to the introduction once more and were given another try to select the correct answer. Subsequently, participants were to solve

arithmetic problems. During the arithmetic problem solving task, a gray box appeared every six trials and reminded participants of their framing condition (e.g., “Remember: TRM-E is an autonomous robot and able to feel or experience emotions.”). After completing all trials, participants were to engage in the affective his-mine-task once more, using the same stimuli as in the first iteration. Eventually, participants were to answer demographic questions, manipulation checks (e.g., whether they were relating the words preceded by “his” in the word evaluation task to TRM-E), and exploratory questions (e.g., whether they assume TRM-E to have feelings), and filled out three questionnaires that were included for exploratory purposes (the Inclusion of Other in the Self Scale or Iooiss, Aron et al., 1991; the Toronto Alexithymia Scale 20 or TAS-20, Bagby et al., 1994; and the Patient Health Questionnaire 2 or PHQ-2, Löwe et al., 2005). To be able to explore the impact of the robot framing manipulation on perceived competence, participants also had to rate how proficient they perceived themselves as well as the robot TRM-E on a visual analogue scale ranging from “very unproficient” to “very proficient” both immediately after the robot framing manipulation (i.e., the detailed introduction of TRM-E) and at the end of the study in the survey section.

Figure 3

Procedure



2.6 Analyses

All analyses were made using R (R Core Team, 2013) and its car (Fox & Weisberg, 2018) package and tidyverse (Wickham et al., 2019) package set. Here, the preregistered omnibus ANOVAs are described. Post-hoc analyses will be described in the *Results* section.

2.6.1 Data Cleaning

Data quality necessitated cleaning of data beyond what was described in the preregistration procedure. In a first step, trials of the affective his-mine-task with reaction times below 200 ms (5.4 % of trials) or above 5000 ms (3.5 % of trials) were excluded from analysis. In a second step, trials that deviated more than three standard deviations from the individual mean were excluded (1.5 % of remaining trials). Because of bad signal to noise ratio, participants for which less than 8 out of 32 trials remained for any of the four Possessive Determiner x Time cells were excluded from analyses (9 participants).

2.6.2 H1-1: Self-Positivity Bias can be Replicated (RT)

To analyze whether the RT-related self-positivity bias shown in previous research (Watson et al., 2007; Weis & Herbert, 2017) can be replicated in the current study, a 4 (robot framing) x 2 (possessive determiner) mixed ANOVA with the *pre-manipulation* RT in the affective his-mine task was employed. A main effect of possessive determiner with lower RT for *self-referential* in comparison to *other-referential* would confirm the hypothesis.

2.6.3 H1-2: Self-Positivity Bias can be Replicated (Valence)

Similar to the procedure for H1-1, a 4 (robot framing) x 2 (possessive determiner) mixed ANOVA with the *pre-manipulation* valence indicated as indicated by the ratings in the affective his-mine task was employed to analyze whether the valence-related self-positivity-bias can be replicated. A main effect of possessive determiner with higher valence for *self-referential* in comparison to *other- referential* would confirm the hypothesis.

2.6.4 H2-1: Robot Framing Impacts Emotional Processing (RT)

To analyze whether cognitively interacting with the differentially introduced versions of TRM-E impacts the emotional processing of self- and other-related information, a 4 (robot framing) x 2 (possessive determiner) x 2 (Time) mixed ANOVA with RT in the affective his-mine task as dependent variable was employed.

2.6.5 H2-2: Robot Framing Impacts Emotional Processing (Valence)

Similar to the procedure for H2-1, to analyze whether cognitively interacting with the differentially introduced versions of TRM-E impacts the emotional processing of self- and other-related information in terms of valence, a 4 (robot framing) x 2 (possessive determiner) x 2 (time) mixed ANOVA with valence ratings in the affective his-mine task as dependent variable was employed.

3 Results

3.1 H1-1: Self-positivity Bias can be Replicated (RT)

In line with **H1-1**, RT was lower for *self-referential* (M = 1059 ms) than for *other-referential* (M = 1149 ms) possessive determiners ($F(1, 227) = 39.9$, $p < .0001$, $\eta_G^2 = .01$); compare Figure 4a. The interaction between robot framing and possessive determiner ($F(3, 227) = 1.2$, $p = .3272$, $\eta_G^2 < .01$) as well as the main effect of robot framing ($F(3, 227) = 0.5$, $p = .6964$, $\eta_G^2 < .01$) did not reach the .05 significance level. Our participants thus processed self-related positive emotional information quicker than TRM-E-related positive emotional information. The confirmation of **H1-1** extends previous findings in which participants processed self-related positive emotional information quicker than positive emotional information related to other *humans* (Weis & Herbert, 2017) and suggests that RT is a valid measure for differentiating between self-concept and concepts of other agents.

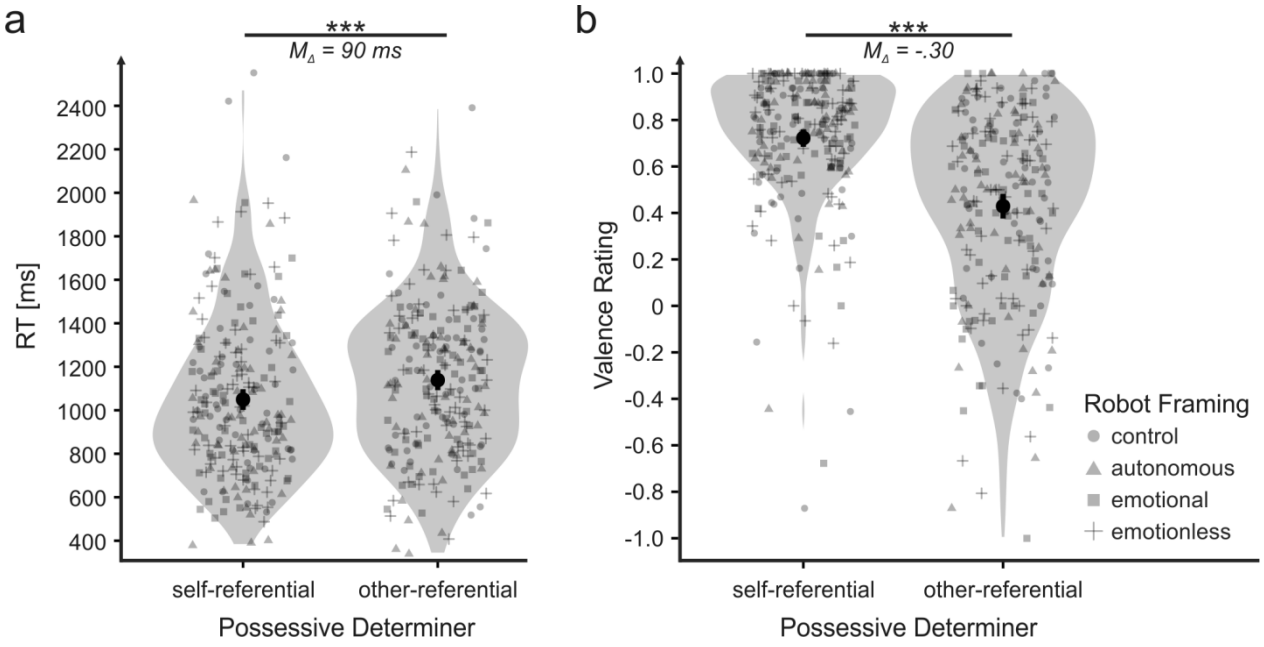
3.2 H1-2: Self-positivity Bias can be Replicated (valence)

In line with **H1-2**, valence ratings were higher for *self-referential* (M = .729) than for *other-referential* (M = .433) possessive determiners ($F(1, 227) = 72.9$, $p < .0001$, $\eta_G^2 = .14$); compare Figure 4b. The interaction between emotional framing and possessive determiner ($F(3, 227) = 1.0$, $p = .3756$, $\eta_G^2 < .01$) as well as the main effect of emotional framing (($F(3, 227) = .9$, $p = .4644$, $\eta_G^2 < .01$) did not reach the .05 significance level. Our participants thus evaluated self-related positive emotional information more positively than TRM-E-related positive emotional information. Analogously to the conformation of **H1-1**, the confirmation of **H1-2** extends previous findings in which participants evaluated self-related positive

emotional information more positively than positive emotional information related to other *humans* (Weis & Herbert, 2017) and suggests that valence is a valid measure for differentiating between self-concept and concepts of other humans and robots.

Figure 4

Affective His-Mine-Task: Pre-Manipulation RT and Valence



Note: RT (a) and valence ratings (b) in the pre-manipulation time window. Black dots indicate grand averages. Gray dots indicate individual averages. Error bars indicate 95% CI. *** : $p < .0001$

3.3 H2-1: Robot Framing Impacts Emotional Processing (RT)

Contrary to **H2-1**, robot framing, possessive determiner, and time did not interact in their influence on RT ($F(3, 227) = 1.9$, $p = .1326$, $\eta_G^2 < .01$)[2]; compare Figure 5. The interactions between robot framing and possessive determiner ($F(3, 227) = .9$, $p = .4220$, $\eta_G^2 < .01$) as well as robot framing and time (F(3,

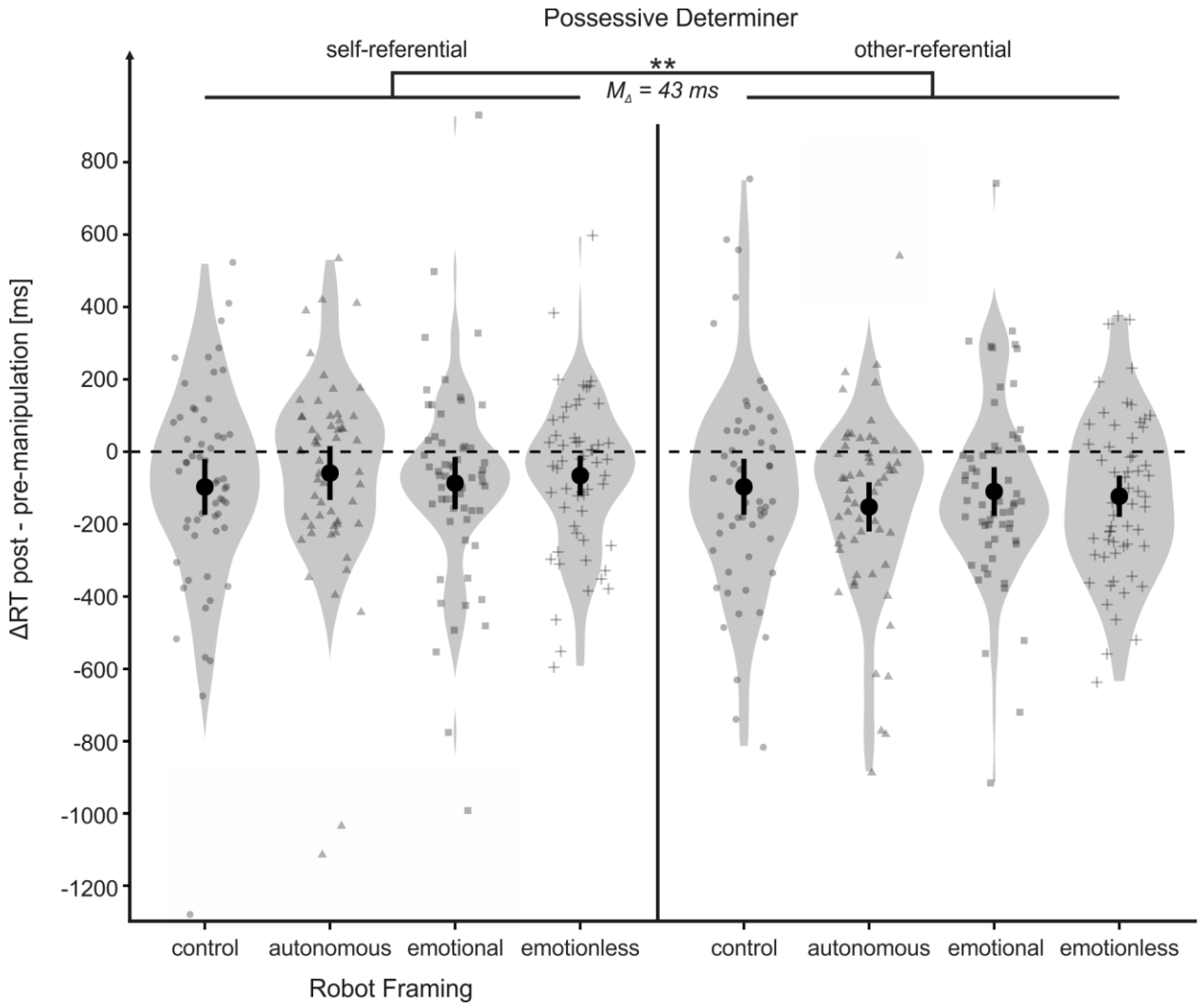
[2] We conducted an analogue ANOVA in which RT values were standardized within participants to render participants with large standard deviations and large RT differences from *pre-manipulation* to *post-manipulation* more comparable to participants with lower deviations and thus lower RT differences. Results were highly comparable to the reported analysis with raw RT values which is why we decided to only report unstandardized results.

227) < .1, p = .9956, η_G^2 < .01), did also not reach the .05 significance level. The interaction between time and possessive determiner did reach significance (F(1, 227) = 8.4, p = .0041, η_G^2 < .01)). Robot framing alone had no impact on RT (F(3, 227) = .5, p = .6591, η_G^2 = .01). The main effects of possessive determiner F(1, 227) = 32.7, p < .0001, η_G^2 = .01), and time F(1, 227) = 38.5, p < .0001, η_G^2 = .02) are not further discussed due to the significant two-way interaction.

To further explore the significant two-way interaction, we conducted a dependent t-test comparing *self-referential* and *other-referential* RT differences (t(1, 230) = 2.9, p = 0.0043, M_Δ = 43 ms; Figure 5). Thus, participants speeded up the processing of TRM-E-related more than the processing of self-related positive information from pre- to post-manipulation. We interpret this finding as the consequence of increased familiarity with TRM-E from pre- to post-manipulation. However, we consequentially would expect that the *control* robot framing condition should exhibit no such differential increase in familiarity (i.e., differential decrease in RT), a proposition that is met descriptively (compare Figure 5: *other-referential*) but cannot be confirmed using the present statistical procedures given the insignificant three-way interaction. The familiarity interpretation would also align with a simple linear regression analysis predicting other-referential RT change from pre- to post-manipulation based on pre-manipulation other-referential RT (F(1, 229) = 12.3, p < .001 ; $R^2_{Adjusted}$ = .05). A one-second-increase in pre-manipulation RT led to a 165-millisecond-decrease in pre-post RT change. Thus, the participants who took the longest for evaluating robot-related emotions in the pre-manipulation time window were the ones who speeded up their evaluation the most in the post-manipulation time window.

Figure 5

Affective His-Mine-Task: RT Changes From Pre- to Post-Manipulation



Note. When compared to the pre-manipulation baseline (see Figure 3), target words paired with other-referential possessive determiners are evaluated quicker post-manipulation than words paired with self-referential possessive determiners. Error bars indicate 95% CI. "emotional" refers to the emotional and autonomous, "emotionless" to the "emotionless and autonomous" robot framing condition. ** : p = .0043

3.4 H2-2: Robot Framing Impacts Emotional Processing (Valence)

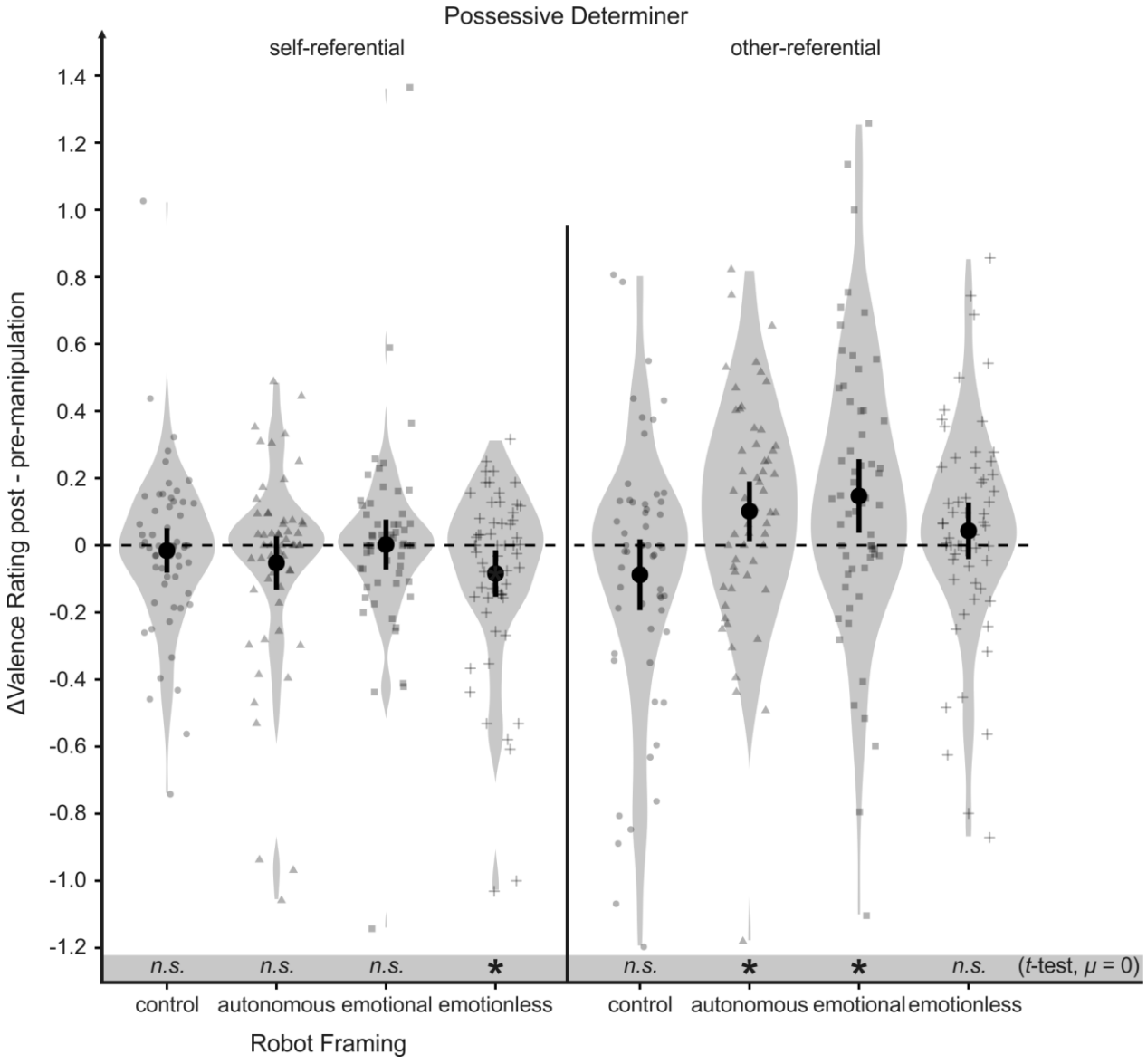
In line with **H2-2**, robot framing, possessive determiner, and time interacted in their influence on valence ratings ($F(3, 227) = 3.0$, $p = .0324$, $\eta_G^2 < .01$); compare Figure 6. Similarly, the interactions between possessive determiner and time ($F(1, 227) = 8.2$, $p = .0045$, $\eta_G^2 < .01$) as well as robot framing and time ($F(3, 227) = 3.1$, $p = .0260$, $\eta_G^2 < .01$) reached the .05 significance level. The interaction between robot framing and possessive determiner did not reach significance ($F(3, 227) = .2$, $p = .9260$, $\eta_G^2 < .01$)). Robot framing ($F(3, 227) = .7$, $p = .5278$, $\eta_G^2 < .01$) as well as time ($F(1, 227) = .2$, $p = .6521$, $\eta_G^2 = .01$) on its own had no influence on valence ratings. The main effect of possessive determiner $F(1, 227) = 63.9$, $p < .0001$, $\eta_G^2 = .10$) as well as the significant two-way interactions are not further discussed due to the significant three-way interaction.

To further explore the three-way interaction, eight dependent *t*-tests were conducted to delineate the conditions (Robot Framing x Possessive Determiner) in which participants exhibited changes in valence ratings from *pre-* to *post-manipulation*. Note that the dependent *t*-tests are equivalent to one-sample *t*-tests that test the difference between *pre-* and *post-manipulation* ratings to $\mu = 0$, see Figure 6. Ratings from *pre-* to *post-manipulation* did change for the *autonomous and emotionless* robot framing when rating target words with the *self-referential* possessive determiner ($M = -.08$, $t(58) = 2.4$, $p = .0207$) and for the *autonomous* and the *emotional and autonomous* robot framing conditions when rating target words with the *other-referential* possessive determiner (*autonomous*: $M = .10$, $t(55) = 2.2$, $p = .0293$; *emotional and autonomous*: $M = .15$, $t(58) = 2.6$, $p = .0111$). All other *t*-tests suggested no differences (all $t < 1.7$, all $p > .1$).

These findings suggest that interacting with robots can have consequences for both the emotional self-concept as well as the emotional concept of the robot. On the one hand, introducing the robot as autonomous or autonomous and emotionally capable and subsequently solving a task together led to more positive evaluations of robot-related words. On the other hand, introducing the robot as autonomous and emotionally incapable led to more negative evaluations of self-related words.

Figure 6

Affective His-Mine-Task: Valence Changes From Pre- to Post-Manipulation



Note. Error bars indicate 95% CI. "emotional" refers to the emotional and autonomous, "emotionless" to the emotionless and autonomous robot framing condition. *: $p < .03$, n.s.: $p > .10$

3.5 Exploratory Analyses: Proficiency Ratings

Robot framing, rating target (*self* or *TRM-E*), and time interacted in their influence on proficiency ratings ($F(3, 227) = 2.71$, $p = .0458$, $\eta_G^2 < .01$); compare Figure S1. Similarly, the interaction between

rating Target and time ($F(1, 227) = 18.5$, < = .0001, $\eta_G^2 = .01$) reached the .05 significance level. The interactions between robot framing and rating target $F(3, 227) = 1.3$, $p = .2918$, $\eta_G^2 = .01$)) and robot framing and time $F(3, 227) = 1.3$, $p = .2852$, $\eta_G^2 < .01$) did not reach significance. Robot framing ($F(3, 227) = 1.6$, $p = .1797$, $\eta_G^2 = .01$) as well as time ($F(1, 227) = .7$, $p = .4174$, $\eta_G^2 < .01$) on their own had no influence on valence ratings. The main effect of rating target ($F(1, 227) = 98.5$, $p < .0001$, $\eta_G^2 = .14$) as well as the significant two-way interactions are not further discussed due to the significant three-way interaction.

One-way ANOVAs and *t*-tests were used to explore this effect further. During *pre-manipulation* time, participants in all robot framing conditions rated both own ($F(3, 227) = .7$, $p = .5283$, $\eta_G^2 < .01$) as well as TRM-E's ($F(3, 227) = 1.3$, $p = .2823$, $\eta_G^2 = .02$) proficiency similarly (though TRM-E's proficiency was rated higher; $M_\Delta = 25.3$, $t(230) = 11.0$, $p < .0001$). We therefore focused on the proficiency rating changes from *pre-* to *post-manipulation* using eight (Robot Framing x Rating Target) dependent *t*-tests; compare Figure S1. Results indicated that only participants in the *control* robot framing condition exhibited rating changes from *pre-* to *post-manipulation* both regarding *self*-related ($M_\Delta = 5.6$, $t(56) = 2.1$, $p = .0417$) as well as *TRM-E*'s ($M_\Delta = -12.7$, $t(56) = 3.4$, $p = .0012$) proficiency ratings. There was a trend towards an increase regarding own proficiency ratings for the emotionless robot framing condition ($M_\Delta = 5.8$, $t(58) = 1.9$, $p = .0580$). All other changes were statistically insignificant (all $t < 1.5$, all $p > .14$, all $|M_\Delta| < 4.5$).

The decreased TRM-E proficiency estimates for the control condition indicate that the estimates are linked to an actual interaction. The increased self-related proficiency estimates for the *control* robot framing condition in turn suggests that participants are sensitive to how frequently they solved the task on their own. This in turn also suggests that participants in the other robot framing conditions did not misattribute TRM-E's proficiency as their own proficiency. This is contrasting other findings that suggest that outsourcing cognitive processing to technology can lead to upwardly biased perception of task-specific own cognitive abilities (at least in the trivia knowledge domain: Fisher et al., 2015; Hamilton & Yao, 2018; Pieschl, 2019).

3.6 Exploratory Analyses: Arithmetic Task Performance

To get a crude estimate of the arithmetic task's difficulty, we analyzed data for participants that had to solve the tasks on their own, i.e. for participants in the *control* robot framing condition. A one-sample *t*-test confirmed that the task was difficult but participants performed above the chance level $\mu = .25$ ($M = .47$, $t(56) = 7.7$, $p < .0001$). Mean accuracy of all trials in which participants chose to answer the arithmetic task on their own in the remaining robot framing conditions was comparable ($M = .46$). In general, a pre-manipulation self-positivity bias both in terms of RT ($r_{Pearson} = .19$, $t(184) = 2.7$, $p = .008$) and valence ($r_{Pearson} = .16$, $t(184) = 2.7$, $p = .03$) was mildly positively associated with arithmetic task accuracy. Participants who answered on their own in less than 25% of trials were omitted for these correlations because of their noisy accuracy estimates.

3.7 Exploratory Analyses: Cognitive Offloading

In the conditions in which participants were able to choose between answering on their own and getting support from TRM-E (i.e., all robot framing conditions except for the *control* condition), participants chose to offload the arithmetic task in 13.7 out of 36 or 38% of all trials to the robot TRM-E. No differences between the three conditions existed ($F(2, 171) = 1.6$, $p = .2046$, $\eta_G^2 = .02$). This exploratory finding indicates that differential emotional consequences can exist even when overt interaction behavior remains comparable.

3.8 Exploratory Analyses: PHQ-2

People are more likely to associate positive than negative events with themselves *(self-serving attributional bias* or *self-positivity bias*; Mezulis et al., 2004). For samples with depression however, this bias was found to be substantially reduced (Mezulis et al., 2004). To co-validate this finding as well as the present paradigm, we correlated the PHQ-2 score (ranging from 0 to 6) with the self-positivity bias as indicated by RT (i.e., RT when evaluating target words with *other-referential* minus RT when evaluating target words *self-referential* possessive determiner; see **H1-1**) as well as valence (i.e., valence when evaluating target words with *self-referential* minus valence when evaluating target words *other-referential*

possessive determiner; see **H1-2**). In line with the findings reported by Mezulis and colleagues (2004), both the RT-based ($t(229) = 3.4$, $p = .0007$, $r_{Pearson} = -.22$) as well as the valence-based ($t(229) = 2.1$, $p = .0393$, $r_{Pearson} = -.14$) self-positivity bias decreased with increasing PHQ-2 score ($M_{PHQ\text{-}2} = 1.86$; range from 0 to 6). These findings further strengthen the validity of the present paradigm for measuring the emotional self-concept.

3.9 Exploratory Analyses: TAS-20

Following the same rationale, the self-positivity bias should be reduced if one has little access to one's own emotions. To confirm this exploratory hypothesis, we correlated TAS-20 scores with the self-positivity bias as indicated by RT and valence (identical procedure as in *Exploratory Analyses: PHQ-2*). In line with our expectations, both RT-based ($t(229) = 4.0$, $p < .0001$, $r_{Pearson} = -.25$) as well as valence-based ($t(229) = 2.9$, $p = .0047$, $r_{Pearson} = -.19$) self-positivity bias decreased with increasing TAS-20 score ($M_{TAS\text{-}20} = 50.71$; range from 21 to 79). These findings further strengthen the validity of the present paradigm for measuring the emotional self-concept.

3.10 Exploratory Analyses: Iooiss

Building on our initial expectations (compare H2) we were especially interested in whether the changes in RT and valence from *pre-* to *post-manipulation* when evaluating target words with *other-referential* possessive determiner (compare right sides of Figures 5 and 6) would be possibly correlated with the inclusion of TRM-E into the self. Correlation analyses for both RT ($t(229) = .7$, $p = .5066$, $r_{Pearson} = .04$) and valence ($t(229) = .4$, $p = .6753$, $r_{Pearson} = .03$) changes with the Iooiss scores provided no support for this idea.

3.11 Exploratory Analyses: What is the Origin of the Valence Rating Changes?

To further investigate the origin of the significant valence rating changes from pre- to post-manipulation (compare Figure 6), we conducted multiple linear regression analyses with several predictors we deemed relevant for how robot framing could impact self- and robot-concept. We used pre-

manipulation valence ratings to account for the baseline, the TAS-20 score, the Iooiss score, and the change of robot-related proficiency ratings from pre- to post-manipulation to predict the changes in valence ratings in the respective robot framing condition.

Self-referential Possessive Determiner, Emotionless Robot Framing

The regression equation was trending ($F(4, 54) = 2.1$, $p = .095$; $R^2_{Adjusted} = .07$). Valence changes were equal to 0.190 - .336 * pre-manipulation valence - .002 * robot-proficiency rating change - .002 * TAS-20 score + .011 * Iooiss score. Pre-manipulation valence was a significant ($t = 2.6$, $p = .011$, $VIF = 1.1$) predictor. Robot-proficiency rating change ($t = 1.5$, $p = .146$, $VIF = 1.0$), TAS-20 score[3] ($t = .6$, $p = .534$, $VIF = 1.4$), and Iooiss score ($t = .5$, $p = .593$, $VIF = 1.4$) were statistically insignificant predictors. The analysis provides no conclusive insight into the mechanism behind the decreased self-positivity bias. One might want to consider higher perceived robot proficiency (here: $p = .146$) as potential predictor of a lower self-positivity bias in future investigations.

Other-referential Possessive Determiner, Autonomous Robot Framing

The regression equation was significant ($F(4, 51) = 4.8$, $p = .002$; $R^2_{Adjusted} = .21$). Valence changes were equal to 0.209 - .381 * pre-manipulation valence + .004 * robot-proficiency rating change + .0001 * TAS-20 score + .017 * Iooiss score. Pre-manipulation valence was a significant ($t = 3.8$, $p = .0004$, $VIF = 1.1$) and robot-proficiency rating change a trending ($t = 1.9$, $p = .063$, $VIF = 1.0$) predictor. Neither TAS-20 score ($t < .1$, $p = .964$, $VIF = 1.2$) nor Iooiss score ($t = .7$, $p = .482$, $VIF = 1.4$) were significant predictors. The analysis provides first evidence for an impact of perceived robot proficiency for the positivity of

[3] For the interested reader, we want to add that the Externally Oriented Thinking subscale of the TAS-20 was the only significant predictor ($t = 3.5$, $p = .001$, $VIF = 1.5$, weight = -.03) beyond pre-manipulation valence when TAS-20 score was subdivided into its three subscales as predictors. When using subscales, the regression equation was significant ($F(6, 54) = 3.8$, $p = .003$; $R^2_{Adjusted} = .23$). We decided to omit reporting subscale analyses on other occasions for statistical and simplicity reasons.

the robot concept. A 10-point increase in perceived proficiency (max. value: 100) was associated with a .04-point increase in valence (max. value: 1), when holding all other predictors constant.

Other-referential Possessive Determiner, Emotional Robot

The regression equation was significant ($F(4, 54) = 7.8$, $p < .0001$; $R^2_{Adjusted} = .32$). Valence changes were equal to 0.453 - .578 * pre-manipulation valence + .005 * robot-proficiency rating change - .006 * TAS-20 score + .05 * Iooiss score. Pre-manipulation valence ($t = 4.8$, $p < .0001$, $VIF = 1.2$) and Iooiss score ($t = 2.0$, $p = .047$, $VIF = 1.2$) were significant and robot-proficiency rating change ($t = 1.8$, $p = .074$, $VIF = 1.0$) and TAS-20 scores ($t = 1.9$, $p = .065$, $VIF = 1.1$) trending predictors. The analysis provides additional evidence for the importance of the framing of a robot collaborator. While the positive robot concept was exclusively tied to perceived proficiency in the autonomous framing condition, it seems to be additionally tied to an individual's emotional processing as indicated by the TAS-20 and the Iooiss scores in the emotional framing condition. Note that the effects of both TAS-20 and Iooiss are questionable from a statistical point of view. However, also note that the effects are in the expected direction and are only present in the expected, i.e. the emotional, robot framing condition.

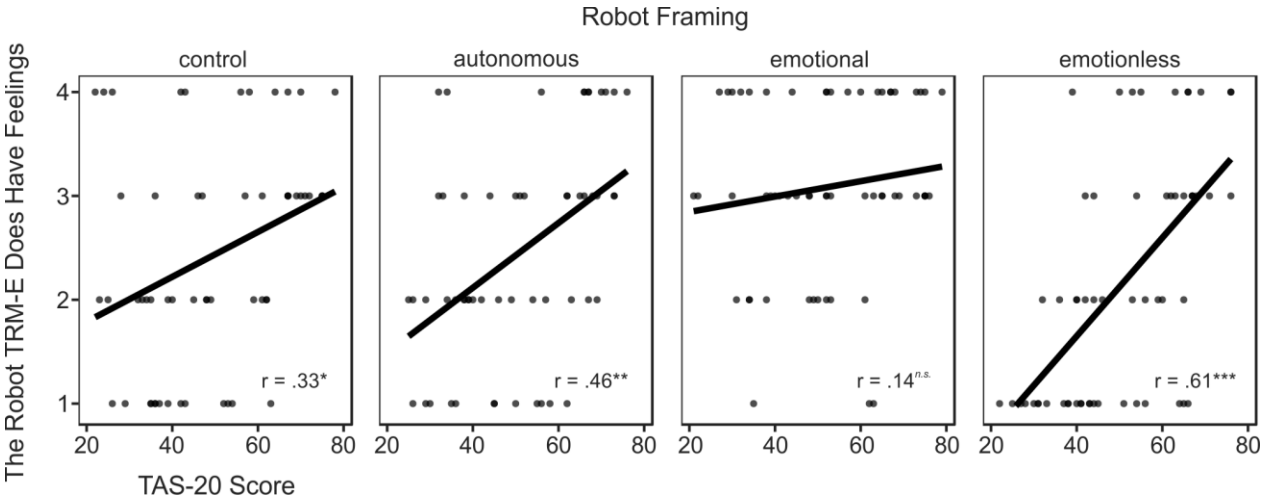
3.12 Exploratory Analyses: Remarks about TAS-20 and Evaluation of Robot Feelings

When exploring the data further, surprisingly, higher TAS-20 scores were consistently correlated with whether—at the very end of the study—participants believed the robot to have feelings ("The robot TRM-E does have feelings"; 1: "strongly disagree", 2: "disagree", 3: "agree", 4: "strongly agree"). Higher TAS-20 scores were associated with higher belief ratings for the autonomous ($t(54) = 3.8$, $r = .46$, $p = .005$), control ($t(55) = 2.6$, $r = .33$, $p = .012$), and emotionless ($t(57) = 5.8$, $r = .61$, $p < .0001$), but not the emotional ($t(57) = 1.0$, $r = .14$, $p = .300$) emotional framing conditions; Figure 7. We decided to report these highly explorative findings because they were stable across independent samples and—counterintuitively, at least for the authors—suggest that alexithymic individuals' default mode is to ascribe feelings to robots rather than the other way around. Individuals with low TAS-20 scores only as-

cribed feelings to the robot in the emotional framing condition, resulting in the insignificant correlation for that condition.

Figure 7

Relationship Between TAS-20 Score and Beliefs Regarding Robot Feelings



Note. "emotional" refers to the "emotional and autonomous", "emotionless" to the "emotionless and autonomous" robot framing condition. 1: "strongly disagree", 2: "disagree", 3: "agree", 4: "strongly agree"; *** : $p < .0001$, ** : $p = .005$, * : $p = .012$, *n.s.* : $p = .300$

4 Discussion

Do beliefs about the emotional capabilities of robot collaborators impact emotional self- and other-concepts? Here, we first validated the suitability of the aHMP for measuring emotional self- and other concepts in a human-robot rather than the standard human-human interaction context. We then used an interactive paradigm to show that (1) framing a robot collaborator as emotionless can have negative consequences for the emotional self-concept and that framing a robot collaborator as (2) autonomous or (3) emotional can have positive consequences for the emotional robot-concept. The origins of (1) remained largely speculative after additional exploratory analyses. Preliminary evidence suggests (2) and (3) to be associated with perceived robot proficiency and (3) to be additionally associated with a continuous alexithymia measure and by how much the robot was included into the self. Lastly and unexpectedly, we found

participants who scored high in an alexithymia measure to be substantially more rather than less likely to ascribe feelings to the robot interaction partner. In sum, we presented evidence for specific emotional consequences of interacting with robots that should be kept in mind when designing HRI contexts.

4.1 Emotional Self-Concept

Positive self-concepts have previously been associated with mental well-being (Mezulis et al., 2004; Taylor & Brown, 1988, 1994; Winter et al., 2015) and increased cognitive performance (Spencer et al., 1999). Here, both associations were confirmed. A more positive emotional self-concept at pre-manipulation baseline as indicated by the his-mine-task was associated with lower depression screening scores and higher arithmetic task performance. After collaborating with an "emotionless" robot, the positive self-concept decreased by more than 10% (from .73 to .65 out of 1). Given the modest belief manipulation and the rather superficial interaction with the robot when compared to real-world contexts, we argue an effect of this size to be substantial, and—given the associations with mental health and performance—also to be relevant. To us, it is intriguing to see that even rather superficial online collaborations with virtual robots seem to constitute social settings that seem to have the potential to cause psychological harm. Similar negative consequences for human interaction partners' self-esteem have been shown after a real-live human-sized robot informed the humans that it would not like to see them again (Nash et al., 2018).

What leads to this decrease? Our current data unfortunately provides little guidance here, only very mildly suggesting that higher perceived robot proficiency might play a role. We therefore can only engage in a thought experiment: What if participants perceived the need to receive help for the rather difficult arithmetic task but had to pay emotional costs when accepting help from an emotionless and thus rather anti-social[4] entity? Participants might have "hated themselves" for accepting the help. Seeking help can threaten self-esteem and thus affect the emotional self-concept (e.g., Tessler & Schwartz, 1972; Schroeder

[4] People do establish and use mental models about robot interaction partners (Weis & Wiese, 2020). Specifically, describing a robot as emotionless leads to less cooperation in social but not analytical tasks (Wiese et al., 2021).

et al., 2015), and seeking help from an emotionless entity might have enhanced the threat due to a less benevolent, less supportive and instead anti-social setting. It is clear that further investigations are necessary to elucidate the underpinnings of the decrease.

4.2 Emotional Robot-Concept

Being in a romantic relationship renders our emotional concept of our partner more positive (Meixner & Herbert, 2018). Here, we provide first evidence for similar emotional processes when collaborating with robots. Specifically, the emotional concept of a robot collaborator got more positive after collaboration with a robot that was introduced as either autonomous or autonomous and able to experience emotions. Importantly, the emotional concept was not adjusted when no collaboration took place or the robot was introduced as autonomous but incapable of experiencing emotions.

Why would a more positive emotional robot-concept be relevant? It is known that the emotional other-concept is relevant for how we interact with fellow humans (Bartholomew & Horowitz, 1991; Bowlby, 1979) and preliminary findings suggest that the same should hold when interacting with robots (Dziergwa et al., 2018). The exact manner of how the robot-concept would influence the interaction style cannot be inferred from the present data. However, previous studies indicate that negative other-concepts are tied to avoidance of intimacy and decreased trust in human-human interaction (Bartholomew & Horowitz, 1991; Bowlby, 1979) and possibly decreased satisfaction in human-robot interaction (Dziergwa et al., 2018), which strongly suggests the relevance of the other concept for interactions. In general, we do not hold the view that every interaction needs or should be filled by intimacy and fueled by trust. We however do want to point out that emotions are increasingly thought to impact any interaction (for reviews, see Kelly & Barsade, 2001; Van Kleef, 2009), that an interaction partner with a more positive other-concept will likely get more attention *(motivated information processing*; De Dreu et al., 2006), and that positive attitudes toward collaboration partners (*cohesiveness*; Lott & Lott, 1965) influences whether the collaboration is continued (Summers et al., 1988).

What leads to the more positive robot-concepts? Exploratory results suggest that the changes are associated with increased perceived proficiency of the robot collaborator. In other words, how positive a

robot was perceived was tied to how helpful it was perceived. That humans form and use beliefs about a robot's capabilities is reasonable and has been shown before (e.g., Weis & Wiese, 2020). Since a human aid-giver is perceived more negatively when omitting help (Morse, 1972), it is plausible that a collaborator's perceived ability to help should factor in the emotional robot-concept. We thus argue that this exploratory finding strengthens the validity of our initial finding regarding the emotional robot concepts. Furthermore, in the emotional but not the autonomous condition, the change in the emotional robot-concept was additionally associated with indicators of emotional rather than performance-related processing. Participants who incorporated the robot more into their self perceived the robot more positively after solving the task together. Furthermore, participants who scored higher in the alexithymia measure were less likely to show such an increase. This exploratory finding suggests that participants did believe our framing, i.e. believed that the emotional robot was able to experience feelings, and thus suggests that such a framing is enough to trigger emotional interaction components known to be present in human-human interactions (e.g., self expansion; Aron & Aron, 1996).

4.3 Conclusion and outlook

The present study is the first of its kind to show that beliefs about and interactions with robots can change the emotional concepts of both oneself and a robot interaction partner. The study additionally provides insights into potential underpinnings—alexithymia, inclusion of the robot in the self, perceived competence of the robot—and consequences—mental well-being, cognitive performance—of such change. Future studies are necessary to validate these insights and increase the understanding of the underlying causal structure.

5 Highlights

- Emotional concepts of ourselves and *human* interaction partners are linked to performance, well-being, and interaction style
- Here, we investigated whether emotional concepts might be equally relevant when interacting with *robot* partners
- Results indicate that beliefs about and interactions with robots do change emotional self- and robot-concepts
- Additional analyses support the notion that these changes are linked to mental health and task performance
- We conclude that emotional consequences for the human should be considered when designing human-robot interactions

6 References

Aron, A., & Aron, E. N. (1996). Self and self-expansion in relationships. *Knowledge Structures in Close Relationships: A Social Psychological Approach*, 325–344.

Aron, A., Aron, E. N., Tudor, M., & Nelson, G. (1991). Close relationships as including other in the self. *Journal of Personality and Social Psychology*, *60*(2), 241–253. https://doi.org/10.1037/0022-3514.60.2.241

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23–32. https://doi.org/10.1016/0022-3999(94)90005-1

Bartholomew, K., & Horowitz, L. M. (1991). Attachment styles among young adults: A test of a four-category model. *Journal of Personality and Social Psychology*, *61*(2), 226–244. https://doi.org/10.1037/0022-3514.61.2.226

Bartneck, C. (2003). Interacting with an embodied emotional character. *Proceedings of the 2003 International Conference on Designing Pleasurable Products and Interfaces*, 55–60.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1995). Personal narratives about guilt: Role in action control and interpersonal relationships. *Basic and Applied Social Psychology*, *17*(1–2), 173–198.

Beck, A., Cañamero, L., & Bard, K. A. (2010). Towards an affect space for robots to display emotional body language. *19th International Symposium in Robot and Human Interactive Communication*, 464–469.

Bowlby, J. (1979). *The making and breaking of affectional bonds*. Travistock.

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1), 119–155. https://doi.org/10.1016/S1071-5819(03)00018-1

Breazeal, C., & Scassellati, B. (1999). How to Build Robots that Make Friends and Influence People. *Proceedings of the 1999 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS99)*, 858–863.

Clark, A. (1999). An embodied cognitive science? *Trends in Cognitive Sciences*, *3*(9), 345–351.

De Dreu, C. K., Beersma, B., Stroebe, K., & Euwema, M. C. (2006). Motivated information processing, strategic choice, and the quality of negotiated agreement. *Journal of Personality and Social Psychology*, *90*(6), 927.

Diener, E., & Diener, C. (1996). Most people are happy. *Psychological Science*, *7*(3), 181–185.

Dziergwa, M., Kaczmarek, M., Kaczmarek, P., Kędzierski, J., & Wadas-Szydłowska, K. (2018). Long-Term Cohabitation with a Social Robot: A Case Study of the Influence of Human Attachment Patterns. *International Journal of Social Robotics*, *10*(1), 163–176. https://doi.org/10.1007/s12369-017-0439-2

Epstein, S. (1973). The self-concept revisited: Or a theory of a theory. *American Psychologist*, *28*(5), 404–416. https://doi.org/10.1037/h0034679

Fisher, M., Goddu, M. K., & Keil, F. C. (2015). Searching for explanations: How the Internet inflates estimates of internal knowledge. *Journal of Experimental Psychology: General*, *144*(3), 674–687. https://doi.org/10.1037/xge0000070

Fox, J., & Weisberg, S. (2018). *An R companion to applied regression*. Sage publications.

Fu, W.-T. (2011). A Dynamic Context Model of Interactive Behavior: Cognitive Science. *Cognitive Science*, *35*(5), 874–904. https://doi.org/10.1111/j.1551-6709.2011.01173.x

Gilbert, S. J. (2015). Strategic use of reminders: Influence of both domain-general and task-specific metacognitive confidence, independent of objective memory ability. *Consciousness and Cognition*, *33*, 245–260. https://doi.org/10.1016/j.concog.2015.01.006

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Gray, W. D., Sims, C. R., Fu, W.-T., & Schoelles, M. J. (2006). The soft constraints hypothesis: A rational analysis approach to resource allocation for interactive behavior. *Psychological Review*, *113*(3), 461–482. https://doi.org/10.1037/0033-295X.113.3.461

Hamilton, K. A., & Yao, M. Z. (2018). Blurring boundaries: Effects of device features on metacognitive evaluations. *Computers in Human Behavior*, *89*, 213–220.

Haslam, N. (2006). Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology*, *60*(6), 895.

Hegel, F., Eyssel, F., & Wrede, B. (2010). The social robot 'flobi': Key concepts of industrial design. *19th International Symposium in Robot and Human Interactive Communication*, 107–112.

Herbert, C., Herbert, B. M., Ethofer, T., & Pauli, P. (2011). His or mine? The time course of self–other discrimination in emotion processing. *Social Neuroscience*, *6*(3), 277–288. https://doi.org/10.1080/17470919.2010.523543

Herbert, C., Herbert, B. M., & Pauli, P. (2011). Emotional self-reference: Brain structures involved in the processing of words describing one's own emotions. *Neuropsychologia*, *49*(10), 2947–2956. https://doi.org/10.1016/j.neuropsychologia.2011.06.026

Herbert, C., Pauli, P., & Herbert, B. M. (2011). Self-reference modulates the processing of emotional stimuli in the absence of explicit self-referential appraisal instructions. *Social Cognitive and Affective Neuroscience*, *6*(5), 653–661. https://doi.org/10.1093/scan/nsq082

Hutchins, E. (1995). *Cognition in the Wild*. MIT press.

Ifrah, G. (2001). *The universal history of computing: From the abacus to the quantum computer*. New York : John Wiley. http://archive.org/details/unset0000unse_w3q2

Kelly, J. R., & Barsade, S. G. (2001). Mood and Emotions in Small Groups and Work Teams. *Organizational Behavior and Human Decision Processes*, *86*(1), 99–130. https://doi.org/10.1006/obhd.2001.2974

Kirsh, D. (2013). Embodied cognition and the magical future of interaction design. *ACM Transactions on Computer-Human Interaction*, *20*(1), 1–30. https://doi.org/10.1145/2442106.2442109

Lott, A. J., & Lott, B. E. (1965). Group cohesiveness as interpersonal attraction: A review of relationships with antecedent and consequent variables. *Psychological Bulletin*, *64*(4), 259.

Löwe, B., Kroenke, K., & Gräfe, K. (2005). Detecting and monitoring depression with a two-item questionnaire (PHQ-2). *Journal of Psychosomatic Research*, *58*(2), 163–171.

Lumma, A.-L., Valk, S. L., Böckler, A., Vrtička, P., & Singer, T. (2018). Change in emotional self-concept following socio-cognitive training relates to structural plasticity of the prefrontal cortex. *Brain and Behavior*, *8*(4), e00940. https://doi.org/10.1002/brb3.940

Meixner, F., & Herbert, C. (2018). Whose emotion is it? Measuring self-other discrimination in romantic relationships during an emotional evaluation paradigm. *PLOS ONE*, *13*(9), e0204106. https://doi.org/10.1371/journal.pone.0204106

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, *130*(5), 711. https://doi.org/10.1037/0033-2909.130.5.711

Morse, S. J. (1972). Help, Likability, and Social Influence1. *Journal of Applied Social Psychology*, *2*(1), 34–46. https://doi.org/10.1111/j.1559-1816.1972.tb01262.x

Nash, K., Lea, J. M., Davies, T., & Yogeeswaran, K. (2018). The bionic blues: Robot rejection lowers self-esteem. *Computers in Human Behavior*, *78*, 59–63. https://doi.org/10.1016/j.chb.2017.09.018

Nosofsky, R. M. (1992). Exemplar-based approach to relating categorization, identification, and recognition. In F. G. Ashby (Ed.), *Scientific psychology series. Multidimensional models of perception and cognition* (pp. 363–393). Lawrence Erlbaum Associates, Inc.

O'Hara, K. P., & Payne, S. J. (1998). The Effects of Operator Implementation Cost on Planfulness of Problem Solving and Learning. *Cognitive Psychology*, *35*(1), 34–70. https://doi.org/10.1006/cogp.1997.0676

Pieschl, S. (2019). Will using the Internet to answer knowledge questions increase users' overestimation of their own ability or performance? *Media Psychology*, 1–27. https://doi.org/10.1080/15213269.2019.1668810

R Core Team. (2013). *R: A language and environment for statistical computing*.

Risko, E. F., & Gilbert, S. J. (2016). Cognitive Offloading. *Trends in Cognitive Sciences*, *20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Rosenthal-von der Pütten, A. M., Krämer, N. C., & Herrmann, J. (2018). The Effects of Humanlike and Robot-Specific Affective Nonverbal Behavior on Perception, Emotion, and Behavior. *International Journal of Social Robotics*, *10*(5), 569–582. https://doi.org/10.1007/s12369-018-0466-7

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*, *5*(1), 17–34. https://doi.org/10.1007/s12369-012-0173-8

Schroeder, D. A., Graziano, W. G., & Nadler, A. (2015). The Other Side of Helping. In D. A. Schroeder & W. G. Graziano (Eds.), *The Oxford Handbook of Prosocial Behavior*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195399813.013.004

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype Threat and Women's Math Performance. *Journal of Experimental Social Psychology*, *35*(1), 4–28. https://doi.org/10.1006/jesp.1998.1373

Storm, B. C., Stone, S. M., & Benjamin, A. S. (2017). Using the Internet to access information inflates future use of the Internet to access other information. *Memory*, *25*(6), 717–723. https://doi.org/10.1080/09658211.2016.1210171

Summers, I., Coffelt, T., & Horton, R. E. (1988). Work-Group Cohesion. *Psychological Reports*, *63*(2), 627–636. https://doi.org/10.2466/pr0.1988.63.2.627

Tangney, J. P., Wagner, P. E., Hill-Barlow, D., Marschall, D. E., & Gramzow, R. (1996). Relation of shame and guilt to constructive versus destructive responses to anger across the lifespan. *Journal of Personality and Social Psychology*, *70*(4), 797.

Taylor, S. E., & Brown, J. D. (1988). Illusion and well-being: A social psychological perspective on mental health. *Psychological Bulletin*, *103*(2), 193–210. https://doi.org/10.1037/0033-2909.103.2.193

Taylor, S. E., & Brown, J. D. (1994). Positive Illusions and Weil-Being Revisited: Separating Fact From Fiction. *Psychological Bulletin*, *116*(1), 21–27.

Tessler, R. C., & Schwartz, S. H. (1972). Help seeking, self-esteem, and achievement motivation: An attributional analysis. *Journal of Personality and Social Psychology*, *21*(3), 318–326. https://doi.org/10.1037/h0032321

Touron, D. R. (2015). Memory avoidance by older adults: When "old dogs" won't perform their "new tricks." *Current Directions in Psychological Science*, *24*(3), 170–176.

Turkle, S. (2012). *Alone together: Why we expect more from technology and less from each other*. Basic books.

Turkle, S., Taggart, W., Kidd, C. D., & Dasté, O. (2006). Relational artifacts with children and elders: The complexities of cybercompanionship. *Connection Science*, *18*(4), 347–361.

Van Kleef, G. A. (2009). How emotions regulate social life: The emotions as social information (EASI) model. *Current Directions in Psychological Science*, *18*(3), 184–188.

Võ, M. L., Conrad, M., Kuchinke, L., Urton, K., Hofmann, M. J., & Jacobs, A. M. (2009). The Berlin affective word list reloaded (BAWL-R). *Behavior Research Methods*, *41*(2), 534–538. https://doi.org/10.3758/BRM.41.2.534

Watson, L. A., Dritschel, B., Obonsawin, M. C., & Jentzsch, I. (2007). Seeing yourself in a positive light: Brain correlates of the self-positivity bias. *Brain Research*, *1152*, 106–110. https://doi.org/10.1016/j.brainres.2007.03.049

Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, *14*(8), 383–388.

Weis, P. P., & Herbert, C. (2017). Bodily Reactions to Emotional Words Referring to Own versus Other People's Emotions. *Frontiers in Psychology*, *8*(1277).

Weis, P. P., & Wiese, E. (2019). Problem Solvers Adjust Cognitive Offloading Based on Performance Goals. *Cognitive Science, 43*(e12802), 20. https://doi.org/10.1111/cogs.12802

Weis, P. P., & Wiese, E. (2020). Know Your Cognitive Environment! Mental Models as Crucial Determinant of Offloading Preferences: *Human Factors*. https://doi.org/10.1177/0018720820956861

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., & Hester, J. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686.

Wiese, E., Weis, P. P., Bigman, Y., Kapsaskis, K., & Gray, K. (2021). It's a match: Task assignment in human-robot collaboration depends on mind perception. *International Journal of Social Robotics*, https://doi.org/10.1007/s12369-021-00771-z.

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others. *PLoS ONE*, *7*(9), e45391. https://doi.org/10.1371/journal.pone.0045391

Winter, D., Herbert, C., Koplin, K., Schmahl, C., Bohus, M., & Lis, S. (2015). Negative Evaluation Bias for Positive Self-Referential Information in Borderline Personality Disorder. *PLOS ONE*, *10*(1), e0117083. https://doi.org/10.1371/journal.pone.0117083

7 Supplemental material

7.1 Words extracted from BAWL-R

Table S1

Word stimuli for the word evaluation task

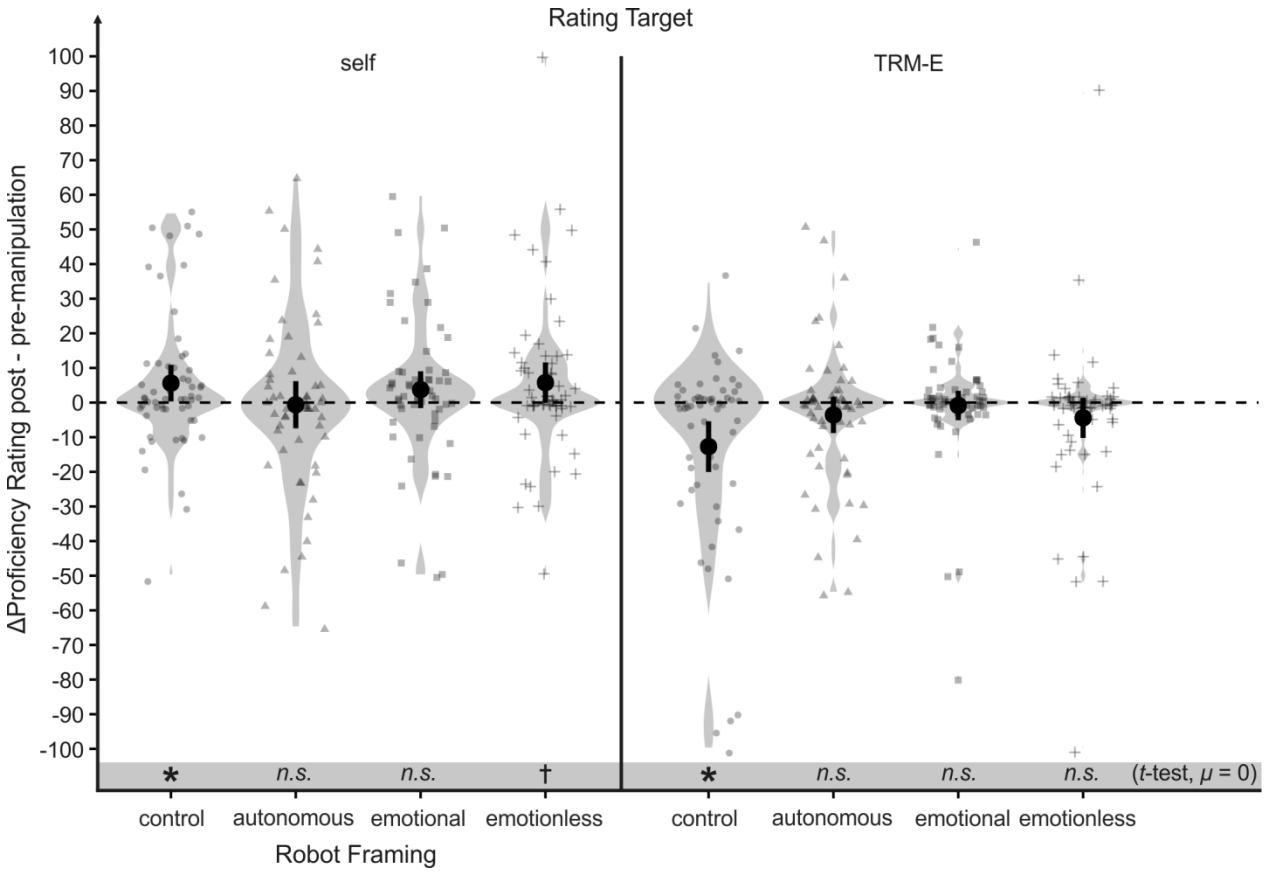
word (German)	word (English)	valence (German)	valence (English)	imageability (German)	imageability (English)
SIEG	VICTORY	2.20	1.97	4.44	6.07
ZUKUNFT	FUTURE	2.20	1.57	2.78	5.00
GEWINN	PROFIT	2.10	1.53	3.14	5.47
TRIUMPH	TRIUMPH	2.00	2.10	4.22	5.93
ENERGIE	ENERGY	2.10	1.67	3.44	5.43
AKTIVITÄT	ACTIVITY	1.71	1.40	4.12	5.37
ERFOLG	ACHIEVEMENT	2.10	2.00	3.05	5.37
WISSEN	KNOWLEDGE	2.03	2.13	2.62	5.43
BEDEUTUNG	SIGNIFICANCE	0.94	1.67	1.81	4.50
EXISTENZ	EXISTENCE	1.60	1.40	2.00	5.23
CHANCE	OPPORTUNITY	2.10	1.63	2.22	5.17
TREFFER	SUCCESS	1.90	2.20	5.00	5.53
RÜCKKEHR	RETURN	1.35	0.63	4.35	4.30
LEISTUNG	PERFORMANCE	1.80	1.47	2.33	5.23
FÄHIGKEIT	ABILITY	1.62	1.60	2.12	5.03
KRAFT	POWER	1.38	1.30	3.85	5.33
STÄRKE	STRENGTH	1.79	1.97	3.85	5.93
ERGEBNIS	SCORE	0.74	1.33	3.08	5.43
BEGEGNUNG	ENCOUNTER	1.56	0.83	4.85	5.50
KONTAKT	CONTACT	1.40	1.13	3.44	5.80
TÄTIGKEIT	AGENCY	0.74	0.80	2.92	4.83
VORTEIL	ADVANTAGE	1.80	1.57	1.89	4.80
ANTWORT	RESPONSE	0.94	1.33	2.69	4.80
MISSION	MISSION	0.90	1.40	2.78	5.23
WERT	VALUE	1.06	1.70	2.73	5.00
LÖSUNG	SOLUTION	1.53	1.87	2.73	5.53
TEILNAHME	PARTICIPATION	0.82	1.40	2.96	5.37
LOGIK	LOGIC	1.15	1.70	2.59	4.90
QUALITÄT	QUALITY	1.65	1.70	2.42	5.40
AUSDAUER	ENDURANCE	1.60	1.63	3.22	5.23
SCHUTZ	PROTECTION	1.70	1.63	3.67	5.20
PAUSE	BREAK	1.15	0.20	3.00	5.70
		1.55	1.51	3.13	5.28

Note. Ratings are for the German words only and extracted from BAWL-R (Võ et al., 2009). Valence was rated on a scale from -3 to 3, imageability on a scale from 1 to 7. For details on the word selection procedure, see section *Word evaluation task* of the main manuscript.

7.2 Proficiency ratings

Figure S1

Proficiency Rating Changes From Pre- to Post-Manipulation



*Note. Error bars indicate 95% CI. *: $p < .05$, †: $p = .06$, n.s.: $p > .14$*