International Journal of Social Robotics (2018) 10:519–535
https://doi.org/10.1007/s12369-017-0455-2

Here Comes the Bad News: Doctor Robot Taking Over

Johan F. Hoorn[1] · **Sonja D. Winter**[2]

Accepted: 6 December 2017 / Published online: 13 December 2017

Abstract

To test in how far the Media Equation and Computers Are Social Actors (CASA) validly explain user responses to social robots, we manipulated how a bad health message was framed and the language that was used. In the wake of Experiment 2 of Burgers et al. (Patient Educ Couns 89(2):267–273, 2012. https://doi.org/10.1016/j.pec.2012.08.008), a human versus robot doctor delivered health messages framed positively or negatively, using affirmations or negations. In using frequentist (robots are different from humans) and Bayesian (robots are the same) analyses, we found that participants liked the robot doctor and the robot's message better than the human's. The robot also compelled more compliance to the medical treatment. For the level of expected quality of life, the human and robot doctor tied. The robot was not seen as affectively distant but rather involving, ethical, skilled, and people wanted to consult her again. Note that doctor robot was not a seriously looking physician but a little girl with the voice of a young woman. We conclude that both Media Equation and CASA need to be altered when it comes to robot communication. We argue that if certain negative qualities are filtered out (e.g., strong emotion expression), credibility will increase, which lowers affective distance to the messenger. Robots sometimes outperform humans on emotional tasks, which may relieve physicians from a most demanding duty of disclosing unfavorable information to a patient.

Keywords Media Equation · CASA · Framing · Language · Communication · Healthcare

1 Introduction

In their Handbook of Research on Computer Mediated Communication, Kelsey and St. Amant [19] state that to regard robots as social, they need to express emotions, show personality, work with natural cues, and be capable of conducting high-level dialogue (p. 867). Their handbook also states that most robots are still underdeveloped in this respect (p. 865). Part of this omission is that we do not understand too well how, for instance, high-level dialogues take place between humans. To advance a little bit into this direction, the current study reports on the delivery of bad health news by a social robot, using framing techniques and language biases that are recommended to human messengers.

Delivering bad news through robots is the test case we used for the deeper aim of this study: A demonstration of testing the reach of Media Equation [35] and Computers Are Social Actors (CASA [32]) fairly by using Bayes as well as frequentist statistics. From the results, our contribution to the field emerged: A moderation of the said theories that people to some extent apply human social criteria to machines but that robots sometimes perform better according to those criteria than humans actually do and that some criteria do not count for robots. This is unexpected for many theories in media and technology, which simply assume that humans are the standard of performance. In application, this study opens up the serious possibility to apply robots in executing emotionally sensitive tasks, both relieving the messenger and the receiver from relational stress.

To progress towards such study, several hurdles have to be taken: in theory, in methods, and practically. Theoretically, we conceive of a social robot as a humanoid software and hardware system that at least in part can make autonomous decisions and expresses social behaviors by

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12369-017-0455-2) contains supplementary material, which is available to authorized users.

✉ Johan F. Hoorn
j.f.hoorn@vu.nl

[1] Department of Communication Science, Vrije Universiteit Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

[2] Psychological Sciences, University of California, Merced, Merced, CA, USA

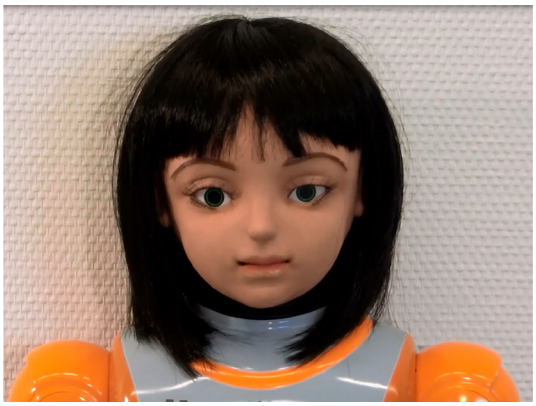

Fig. 1 Hanson Robokind Alice R50 bringing bad health news

electro-mechanical means. In our case, we used the Robokind R50 "Alice," designed by David Hanson and equipped it with our own software (Fig. 1). As a starting point, we relied on theorists such as Reeves and Nass [35], who formulated the Media Equation, suggesting that humans hardly distinguish between other humans and machines in the way they behave toward them. Reeves and Nass report that, for example, applications that stimulate extroversion are more appreciated by extroverts, whereas introvert people would prefer applications that demand more introversion. With regard to robots, certain studies indeed find that similar personalities attract (e.g., [42]). However, others find that complementarity attracts (e.g., [23]) but this may still be in line with Media Equation if humans respond similarly to humans who complement them. In the same vein did Nass and Moon [32] assert that Computers Are Social Actors, the CASA view. While interacting with computers, users would automatically or 'mindlessly' apply human-inspired heuristics and rules because certain anthropomorphic aspects of the machine remind them of human conduct—even if they are aware of the machine (cf. the emotional expressive systems of [10]). And indeed, Lee et al. [26], for instance, studied a robot moving around an office space while it dispensed snacks to the employees. Over time, the robot became more of a co-worker than a vending machine in the eyes of the employees, inspiring all kinds of social dynamics such as becoming more polite to the robot, protecting it, making social comparisons, and becoming jealous of people that supposedly were preferred by the machine.

We may call the Media Equation and CASA view the *similarity hypothesis*, indicating that people treat robots more-or-less the same as human beings with regard to the application of social rules and overlearned social responses (also in cases of complementarity). Note that Media Equation and CASA do not mean that humans interacting with media behave in the exact same way as they do towards other humans but they do so roughly. Although the theory says 'Equation,' suggesting that behaviors are identical, we better stick to similarity, leaving some room for variance. Nevertheless, if Media Equation and CASA want to refrain from vacuity, results of a comparison between human–human and human–robot interactions should be 'about the same' and not 'different.'

There are also authors who believe that humans and robots are not treated equally (e.g., [11,13]). For example, Takayama and Go [41] observed that certain people categorized mobile telepresence machines as indeed another person, arguing with it and making hand gestures (cf. CASA). Yet, other people saw it as a disabled person because it could not open the door or failed otherwise. A third group saw the telepresence machine as robot, a substitute of the actual person, and the fourth group merely saw an object, referring to it as 'it' and 'thing,' sitting on it like a chair or leaning against it like a railing (ibid.).

Such studies underscore that not all perceivers transfer the human model to robots without more (some see the machine-side rather than the human likeness) or at least incompletely (robot is a substitute for the real thing) or qualified by different scores (robot is clumsier). For instance, Küster and Świderska [21] reported evidence that humans do attribute a certain amount of morality to humanoid robots but also that systematic differences with humans occurred. For instance, humans were regarded as more experienced and more conscious than robots (p. 341). According to this kind of *dissimilarity hypothesis*, robots better should be seen as novel social entities with non-human but highly appreciated qualities of their own [14].

Where do we stand then? Do we treat social robots as if they were real people (similarity hypothesis)? There may be some margin assumed but not so large that the difference between treating robots and treating human beings becomes statistically significant. If similar, then all of the rules and subtleties of human communication apply and people supposedly are sensitive to whether the robot complies with those rules or not. Or are our human responses modified by the fact that it is a robot (dissimilarity hypothesis)? In that case, human behavioral rules would roughly apply but not systematically and unequal from responding to a human partner.

So far so good for theory because the comparison between human–human and human–robot interactions has its challenges in methods and analysis as well. From a methods point of view, ideally one would like to compare a physically present human (the 'confederate') with a physically present robot, interacting with a test participant. However, human confederates are hardly capable of delivering the same performance time and again across a sample of participants; let alone if that person has to act in two or more different ways (e.g., acting happy vs. sad). Woods et al. [48,49] recommend the use of video clips in cases where interaction is low (such as unidirectionally delivering a message). Therefore, human–human studies often rely on videotaped materials that can be

repeated multiple times without changing the contents. That may sound as yielding to methods at the cost of ecological validity but if we see the movie clips as a proxy of a video-conference (cf. telemedicine), we may not deviate too far from reality, particularly in view of future developments.

Given that humans should be videotaped to deliver the same performance, to keep things comparable, robots also should be videotaped. That of course, does something for the aspect of interaction (as with humans) and for autonomous decision making (as with humans). Nonetheless, we should start somewhere and explore the 'simpler' situations first before entering the methodological complications of real-life interactions. And admittedly, using movie clips is way more practical and cost-efficient than hiring professional actors and far less work than handling a robot, which commonly shows all kinds of technical flaws that need to be repaired on the fly [48,49]. Fortunately, there are situations where little interaction and decision making is required, namely in one-way monologues where someone tells the other about a state-of-affairs, such as notifying employees of their layoff or breaking bad news to a patient, which is the topic of the current study.

With reference to statistics, the human–human versus human–robot comparison also differs from what is common practice. In conventional (frequentist) statistical techniques (e.g., *t* test, ANOVA), data are always compared to a null hypothesis (H0). This works fine when one has no idea of what could be going on in the data. However, this is not often the case. Researchers have theories, expectations, and previous studies to build their case (prior knowledge). Bayesian statistics allows prior knowledge (the prior) to support the estimation of a model and to test hypotheses. Using this method, one can build on earlier research, instead of starting from scratch every time anew. Moreover, frequentist statistics assume that but one true population parameter (fixed) exists in the population.

Rejecting the null actually means that in the data significant differences were found between two groups, confirming the H1. In other words, frequentist analyses favor in the case of Media Equation and CASA, the dissimilarity hypothesis (H1: robots and humans are perceived as significantly different). Because if H1 is rejected and the H0 remains, it merely says 'we found no difference' rather than 'we found similarity.' Hence, to be fair to the similarity hypothesis as well, analyses should also be performed with Bayesian statistics. With Bayes, all unknown parameters can be defined by a probability distribution. Thus, Bayesian statistics do not result in a point estimate, but rather in an interval with a certain probability that the true coefficient is part of that interval [12,44]. Bayesian analysis exists of three parts: the prior distribution, the data (the likelihood), and the posterior distribution. The posterior distribution is a combination of the prior distribution and the data, an updated understanding of the theory under question (ibid.). In other words, with Bayes the H0 *can* be tested and if H1 (dissimilarity) is refuted and H0 remains, we can assume that robots are treated similarly to humans as Media Equation and CASA would have it.

1.1 Related Work

Media Equation and CASA are directed at media and computers. However, other artifacts are humanized and attributed agency just as well [16,22]. This also holds for household appliances such as vacuums [40]. With agency assumed, people hold computers for a messenger [39], employing different communication strategies [2–4]. However, not every person responds in the same way [25]. Therefore, tailored approaches attempted to relate the robot to the user's personality [42]. In application, the robot's performance in real-life situations was studied by Severinson-Eklundh et al. [37] for services, by Mutlu and Forlizzi [31] for organizations, by Kanda et al. [18] for tutoring, and by Sabelli et al. [36] for eldercare.

As the latter study illustrates, an area where social robots are advancing quickly is health care. To alleviate the care burden, robots are employed to remind people of medicine intake, to help exercising, and they are used as companions (e.g., [27]).

With respect to health communication, one of the toughest duties of a doctor is to break bad news about health to a patient. On the Internet and in medical journal articles alike, tips, tricks, and training protocols are shared on how to communicate with patients in a compassionate manner (e.g., [1,29]). To relieve the doctor from a challenging task, could a bad news message be brought by a robot equally well (the similarity hypothesis)? Or can we not and shall we not leave such a sensitive matter to robots, incapable of real understanding of the user's emotional state (the dissimilarity hypothesis)? After all, bad news may excite anxiety reactions and stress and if delivered wrongly, it can even aggravate the illness [38]. So how to deliver a bad health message properly?

Burgers et al. [6] looked into the effects of framing a health message positively or negatively, while using affirmative or negative language. These authors worked from Prospect Theory [17], measuring how people responded differently to four movie clips of a doctor communicating positive versus negative outcomes of a health diagnosis (the frame) phrased in affirmative language versus the use of negations. Participants liked it best when the videotaped doctor said that things "go well" (affirmative wording, positive outcome) but they disliked most a positive outcome that was brought with negations ("The news is not bad"). If things went badly, they did not want to get it straight into their face ("It's going bad") but rather preferred the denial of the positive outcome ("It's not going well"). Burgers et al. [6] found that if done in the wrong way, people felt negative about the doctor, about the

health message, expected a lower quality of life, and did not feel like following up on the doctor's advice. Their Experiment 2 will be the entrance point for our robot doctor to communicate bad news.

1.2 Research Question and Hypotheses

Question is of course, will the effects of health-message framing by a human-doctor also be established by a robot-doctor (the similarity hypothesis) or will it be different (the dissimilarity hypothesis)? And if different, in what way? Worse? Better? Or what?

At face value, one expects that a patient would rather talk to a human doctor than a machine. In earlier work, however, communication with robots sometimes is preferred over human communication, even or sometimes precisely because emotional concerns are at stake (e.g., [14,50]). One could argue then that because the doctor is human, patients expect the highest accomplishment; better than a robot. If humans do not live up to expectations, the human is deeply disappointing; more than a robot. Because we expect less of robots, they conceivably occupy the middle ground: not as good as well-performing human doctors, not as bad a poor-performing humans (cf. [46]).

Therefore, we wanted to repeat Experiment 2 of Burgers et al. [6] and replace the movie clips of a human doctor by clips of a Robokind R50 Alice, bringing bad news about Bekhterev's rheumatic disease and compare our data with the original human-doctor data of [6]. The *similarity* hypothesis consisted of the constellation of findings in [6]. Because no substantial differences were to be expected between human and robot, we predicted the H0 for each effect that included the type of doctor (human vs. robot). Hence, we anticipated a significant interaction between frame and language for both doctors: For news that is framed as fortunate but phrased in a negative manner (a positive frame with negations: "Things are not too bad"), we thought that participants would rate the health message, the doctor, the quality of life they expected, and the intention to adhere to the medical advice lower than when affirmations were used (positive frame, affirmative language: "Things go well"). Additionally, the generally lower scores to bad news would be mitigated when negations were used (negation of a negative frame: "Not good") as compared to the affirmation of negative frames ("It's going bad").

For the *dissimilarity* hypothesis (H1), we followed [46] in predicting that when a human does it right (affirmation of a positive frame and negation of a negative frame), the human will outperform the robot at the message-related outcome variables. Yet, when the human doctor does it wrong (negation of a positive frame and affirmation of a negative frame), the robot outdoes the human doctor because the robot is 'forgiven' for being communicatively and emotionally not so proficient. In other words, the dissimilarity hypothesis was fine-tuned by expecting that it was *better to have a good robot than a bad human* but that a good human would supersede everyone.

Apart from the message-related outcome variables (i.e. message evaluation, doctor evaluation, quality of life, and medical adherence), we additionally measured a number of experiential variables not sampled in the Burgers study [6]. These measures were taken from an empirically well-validated model of perceiving and experiencing fictional characters [45] such as movie characters, media figures, game characters, and robots. In other words, we also explored an extra research question (RQ1) on how the robot doctor was experienced in terms of ethics (a good or a bad robot), its affordances (is the robot skilled or not), feeling involved with the robot (friendly feelings towards the robot) or feeling at a distance (cold feelings), and use intentions (to what extent people wanted to consult the robot doctor again).

In sum, if robots follow the same pattern (H0) as for human doctors (cf. Media Equation), they should not say, for example: "Things are not going bad." But rooted in the assumption that robots are 'cold,' people may dislike a robot delivering the bad news ("Inhumane"). In other words, the pattern may be the same as for humans but on a lower level: more negative about the doctor, about the health message, about the expected quality of life, and less inclined to follow the robot's advice. Moderating this, however (H1), could be the effect that a senseless machine that tries its best is forgiven. To test these hypotheses (H0, H1) and to explore our extra research question (RQ1), we received the data from the human doctor in Experiment 2 in Burgers et al. [6] directly from the authors and collected additional data, using a robot doctor.

2 Method

2.1 Participants and Design

Participants ($N = 134$; $M_{age} = 28.02$, $SD_{age} = 11.98$, 61.2% female, Dutch nationality) were randomly assigned to a 2 (Language: affirmation vs. negation) $\times 2$ (Frame: positive vs. negative) between-subjects experiment, receiving course credits or a small monetary compensation. Seventy percent of the participants had higher vocational or university training; sixteen people were familiar with Bekhterev's disease, 75 people had experience with doctor-patient conversations. For a comparison with [6] (also all Dutch citizens), see Table 1. The Analysis section shows that no significant differences were found in background information between condition cells. Hence, we may assume that the two sets of data were collected from the same population with comparable individual differences.

Table 1 Participant characteristics

Doctor	Human doctor ($N = 115$)												Robot doctor ($N = 134$)											
Language	Negation ($n = 59$)						Affirmation ($n = 56$)						Negation ($n = 68$)						Affirmation ($n = 66$)					
Frame	Negative ($n = 32$)			Positive ($n = 27$)			Negative ($n = 26$)			Positive ($n = 30$)			Negative ($n = 33$)			Positive ($n = 35$)			Negative ($n = 31$)			Positive ($n = 35$)		
	n	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*
Familiar Bekhterev?																								
Yes	3			5			6			4			5			2			5			4		
No	29			22			20			26			28			33			26			31		
Experience with doctor/patient conv.																								
Yes	21			18			19			24			21			17			17			20		
No	11			9			7			6			12			18			14			15		
Sex																								
Male	9			8			7			9			13			14			13			12		
Female	23			19			19			21			20			21			18			23		
Age		28.09	13.74		26.74	13.37		27.00	11.72		26.87	11.78		28.52	13.37		29.60	13.63		28.26	11.27		25.86	9.81
Doctor Evaluation		3.65	1.12		3.27	1.19		3.31	1.18		4.03	1.37		4.32	1.44		4.25	1.23		4.06	1.34		3.82	1.37
Message Evaluation		4.00	1.66		2.56	1.51		4.05	1.28		3.60	1.48		4.81	1.51		5.05	1.40		5.25	1.28		4.87	1.31
Expected Quality of Life		4.98	1.28		3.63	1.40		4.42	1.24		3.37	1.14		5.06	1.49		4.61	1.63		3.18	1.41		3.16	1.16
Medical Adherence		5.50	1.34		4.98	1.40		5.31	1.52		5.32	1.53		5.86	1.1		5.73	1.07		5.39	1.21		5.46	1.2
Ethics														5.13	0.87		4.94	0.85		4.94	1.03		4.59	1.08
Involvement														2.97	0.86		2.90	0.84		2.62	0.93		2.72	0.92
Distance														3.97	1.37		4.19	1.14		4.68	1.01		4.44	1.35

	Robot doctor ($N = 134$)							
Language	Negation ($n = 35$)				Affirmation ($n = 33$)			
Frame	Negative ($n = 16$)		Positive ($n = 19$)		Negative ($n = 16$)		Positive ($n = 17$)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Affordances	3.68	0.79	3.48	0.70	3.36	0.81	3.30	0.86
Use Intentions	3.26	0.35	2.99	0.42	3.16	0.47	3.39	0.46

The message-related outcome variables were taken from [6]. The experiential variables were taken from [45]. All values represent raw, non-standardized scores

2.2 Procedure

Participants were invited through Facebook to open a link to a Qualtrics environment, used for administering surveys and experiments. In [6], people watched one out of 4 movie clips of a human doctor bringing health news in Dutch (Online Resource 1). Mimicking [6] as much as possible, the current participants watched a YouTube clip (1.27 m) that showed a robot doctor facing the camera and talking to the participant the same way, according to the same scripts, and using the same language as the human doctor in [6], Experiment 2. Qualtrics randomly distributed participants over four clips that represented the four experimental conditions (i.e. affirmative-positive, affirmative-negative, negation-positive, negation-negative) (Online Resource 2). The robot was a Hanson Robokind R50 "Alice" with a human-like girlish face and mechanical bodywork, which was visible in half total (Fig. 1), the same way as in [6]. We chose this machine because of its expressive face and the good results we booked in previous studies [14]. There was also a practical reason: This was at the time the only machine we could work with. Texts were between 167 and 173 words long and were run through an ACAPELA speech engine, speaking with a young adult female voice. In the first part of the text, the robot introduced the topic, in the second part it did the diagnosis, the third part gave a prognosis and the doctor's advice on medicine intake. In the fourth part, the robot doctor offered the option to consult her again. The stimulus text with variations in framing (bringing the message positively or not) and language (using negations or not) go next (translated from the Dutch):

> Good morning, please sit down. I will not beat around the bush: The news I have to bring is **(not) good / (not) bad.** As you know, we have done a lot of tests in the past week to diagnose your complaints. We made an X-ray and took a Schober test. From all these tests the same results were obtained. You have been tested positively for Bechterew's disease. In view of the circumstances, I think these results are **(not) good / (not) bad.** I understand you are full of questions right now. For now it is important to know that Bekhterev's disease is a genetic disease. Most patients **(do not)** find it **easy / hard** to live with this disease. I will prescribe you specific medication. With this medication your quality of life will probably **(not) progress / (not) deteriorate** in the coming weeks. I recommend reading this information leaflet about Bekhterev's disease. In addition, I would like to make a new appointment for about two weeks to evaluate the treatment.

This structure was the same as in [6]. After watching the clip, participants completed a structured questionnaire that was presented in blocks with pseudo-random sequences of items within blocks.

2.3 Measures

Measurement scales were composed of indicative and counter-indicative Likert-type items, rated on a 7-point scale (1 = totally disagree, 7 = totally agree). The message-related outcome variables were identical to [6], Experiment 2. Sample items from the scales were (translated from the Dutch) "The doctor was tactful" for *Doctor Evaluation*, "The doctor's message was clear" for *Message Evaluation*, "I think the medication will work out well" for *Expected Quality of Life*, and "It is a good idea to follow the treatment advice" for *Medical Adherence.* The experiential variables were drawn from [45]. Sample items from the scales were "This robot-doctor is fair" for *Ethics*, "This robot-doctor raises warm feelings" for *Involvement*, "… cold feelings" for *Distance*, "This robot-doctor is incompetent" for *Affordances*, and "I would like to visit this robot-doctor again" for *Use Intentions*.

2.3.1 Message-Related Outcome Variables

To assess the factor structure of the four outcome variables *Doctor Evaluation*, *Message Evaluation*, *Expected Quality of Life*, and *Medical Adherence*, we executed an Exploratory Factor Analysis (EFA) (Maximum likelihood estimation) with Promax rotation, expecting 4 factors. Model fit was good ($\chi^2 = 85.75$, $p < .001$, RMSEA = .066). All items neatly loaded onto their own scale, with the exception of the recoded *Message Evaluation* item "Took my hope away" and the first *Expected Quality of Life* item "On the basis of this conversation I think the quality of my life will be high." In Winter and Hoorn [47], further scale analysis (Online Resource 3) indicated that we should construct an *Expected Quality of Life* scale that was different from [6], because that study seemed to have a flaw in the analysis (Online Resource 3).

This left us with the following scales: *Message Evaluation* (3 items, Cronbach's alpha = .89), *Doctor Evaluation* (5 items, Cronbach's alpha = .88), *Expected Quality of Life* (2 items, one originally from *Message Evaluation*, Cronbach's alpha = .65), *Medical Adherence* (2 items; Cronbach's alpha = .84).

2.3.2 Experiential Variables

The experiential variables in the Robot Doctor study (N = 134) were *Ethics*, *Involvement*, and *Distance*. *Ethics* pertained to the doctor being morally good, just, and the like, *Involvement* pertained to becoming friends with the robot, and *Distance* was about having cold feelings for the robot doctor and about feeling aloof.

To assess the factor structure of *Ethics*, *Involvement*, and *Distance*, we executed an EFA (Maximum likelihood estimation) with Promax rotation, expecting 3 factors. Model fit was good ($\chi^2 = 102.15$, $p = .001$, RMSEA = .068). The general factor structure looked good, with some exceptions. First, "This robot doctor is kind" and "This robot doctor is of good will" had a low loading on their own *Ethics* factor, but a high loading on the *Involvement* factor. Looking at the content of these items, it is not unexpected that they clang to the *Involvement* items (e.g., "I have a good feeling about this doctor"), thus we added them to the *Involvement* scale. This way, we ended up with: *Ethics* (5 items, Cronbach's alpha = .89), *Involvement* (6 items, two originally from *Ethics*, Cronbach's alpha = .82), and *Distance* (4 items, Cronbach's alpha = .83).

Due to a mishap in the online survey, the experiential variables *Affordances* and *Use Intentions* were measured for a mere sub sample of participants in the Robot Doctor study ($n = 68$). *Affordances* related to what the user thinks the robot is capable of doing (e.g., being skillful, knowledgeable) and *Use Intentions* pointed at the willingness of the user to engage with the robot again in the future (e.g., a new consultation). An EFA assessed the factor structure of these two variables (Maximum likelihood estimation) with Promax rotation, expecting 2 factors. Whereas some of the items clearly loaded onto their respective factors, other showed cross-loadings, whereas yet other loaded on a factor they theoretically did not belong to (Online Resource 3). Particularly, three *Affordances* items had a low loading on their own factor (< .30), and a higher loading on the *Use Intentions* factor. These three items all had to do with how inadequate the doctor was, and we excluded them from the scale. This resulted into the following scales: *Use Intentions* (6 items, Cronbach's alpha = .93), and *Affordances* (6 items, Cronbach's alpha = .95). For further details, see [47] (Online Resource 3).

3 Results

3.1 Message-Related Outcome Variables

3.1.1 Preliminary Analyses

To control for differences in background information between condition cells, we combined our robot data ($N = 134$) with the data set of [6] for human doctors ($N = 115$ in their Experiment 2). An ANOVA with age as the dependent variable showed no main or interaction effects of the condition variables (Doctor, Frame, Language). We also created a variable that combined the three condition variables and explored the differences with Chi-square, showing no significant dependencies between conditions, neither with familiarity with Bekhterev's disease ($\chi^2 = 5.24$, $p = .630$, $\varphi = .145$), nor experience with doctor-patient conversations ($\chi^2 = 9.64$, $p = .210$, $\varphi = .197$), or gender ($\chi^2 = 3.37$, $p = .849$, $\varphi = .116$). Therefore, we dismissed the background variables from further analyses. Table 1 provides the descriptive statistics.

3.1.2 MANOVA

In testing our hypotheses for the entire data set, we ran a 2 (Doctor: Human vs. Robot) × 2 (Language: Negation vs. Affirmation) × 2 (Frame: Negative vs. Positive) between-subjects MANOVA. *Message Evaluation*, *Doctor Evaluation*, *Medication Adherence*, and *Expected Quality of Life* were the dependent variables.

3.1.3 Multivariate Results

The multivariate results showed that all main effects were significant (Doctor: $\lambda = .79$, $F_{(4,238)} = 15.74$, $p < .001$, $\eta_p^2 = .21$; Language: $\lambda = .85$, $F_{(4,238)} = 10.24$, $p < .001$, $\eta_p^2 = .15$; Frame: $\lambda = .89$, $F_{(4,238)} = 7.33$, $p < .001$, $\eta_p^2 = .11$). Two two-way interaction effects were significant as well (Doctor * Language: $\lambda = .94$, $F_{(4,238)} = 3.78$, $p = .005$, $\eta_p^2 = .06$; Doctor * Framing: $\lambda = .92$, $F_{(4,238)} = 5.33$, $p < .001$, $\eta_p^2 = .08$).

3.1.4 Univariate Results

Doctor had a univariate effect on *Doctor Evaluation* ($F_{(1,241)} = 11.18$, $p < .001$, $\eta_p^2 = .04$), *Message Evaluation* ($F_{(1,241)} = 61.85$, $p < .001$, $\eta_p^2 = .20$), and *Medical Adherence* ($F_{(1,241)} = 4.09$, $p = .044$, $\eta_p^2 = .02$) but not on *Expected Quality of Life* ($F < 1$). The Robot Doctor scored higher than the Human Doctor: *Doctor Evaluation* was higher for the Robot Doctor ($M = 4.11$, $SE = .11$, 95% CI 3.89–4.33) than for the Human Doctor ($M = 3.57$, $SE = .12$, 95% CI 3.32–4.80). *Message Evaluation* was higher for the Robot Doctor ($M = 4.99$, $SE = .12$, 95% CI 4.75–5.24) than for the Human Doctor ($M = 3.55$, $SE = .14$, 95% CI 3.29–3.82). *Medical Adherence* also was higher for the Robot Doctor ($M = 5.61$, $SE = .11$, 95% CI 5.39–5.83) than the Human Doctor ($M = 5.28$, $SE = .12$, 95% CI 5.04–5.51).

Language had a univariate effect on *Expected Quality of Life* ($F_{(1,241)} = 36.16$, $p < .001$, $\eta_p^2 = .13$): Negation ($M = 4.57$, $SE = .12$, 95% CI 4.33–4.81) yielded higher scores than Affirmation ($M = 3.53$, $SE = .12$, 95% CI 3.29–3.78).

Framing had an effect on *Expected Quality of Life* ($F_{(1,241)} = 17.27$, $p < .001$, $\eta_p^2 = .07$) and *Message Evaluation* ($F_{(1,241)} = 7.73$, $p = .006$, $\eta_p^2 = .03$), both showing

higher scores for Negative Frames. *Expected Quality of Life* was higher in a Negative Frame ($M = 4.41$, $SE = .12$, 95% CI 4.17–4.66) than in a Positive Frame ($M = 3.69$, $SE = .12$, 95% CI 3.45–3.93).[1] *Message Evaluation* also was higher in a Negative Frame ($M = 4.53$, $SE = .13$, 95% CI 4.27–4.78) than in a Positive Frame ($M = 4.02$, $SE = .13$, 95% CI 3.77–4.27).

The interaction between Doctor and Language had a significant effect on *Expected Quality of Life* ($F_{(1,241)} = 13.20$, $p < .001$, $\eta_p^2 = .05$). Pairwise comparisons showed that there was no effect of Language for the Human Doctor ($\Delta M = .41$, $SE = .25$, $p = .106$, 95% CI -0.09 to 0.92), whereas there was an effect of Language for the Robot Doctor ($\Delta M = 1.67$, $SE = .24$, $p < .001$, 95% CI 1.21–2.13). When the Robot Doctor used Negation, *Expected Quality of Life* was on average higher than when the Robot Doctor used Affirmation. Yet, no significant effect of Language on *Expected Quality of Life* was found for the Human Doctor; the means hardly differed ($\Delta M = .41$).

The interaction between Doctor and Frame was significant for *Expected Quality of Life* ($F_{(1,241)} = 7.88$, $p = .005$, $\eta_p^2 = .03$) and *Message Evaluation* ($F_{(1,241)} = 5.73$, $p = .017$, $\eta_p^2 = .02$). Pairwise comparisons for *Expected Quality of Life* showed that there was a significant effect of Frame for the Human Doctor ($\Delta M = 1.21$, $SE = .25$, $p < .001$, 95% CI 0.71–1.71), but not for the Robot Doctor ($\Delta M = .23$, $SE = .24$, $p = .322$, 95% CI -0.23 to 0.70). When the Human Doctor used Negative Frames, *Expected Quality of Life* was higher than when the Human Doctor used Positive Frames. For the Robot Doctor, the differences were marginal ($\Delta M = .23$).

The interaction effect of Doctor and Frame on *Message Evaluation* showed a similar pattern. Pairwise comparisons showed that there was a significant effect of Frame for the Human Doctor ($\Delta M = .95$, $SE = .27$, $p = .001$, 95% CI 0.42–1.48), but not for the Robot Doctor ($\Delta M = .07$, $SE = .25$, $p = .777$, 95% CI -0.42 to 0.56). When the Human Doctor used Negative Frames, *Message Evaluation* was higher than when he used Positive Frames. For the Robot Doctor, this difference was near absent ($\Delta M = .07$). As said, however, the *main* effect of Doctor on *Message Evaluation* was higher for the Robot Doctor across all Framing conditions.

3.2 Bayes Factors

Whereas the above frequentist analyses tested for differences between human and robot (H1), the Bayes analyses presented next will test for equality (H0). With Bayes, we looked for the amount of evidence in the data for the human and robot doctor to be perceived 'as similar,' while they used the same communication frames and language.

To compare the various models (not effects) of equality, we calculated Bayes Factors (BF), which is a non-binary, continuous statistical index that quantifies the evidence for a hypothesis in comparison to the alternative hypothesis. Hence, there are no cut-off points (i.e. the frequentist .05 rejection area), which makes its interpretation context-dependent and not a mere yes or no. For the meaning and interpretation of BF, we follow Lee and Wagenmakers's [24, p. 122] reading of Jeffreys [15] that $3 < BF_{10} < 30$ is considered moderate to strong evidence for H1 and $1/30 < BF_{10} < 1/3$ is moderate to strong evidence for H0.

We calculated BF in JASP (Version 0.6 [28]) and BayesFactor (Version 0.9.10-2 [30]). Note that JASP does not return a Bayes Factor for each effect in a full model (a model with all main and interaction effects) but returns a Bayes Factor for each constituent model, building from the Null (no predictors) to a model with all main effects and interactions included. For each step, JASP compares the current model to the original Null and computes a Bayes Factor based on the difference in model fit of the two models. JASP can produce two types of Bayes Factors, one that quantifies evidence in favor of the Null model as compared to the Alternative (BF_{01}), and another that quantifies the opposite evidence in favor of the Alternative model as compared to the Null model (BF_{10}).

3.2.1 Doctor Evaluation

Table 2 shows partial output of the Bayesian ANOVA for both types of Bayes Factors. First, all main effects were tabulated individually, after which they were incorporated into one model. Then all 2-way interactions were added to that model and then the 3-way interaction effect.

The similarity hypothesis predicts no main effect of Doctor, i.e. that the BF for the Null model (BF_{01}: robot and human perform about equally) would be higher than a model including the Doctor main effect (BF_{10}: robot and human perform differently). For the model that did include the single main effect of Doctor, BF_{01} was < 1, indicating that there was no evidence in favor of the Null. Instead, it obtained a high BF_{10}, explaining *Doctor Evaluation* better than a model without any predictors.

In fact, the model with the single Doctor main effect had the highest BF_{10} of all models tested (i.e. robot and human differ), including models tested but not reported in Table 2. Compared to a model with all main effects, the model with the single Doctor main effect was preferred by a Bayes Factor of 49.09 (BF_Doctor / BF_AllMain). This preference became even more pronounced when the Doctor-only model

[1] This is the reverse result of [6], which probably had a scale-analysis issue [see Online Resource 3, section *Reanalysis of Burgers et al.* (2012)].

Table 2 Bayesian ANOVA results for Doctor Evaluation, Message Evaluation, Medical Adherence, and Expected Quality of Life

No.	Included effects on *Doctor Evaluation*	BF_{01}	BF_{10}
1.	Doctor	.059	17.084
2.	Language	6.395	.156
3.	Frame	7.190	.139
4.	All main effects	2.874	.348
5.	No. 4. + all 2-way interactions	35.852	.028
6.	No. 5. + 3-way interaction	24.130	.041
No.	Included effects on *Message Evaluation*	BF_{01}	BF_{10}
1.	Doctor	1.35E−10	7.40E+09
2.	Language	2.922	.342
3.	Frame	1.021	.979
4.	All main effects	1.20E−10	8.36E+09
5a.	Doctor + Frame + Doctor * Frame	2.71E−11	3.69E+10
5b.	No. 4. + all 2-way interactions	6.19E−10	1.62E+09
6.	No. 5. + 3-way interaction	3.48E−10	2.88E+09
No.	Included effects on *Medical Adherence*	BF_{01}	BF_{10}
1.	Doctor	1.128	.886
2.	Language	4.232	.236
3.	Frame	5.335	.187
4.	All main effects	25.364	.039
5.	No. 4. + all 2-way interactions	795.698	.001
6.	No. 5. + 3-way interaction	2481.664	4.03E−04
No.	Included effects on *Expected Quality of Life*	BF_{01}	BF_{10}
1.	Doctor	6.257	.16
2.	Language	2.58E−07	3.88E+06
3.	Frame	.010	98.335
4.	All main effects	1.79E−08	5.58E+07
5a.	No 4. + Doctor * Language and Doctor * Frame	6.59E−11	1.52E+10
5b.	No. 4. + all 2-way interactions	1.99E−10	5.03E+09
6.	No. 5. + 3-way interaction	8.88E−10	1.13E+09

was compared to a model including all 2-way interactions as well as the 3-way interaction (Bayes Factor = 416.68).

3.2.2 Message Evaluation

Table 2 (second panel) shows partial output of the Bayesian ANOVA ran in JASP. We calculated all main effects alone, after which they were compacted into one model. Subsequently, the model with the highest BF_{10} was included (Doctor and Frame main effects and their interaction). Then we added all 2-way interaction effects and finally, the 3-way interaction.

Again, the similarity hypothesis expects that the BF for the Null would be high compared to a model with the Doctor main effect. However, for the model with the single main effect of Doctor, $BF_{01} < 1$, rejecting the Null. This was confirmed by a high BF_{10}, evidencing that the Doctor main effect explained *Message Evaluation* better than a model with no predictors.

While the BF_{10} for the All-main-effects-model was high, one model resulted into an even higher BF_{10}: This model included the main effects of Doctor and Frame, and their interaction. Compared to the All-main-effects-model, this model was preferred by a Bayes Factor of 4.42.

3.2.3 Medical Adherence

Table 2 (third panel) shows main effects on their own, then compiled into one model, after which the model with the highest BF_{10} was included (Doctor and Frame main effects and their interaction). Then we included the 2-way interactions and lastly, the 3-way interaction.

As before, we predicted the Null against the Doctor main effect. Yet, the BF_{01} for the latter (Table 2) was 1.128, indicating some evidence in favor of the Null that robot and human acted about equally well, which was supported by a low BF_{10}. In predicting *Medical Adherence* scores, then, human and robot doctor were about the same. None of the other BF_{10} reached a level higher than 1. Instead, adding more predictors to the model resulted in an ever-declining Bayes Factor.

3.2.4 Expected Quality of Life

Table 2 (fourth panel) shows the main effects, followed by all main effects combined into one model. The model with the highest BF_{10} was included (All-main-effects plus the two interaction effects that included Doctor) and then we added all 2-way interactions, and the 3-way interaction.

The similarity hypothesis assumes the absence of a main effect of Doctor, so that the BF would favor the Null. Indeed, the BF_{01} for the main effect of Doctor was 6.257 with a low BF_{10}, suggesting evidence for the Null model: It seems that Robot performed similar to Human in predicting *Expected Quality of Life* scores. However, this was not the best fit. The BF_{10} for the model including all main effects and two of the two-way interactions was highest for all tested models. Compared to a model with main effects only, this model was preferred by a Bayes Factor of 272.29, and to a model with all two-way interaction effects by a Bayes Factor of 3.02.

3.3 Effect of Ethics, Involvement, and Distance on Message-Related Outcomes: MANCOVA

In the Robot Doctor condition ($N = 134$), we also surveyed items on the doctor's *Ethics*, the *Involvement* she stimulated, and affective *Distance* she provoked (RQ1). To test the effect of these dependents on the message-related outcome variables, we treated them as covariates in a 2 (Language: Negation vs. Affirmation) × 2 (Frame: Negative vs. Positive) MANCOVA with all four outcome variables such as *Doctor Evaluation*. This was a frequentist analysis. Descriptives can be found in Table 1.

3.3.1 Multivariate Results

Two covariates were significant: *Ethics* ($\lambda = .90$, $F_{(4,124)} = 3.38$, $p = .012$, $\eta_p^2 = .10$) and *Involvement* ($\lambda = .68$, $F_{(4,124)} = 14.43$, $p < .001$, $\eta_p^2 = .32$). The main effect of Language also was significant ($\lambda = .76$, $F_{(4,124)} = 9.59$, $p < .001$, $\eta_p^2 = .24$), which was to be expected in view of the earlier MANOVA.

3.3.2 Univariate Results

Ethics had a univariate effect on *Doctor Evaluation* ($F_{(1,127)} = 4.21$, $p = .042$, $\eta_p^2 = .03$) and *Message Evaluation* ($F_{(1,127)} = 9.43$, $p = .003$, $\eta_p^2 = .07$): A higher *Ethics* score was related to a higher score on the outcome variables. *Involvement* had a significant effect on *Doctor Evaluation* ($F_{(1,127)} = 51.04$, $p < .001$, $\eta_p^2 = .29$), *Message Evaluation* ($F_{(1,127)} = 7.80$, $p = .006$, $\eta_p^2 = .06$), and *Medical Adherence* ($F_{(1,127)} = 12.64$, $p = .001$, $\eta_p^2 = .09$). Here too, a higher score on *Involvement* was related to a higher score on the outcome variables.

3.4 Effect of Conditions on Ethics, Involvement, and Distance: MANOVA

In the Robot Doctor condition ($N = 134$), we also tested whether the experimental conditions affected *Ethics*, *Involvement*, and *Distance*. Therefore, we performed a 2 (Language: Negation vs. Affirmation) × 2 (Frame: Negative vs. Positive) MANOVA on the three experiential variables. However, none of the multivariate effects were significant ($p > .154$), indicating that the Language and Framing conditions did not affect *Ethics*, *Involvement*, and *Distance*.

3.5 Effect of Affordances and Use Intentions on Message-Related Outcomes: MANCOVA

For a subset of participants that were in the Robot Doctor condition ($N = 68$), two more experiential variables were measured, namely *Affordances* and *Use Intentions*. To test their effect on the outcome variables such as *Doctor Evaluation*, we included them as covariates in a 2 (Language: Negation vs. Affirmation) × 2 (Frame: Negative vs. Positive) MANCOVA with all four outcome variables. Table 1 shows the descriptives.

3.5.1 Multivariate Results

Both covariates were significant: *Affordances* ($\lambda = .83$, $F_{(4,59)} = 2.93$, $p = .028$, $\eta_p^2 = .17$), *Use Intentions* ($\lambda = .61$, $F_{(4,59)} = 9.34$, $p < .001$, $\eta_p^2 = .39$). Given the earlier results, obviously, the main effect of Language also was significant ($\lambda = .44$, $F_{(4,59)} = 18.49$, $p < .001$, $\eta_p^2 = .56$).

3.5.2 Univariate Results

Affordances had a significant effect on *Message Evaluation* ($F_{(1,62)} = 7.44$, $p = .008$, $\eta_p^2 = .11$) and on *Medical Adherence* ($F_{(1,62)} = 8.09$, $p = .006$, $\eta_p^2 = .12$). In both cases,

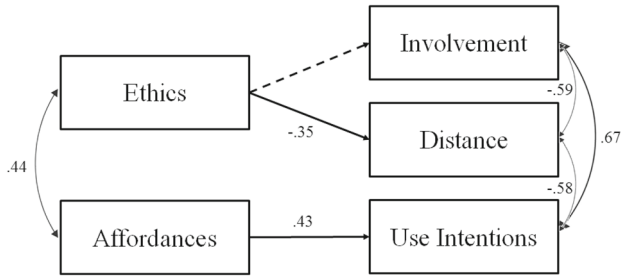

Fig. 2 Revised model for experiencing the robot doctor. Standardized estimated Bayesian parameter coefficients. The dashed path is not significant

higher *Affordances* was related to higher scores on the outcome variables.

The covariate *Use Intentions* had a significant effect on *Doctor Evaluation* ($F_{(1,62)} = 32.65$, $p = .000$, $\eta_p^2 = .35$) and *Message Evaluation* ($F_{(1,62)} = 4.13$, $p = .047$, $\eta_p^2 = .06$). Here too, a higher score on *Involvement* related to higher scores on the outcome variables.

3.6 Effect of Conditions on Affordances and Use Intentions: MANOVA

For the participants that were in the Robot Doctor condition with the extra two variables ($N = 68$), we tested whether *Affordances* and *Use Intentions* with regard to Robot Doctor were affected by the experimental conditions. We performed a 2 (Language: Negation vs. Affirmation) × 2 (Frame: Negative vs. Positive) MANOVA with both experiential variables. Yet, none of the multivariate effects were significant ($F < 1$, $p > .373$), indicating that the Language and Frame conditions did not affect *Affordances* and *Use Intentions.*

3.7 Bayesian Path Models

The experiential variables in unison are part of a nine-factor model for perceiving and experiencing fictional and virtual others such as robots [45]. In as far as possible, we retrieved this model in our data, while comparing two versions. The first did not allow *Ethics* and *Affordances* to covary (in line with [45]); the second did allow for such covariance (see Fig. 2). We used Mplus 7 to assess whether this model converged (for settings, see [47] in Online Resource 3).

The alternative model, which allowed for covariances, had a smaller Bayesian Information Criterion value (BIC = 1478.91) and a smaller Deviance value (DIC = 1426.13) than the original [45] version (BIC = 1487.40, DIC = 1437.81). This indicates that the original model fit the data worse than the alternative model. Raftery [34] stated that a BIC difference of > 10 is strong evidence against the model with a higher BIC value (in this case, the original [45] model). Even though the BIC Difference of 8.49 was below 10, the DIC Difference of 11.68 was higher than 10.

Figure 2 shows the (standardized) estimated Bayesian parameter coefficients of the revised engagement model for the paths that were significant in a Bayesian sense (i.e. the 95% confidence interval excludes 0). For details, see [47] (Online Resource 3).

One path was not significant: *Ethics* showed no effect on *Involvement* but instead, was a negative predictor of *Distance*, such that higher evaluations of the robot's *Ethics* were related to lower feelings of *Distance*. *Affordances* were a positive predictor of *Use Intentions*, indicating that a positive opinion of the robot's affordances were related to a higher evaluation of its *Use Intentions*.

Affordances and *Ethics* were positively correlated, meaning that the positive evaluations of the robot doctor's *Ethics* and *Affordances* often went together. *Distance* was negatively correlated with *Involvement* and *Use Intentions*, whereas *Involvement* and *Use Intentions* were positively correlated. The model explained 13.0% of the variance in *Distance*, 3.4% of the variance in *Involvement* (a non-significant predictor can still explain a limited amount of variance), and 21.3% of *Use Intentions.*

3.8 Interim Conclusions

With respect to the original hypotheses, then, the *similarity* hypothesis (H0) stated that scores for the robot doctor would follow the same pattern as for the human doctor. The *dissimilarity* hypothesis (H1) suspected for the doctor who does it right that lower scores would be given to the robot than to the human. When the doctor does it wrong, however, H1 expected higher scores for the robot. The latter was instilled by the idea that a robot—unskilled in human communication—would be pardoned for its ineptness; with higher scores than a human who did it wrong.

However, both hypotheses were strongly refuted by the main effects on *Doctor Evaluation*, *Message Evaluation*, and *Medical Adherence*, where the Robot Doctor gained significantly higher scores than the Human Doctor. In other words, Doctor Robot outperformed humans at these matters and did not 'follow the same pattern,' thus rejecting H0. And although there was a different level of performance between the two, it was into another direction than predicted by H1. It seems that Doctor Robot did not need to be forgiven because no matter what, it beat the human doctor at almost every dimension.

With respect to *Expected Quality of Life*, we found that it was higher in negative than in positive frames, which is the reverse result of [6] (see Sect. 4.1). The interaction between Doctor and Language was significant, indicating that when the Robot Doctor used negations instead of affirmations, scores would increase. This effect was absent for the Human Doctor, which is evidence against H0. There were no signif-

icant effects of Frame for the Robot Doctor but they were present for the Human Doctor. Hence, different patterns in scores to human and robot behavior countered H0 once more.

Bayesian analyses confirmed these conclusions. *Doctor Evaluation* scores were best explained by who the doctor was (robot raising higher scores), whereas Language and Frame played no role here. *Message Evaluation* also was best explained by the robot raising higher scores than the human, together with how the message was framed, and by the interaction between the two. Language played no role here. The frequentist analysis showed that negative framing was important for human doctors, leading to higher *Message Evaluation* scores. For robot doctors, framing did not affect *Message Evaluation*. Together, these results countered H0.

With respect to *Medical Adherence*, scores in the Bayesian analyses were not explained by Doctor, Language, or Frame. These results were not in agreement with the frequentist analyses, where we found that *Medical Adherence* was significantly higher for the Robot Doctor than for the Human Doctor. This discrepancy can be explained by the relatively small mean difference of 0.33 points between the groups combined with a relatively large sample size. Using frequentist methods with a large sample size can result in spurious significant effects that are not necessarily meaningful. Bayesian estimation can lead to decreases in BF with increasing sample size, if the mean difference is not big and specific (low variance within groups) enough to be meaningful [20]. Therefore, if we believe the Bayesian analysis, we can confirm that for *Medical Adherence*, Doctor Robot performed similarly to a Human Doctor, confirming H0. If we follow the frequentist analysis, Doctor Robot outdid the human. That Language and Frame should play a role in *Medical Adherence* could not be confirmed for the Robot Doctor, rejecting H0.

In the Bayesian analyses, we also saw that *Expected Quality of Life* scores were best explained by a combination of who the Doctor was (Human or Robot), what Language was used (Affirmation or Negation), and how the message was Framed (Positive or Negative), and the interaction between Doctor * Language and Doctor * Frame. The interaction between Frame and Language seemed unimportant. These results are partially in agreement with the frequentist results, the exception being that the main effect of Doctor was not statistically significant in the frequentist MANOVA. This is where the difference between Bayesian ANOVA and frequentist MANOVA becomes clear. Instead of testing each effect on its own, the Bayesian ANOVA tries to find the best fitting overall model for explaining, here, the *Expected Quality of Life* scores. In this context, the main effect of Doctor did result in an improved model fit. Therefore, it was included even if its individual effect was non-existent (see $BF_{01} > 1$ for Doctor-only model). Hence, if we focus on the high BF_{01} for the Doctor-only model, we can confirm that Robot and Human worked comparably well at raising *Expected Quality of Life*, which could count as a corroboration of H0. Yet, we could not confirm that a combination of Affirmative Language and Positive Frames affected *Expected Quality of Life*, as the preferred model did not contain this interaction. In fact, Affirmation was associated with lower *Expected Quality of Life* scores ($M = 3.53$, $SE = .12$, 95% CI 3.29–3.78) than Negation ($M = 4.57$, $SE = .12$, 95% CI 4.33–4.81). Regarding Frame, we found that Positive Frames actually lowered *Expected Quality of Life* scores ($M = 3.69$, $SE = .12$, 95% CI 3.45–3.93) compared to Negative Frames ($M = 4.41$, $SE = .12$, 95% CI 4.17–4.66).

Finally, RQ1 wondered how participants might have experienced the robot doctor, measuring its moral fiber and the like. Was Doctor Robot seen as distant and clumsy? Actually not: *Ethics* had a positive effect on *Doctor Evaluation* and *Message Evaluation*. The attribution of better moral behavior made the robot a better doctor, delivering a better message. The Robot Doctor was not seen as cold since the multivariate effects of Distance were not significant ($\lambda = .95$, $F_{(4,124)} = 1.79$, $p = .135$). Involvement with 'warm' Doctor Robot had (sometimes strong) positive effects on *Doctor Evaluation*, *Message Evaluation*, and *Medical Adherence*. Feeling more involved with the robot made it a better doctor, delivering a better message that would be followed up. Doctor Robot appeared to be skillful: Affordances had positive effects on *Message Evaluation* and *Medical Adherence*. Better skills and action possibilities made a better robot doctor, offering advice that was followed. Participants also wished to see 'warm' Doctor Robot more often: Use Intentions had positive effects on *Doctor Evaluation* and *Message Evaluation*. Increased intentions to use the robot made it a nicer doctor, delivering a better message.

The model of perceiving and experiencing fictional characters and virtual others (Fig. 2) based on [45] showed that *Ethics* and *Affordances* were interconnected (good was also skillful) and (again) they were main determinants of engagement, lowering *Distance* when the robot was seen as morally good and raising *Use Intentions* when the robot was seen as competent.

4 Discussion

Different from Kelsey and St. Amant [19, p. 867], to regard robots as social, they do not need to express emotions too much, do not have to show a specific personality, nor work with natural cues, and do not have to be capable of conducting high-level dialogue. We repeated Experiment 2 of [6] with positive-negative framing and affirmative-negative language, confronting 134 participants with a robot doctor that brought bad news about Bekhterev's rheumatic disease. H0 expected effects to be quite similar to the assorted effects of a human

doctor. H1 predicted mediocre scores for robots with human doctors obtaining the highest and lowest scores dependent on the way they communicated the health message.

At three out of four dimensions (*Doctor Evaluation*, *Message Evaluation*, and *Medical Adherence*), the robot beat the human doctor (main effects). At *Expected Quality of Life*, the two tied. To increase *Expected Quality of Life*, the robot had to use negations, whereas the human had to use negative frames. The human also had to use negative frames to enhance *Message Evaluation*. Looking into the stimulus materials, *Expected Quality of Life* was said to (not) progress or (not) deteriorate. If a robot is to use negations (not progress and not deteriorate), the robot actually is preferred when things remain stable; when nothing changes compared to how it was. If for a human the emphasis is on the negative outcome but not so much on the language (not progress and deteriorate), then a human doctor seems the preferred option; when things get serious and worsen. Note, however, that there were some issues with the measurement of *Expected Quality of Life* (Online Resource 3), so that results should be taken with caution.

Putting it differently, robot and human did equally well at times, supporting H0; or the robot surpassed the human, even in cases where H1 did not expect it. Ergo, the human doctor never outperformed Doctor Robot except (if we trust the measurement) when someone's quality of life was expected to worsen.

There may be some nuance to these conclusions, dependent on the type of analysis one adheres to. If the reader follows the frequentist analyses, proper framing seemed more of an issue for humans (*Message Evaluation, Expected Quality of Life*) than for robots. If the reader follows Bayesian analysis, robots had similar restrictions on how to convey the news. At one point (*Expected Quality of Life*), language was important for robots (i.e. negation use) but not for humans. In other words, robots could frame it and say it any way they wanted with sometimes an exception. Nevertheless, the robot outperformed the human at *Doctor Evaluation* (robot was nicer) and *Message Evaluation* (robot delivered a better message), and performed about equally well as a human on *Medical Adherence* (if you do not believe Bayes, then the robot performed even better). The main effects on *Expected Quality of Life* were insignificant so that may count as a tie.

We also measured how the robot was experienced in terms of *Ethics*, *Affordances*, *Involvement*, *Distance*, and *Use Intentions* [45]. Doctor Robot was not seen as distant but rather involving, morally good, skillful, and evoking the willingness to see her again.

Yet, it is strange that we obtained these results although the robot lacked interactive capacity and just delivered a prerecorded message. Imagine what the robot would be capable of if it adapted its tone of voice and facial expression to the gravity of the diagnostic, show empathy, or adaptive behaviors [8]? Would that improve the appreciation of the robot even more? The question that arises is why the robot doctor performed so well at her communication tasks in spite of its lack of interaction and why, moreover, it was regarded as 'ethical' (although she made no moral statements), not cold and distant but rather warm (although no emotion simulator was running), and involving (although it spoke with monotonous voice)? Doctor Robot was considered 'skilled' (but merely recited a predefined message) and she invited future use although the electronic girlish machine hardly resembled a realistic physician.

In view of the Media Equation [35] and Computers Are Social Actors, the CASA paradigm [32], our results may be interpreted that indeed people listen to a robot as if it were human. However, only occasionally do they take the same factors into consideration and evaluate the performance as high. The *as if* is important here, because people are not mistaken by who the source of communication is (i.e. a nonhuman) and hence apply a different yardstick [9] as they do with pet animals for instance. In other words, human-like behaviors evoke human-like, not human-equal, evaluations.

Why was Doctor Robot a better messenger of bad news than the human doctor? Doctors prefer to sweeten the bitter pill by using negations such as "The news is not so good," whereas patients wish to know what doctor actually thinks [7]. Could it be that the robot is not suspected of doubt and reservation? That the robot is not expected to want to regulate its own emotions nor those of its patient?

Perhaps that, reversely, the patient does not want to regulate the emotions of the doctor. After all, delivering bad news is as demanding for the messenger as it is for the patient. If the patient has to deal with more emotions than his or her own, less cognitive capacity will be available to process the information of the message. Perhaps that a doctor's well-meant empathy and his sorry face are too much. Maybe professional distance should be restored so that the patient can share her emotions with the ones she loves, not with the one who brings her the bad news. In facing a robot, you do not have to feel embarrassed if you start crying nor will it have second thoughts about you.

Robots are not humans and perhaps people assume that robots have no desires of their own. Therefore, people probably believe that robots have no hidden agendas. They are not judgmental, have no critique. They merely deliver the message, without further ado. And without these negative features, they may gain more credibility. The absence of negative qualities may heighten the acceptance of the message (cf. [33]).

For example, only twice were there effects of framing for the robot (according to Bayes) and at one time did language play a role. It seemed that how a message was framed was important when the source of the message was a human, less so when it was a machine. In that case, predictions based on

Prospect Theory [17] are limited by who the sender of the message is (i.e. humans).

Buhlmann and Gisler [5] observed that messengers receive more credibility when they are direct, brief, and result-oriented, things a robot does naturally. The absence of goals of its own may perhaps count as an indicator of ethical behavior of the robot (e.g., honest, sincere, trustworthy). Based on this assumed integrity, receivers would tend to believe the message. And even if they did not remember the message content too well, they probably did recall how the messenger made them feel (cf. [5]); in the case of the robot doctor, as an ethical entity, which reduced the receiver's affective distance (Fig. 2).

Would it be ethical, then, to employ a robot for bad news conversations? For now, this question comes too early because we do not know the ins and outs of what the robot is allowed to do morally. Yet, certain people may regard the robot as more ethical than humans because it has no hidden agenda. For others, it would not be a sign of 'good care' to leave sensitive matters to machines. However, what is always good is that people have the opportunity to choose between a human and a machine as the bearer of bad news. And for the human messenger, that might be a relief as well.

It seems, then, that direct, almost impolite, communication works best for robots. That is to say, Torrey et al. [43] found that when robot as well as human cooking assistants communicated with or without 'hedges' (words that mitigate the impact of an utterance), respondents found them more friendly, empathic (i.e. considerate), and less controlling. Robots even outdid human helpers on these dimensions when they modulated their communication through hedges. On the one hand, this supports our finding that robots can be better communicators than humans. On the other hand, the robot in the cooking study achieved this result by using subtle communications whereas in our study, it did so in the absence of communicative modifications. Perhaps that the different tasks explain the different communications required: If you help a novice to achieve a modest goal (e.g., to bake muffins), it is good to be gentle but if bad news about serious matters (i.e. illness) has to be delivered, perhaps that is seen as beating around the bush and people prefer to get to the point straight away.

4.1 Limitations

We could replicate the response patterns for the message-related outcome variables in the human-doctor data of [6], Experiment 2, but not always for *Expected Quality of Life*. In [6], negative frames yielded higher scores than positive frames but not for *Expected Quality of Life*: positive was higher than negative. In our data set (both human and robot), *Expected Quality of Life* always raised higher scores in negative frames. We suspected a difference in scale construction. Therefore, Winter and Hoorn [47] analyzed all four message-related outcome variables in three ways (Online Resource 3): We established divergent validity of *Expected Quality of Life* through EFA—something [6] did not do; we did a full replication of [6], using their exact scale items, and we followed a best-performing single-item approach. There were but a few differences in the results but one of them was crucial. In the full replication of [6], the effects of Frame on *Expected Quality of Life* were not significant and the Doctor * Language interaction effect on *Message Evaluation* was significant only in the general ANOVA, not in the pairwise contrasts. That Positive Frame scored higher on *Expected Quality of Life* than Negative Frame could only be replicated for the human doctor using a single item, not with the exact same scale [6] used, which rendered insignificant results (Online Resource 3, [47, Table 16]). This led us to think that we should stick to our own approach of calculating divergent validity and constructing the scale based on factor analysis, and hence to conclude that Negative Frames were preferred for each of the four outcome variables.

Nonetheless, the finding remains that irrespective of interaction effects, the robot excelled at delivering bad news; comparable or even better than a human. Of course, this result is limited by the specific robot we used and the specific human that was tested. We did not show it for other robots or other humans. But according to good Popperian tradition, finding one example refuting the theory already is scientific progress and it is still remarkable that even one single robot can do better than a human in affect-sensitive communications (which is support of H1).

There might have been confounds: Maybe the Dutch culture favored the rather direct way of news delivery by the robot. In Asian cultures, such directness may be seen as rude and may not be the preferred communication style. Maybe a facially expressive robot does the job but a machine with fewer human features not. Or it might be that less humanness increases credibility because humans are associated with hidden intentions and being judgmental? Perhaps that the robot exerted a novelty effect that boosted the scores to its performance. Perhaps it was enthusiasm for robotics for its own sake, surprise, self-efficacy, locus of control or lack thereof, and so forth. However, it is infeasible to bother participants with each possible variable that might confuse results although admittedly, adding a novelty scale would have done no harm. Yet, although we cannot exclude the interference by some sort of confounding variable and hence we cannot conclude what exactly made the robot a better messenger of bad news, the fact remains that the robot *was* the better messenger. Was it then that Doctor Robot did better because both doctors were on film and not present in real life? The human doctor was on film to exactly replicate the stimulus. Yet, maybe a real doctor in real life has so much presence that s/he is still preferred over a robot.

That begs the question what the robot would do in a real hospital setting with real patients? How would people react if they are told by the robot that they are diagnosed for cancer, for instance? Probably for severe cases, people should be informed by a human but for the minor ailments (e.g., skin condition, constipation, or haemorrhoids), the robot might suffice.

Yet, we should take one thing at a time. Perhaps the current study should be regarded as a good first start. Next we should conduct a fully controlled study in a lab-like environment or at least a computer classroom-like environment where multiple participants can engage in experimental sessions with their own terminals. The proof of the pudding would be to conduct yet another study with a real doctor in an actual setting contrasted with a (lookalike) robot doctor in the same setting. We could also measure the experiential variables (ethics, affordances, etc.) for the human doctor to provide more backdrop to the message-related outcomes.

5 Conclusion

In all, the Media Equation thesis holds that people treat computers as social actors. Similarly do people treat robots as social actors; not like humans but as species of their own. People are not as simple as Reeves and Nass [35] thought. That the robot shows no signs of a personal ego with its selfish goals and concerns may make it an ethical social entity. Perhaps that in addition, reduced emotional expressions puts the locus of attention on the information (here, health) and less so on the inner struggle of the sender. In view of our results, telemedicine with a human may not always be the preferred way of communication: A humanoid robot on a screen may work just as well, sometimes even better. Media do not equal real life but they come close and in the case of social robots their near-humanness is what makes them stand out against 'conventional' humans with all their personal issues.

We started this study from the wisdom that it was better to have a good robot than a bad human. But reality is more harsh on humanist and human-centric certainties: Our results suggest that the new saying should be that *a robot a day keeps the doctor away.* No one likes having to bring bad news, but a robot does not care. And we think that is why it does it better.

Acknowledgements This study is part of the Services of Electromechanical Care Agencies (SELEMCA) project and was supported by a grant from the Dutch Ministry of Education, Culture, and Science (Grant Number NWO 646.000.003). Many thanks go to Christian Burgers for making available the data on the human doctor and for his comments and advices. Marcel Nihot is kindly thanked for data collection in the robot experiment. We acknowledge with much appreciation Elly A. Konijn who reviewed an earlier draft of this paper. We gratefully acknowledge the comments and suggestions of the anonymous reviewers for the profound improvement of this paper.

References

1. Baile WF, Buckman R, Lenzi R, Glober G, Beale EA, Kudelka AP (2000) SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer. Oncologist 5(4):302–311
2. Bickmore T, Cassell J (2001) Relational agents: a model and implementation of building user trust. In: Jacko J, Sears A (eds) Proceedings of the SIGCHI conference on human factors in computing systems (CHI '01), Seattle, WA, 31 March–05 April 2001. ACM, New York, NY, pp 396–403
3. Bickmore T, Picard R (2005) Establishing and maintaining long-term human–computer relationships. Trans Comput Hum Interact 59(1):21–30
4. Brave S, Nass C, Hutchinson L (2005) Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. Int J Hum Comput Stud 62:161–178
5. Buhlmann H, Gisler A (2006) A course in credibility theory and its applications. Springer, New York
6. Burgers C, Beukeboom CJ, Sparks L (2012) How the doc should (not) talk: when breaking bad news with negations influences patients' immediate responses and medical adherence intentions. Patient Educ Couns 89(2):267–273. https://doi.org/10.1016/j.pec.2012.08.008
7. Brown VA, Parker PA, Furber L, Thomas AL (2011) Patient preferences for the delivery of bad news: the experience of a UK cancer centre. Eur J Cancer Care 20(1):56–61
8. Culnan MJ, Markus ML (1987) Information technologies: electronic media and interorganizational communication. In: Jablin FM, Putnam LL, Roberts KH, Porter LW (eds) Handbook of organizational communication: an interdisciplinary perspective. Sage, Newbury Park, pp 420–443
9. Derks D, Bos AER, Von Grumbkow J (2008) Emoticons in computer-mediated communication: social motives and social context. CyberPsychol Behav 11:99–101
10. De Melo CM, Carnevale P, Gratch J (2011) The effect of expression of anger and happiness in computer agents on negotiations with humans. In: Tumer K, Yolum P, Sonenberg L, Stone P (eds) Proceedings of the 10th international conference on autonomous agents and multiagent systems (AAMAS '11), vol 3, Taipei, Taiwan, 02–06 May 2011. IFAAM, Richland, pp 937–944
11. Fischer K, Foth K, Rohlfing K, Wrede B (2011) Mindful tutors: linguistic choice and action demonstration in speech to infants and a simulated robot. Interact Stud 12(1):134–161
12. Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Chapman & Hall/CRC, Boca Raton
13. Gray K, Wegner DM (2014) Feeling robots and human zombies: mind perception and the uncanny valley. Cognition 125(1):125–130
14. Hoorn JF, Konijn EA, Germans DM, Burger S, Munneke A (2015) The in-between machine: the unique value proposition of a robot or why we are modelling the wrong things. In Loiseau S, Filipe J, Duval B, Van den Herik J (eds) Proceedings of the 7th international conference on agents and artificial intelligence (ICAART), Lisbon, Portugal, 10–12 Jan 2015. ScitePress, Lisbon, pp 464–469
15. Jeffreys H (1961) Theory of probability, 3rd edn. Oxford University, Oxford

16. Johnson SC (2003) Detecting agents. Philos Trans R Soc B Biol Sci 358(1431):549–559
17. Kahneman D, Tversky A (2000) Choices, values and frames. Cambridge University Press, Cambridge
18. Kanda T, Hirano T, Eaton D, Ishiguro H (2004) Interactive robots as social partners and peer tutors for children: a field trial. Hum Comput Interact 19:61–84
19. Kelsey S, St. Amant K (2008) Handbook of research on computer mediated communication. Information Science Reference, Hershey
20. Konijn EA, Van de Schoot R, Winter SD, Ferguson CJ (2015) Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. Commun Methods Meas 9(4):280–302
21. Küster D, Świderska A (2016) Moral patients: what drives the perceptions of moral actions towards humans and robots? In: Seibt J, Nørskov M, Schack Andersen S (eds) What social robots can and should do: proceedings of robophilosophy, frontiers of artificial intelligence and applications (TRANSOR 2016), Aarhus, Denmark, 17–21 Oct 2016. IOS Press, Amsterdam, pp 340–343. https://doi.org/10.3233/978-1-61499-708-5-340
22. Latour B (1992) Where are the missing masses? The sociology of a few mundane artifacts. In: Bijker WE, Law J (eds) Shaping technology, building society: studies in sociotechnical change. MIT, Cambridge, pp 225–258
23. Lee KM, Peng W, Jin S-A, Yan C (2006) Can robots manifest personality? An empirical test of personality recognition, social responses, and social presence in human–robot interaction. J Commun 56:754–772
24. Lee MD, Wagenmakers E-J (2013) Bayesian modeling for cognitive science: a practical course. Cambridge University, New York
25. Lee MK, Kiesler S, Forlizzi J (2010) Receptionist or information kiosk: How do people talk with a robot? In: Inkpen K, Gutwin C, Tang J (eds) Proceedings of the 2010 ACM conference on computer supported cooperative work (CSCW '10), Savannah, GA, 06–10 Feb 2010. ACM, New York, pp 31–40
26. Lee MK, Kiesler S, Forlizzi J, Rybski P (2013) Ripple effects of an embedded social agent: a field study of a social robot in the workplace. In: Mackay WE, Brewster S, Bødker S (eds) Proceedings of the SIGCHI conference on human factors in computing systems, Paris, France, 27 April–02 May 2013. ACM, New York, pp 695–704
27. Lorenz T, Weiss A, Hirche S (2016) Synchrony and reciprocity: key mechanisms for social companion robots in therapy and care. Int J Soc Robot 8(1):125–143. https://doi.org/10.1007/s12369-015-0325-8
28. Love J, Selker R, Verhagen J, Marsman M, Gronau QF, Jamil T, Šmíra M, Epskamp S, Wild A, Morey R, Rouder J, Wagenmakers EJ (2015) Java based statistical processor (JASP) [computer software]. https://jasp-stats.org/. Accessed 16 May 2016
29. Monden KR, Gentry L, Cox TR (2016) Delivering bad news to patients. Proc (Bayl Univ Med Cent) 29(1):101–102
30. Morey RD, Rouder JN (2015) BayesFactor [computer software]. http://bayesfactorpcl.r-forge.r-project.org/. Accessed 12 May 2016
31. Mutlu B, Forlizzi J (2008) Robots in organizations: the role of workflow, social, and environmental factors in human–robot interaction. In: Fong T, Dautenhahn K, Scheutz M, Demiris Y (eds) Proceedings of the 3rd ACM/IEEE international conference on human robot interaction (HRI '08), Amsterdam, The Netherlands, 12–15 March 2008. ACM, New York, pp 287–294
32. Nass C, Moon Y (2000) Machines and mindlessness: social responses to computers. J Soc Issues 56(1):81–103
33. Ohanian R (1990) Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. J Advert 19(3):39–52. https://doi.org/10.1080/00913367.1990.10673191
34. Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 25:111–163
35. Reeves B, Nass C (1996/2002) The Media Equation: how people treat computers, television, and new media like real people and places. CSLI, Stanford
36. Sabelli AM, Kanda T, Hagita N (2011) A conversational robot in an elderly care center: An ethnographic study. In: Billard A, Kahn P, Adams JA, Trafton G (eds) Proceedings of the 6th international conference on Human–robot interaction (HRI'11), Lausanne, Switzerland, 06–09 March 2011. ACM, New York, pp 37–44
37. Severinson-Eklundh K, Green A, Hüttenrauch H (2003) Social and collaborative aspects of interaction with a service robot. Robot Auton Syst 42(3–4):223–234
38. Street R, Makoul G, Arora N, Epstein R (2009) How does communication heal? Pathways linking clinician–patient communication to health outcomes. Patient Educ Couns 74(3):295–301
39. Sundar SS, Nass C (2000) Source orientation in human–computer interaction. Commun Res 27(6):683–703
40. Sung J-Y, Guo L, Grinter RE, Christensen HI (2007) "My Roomba is Rambo": intimate home appliances. In: Krumm J, Abowd GD, Seneviratne A, Strang T (eds) UbiComp 2007: ubiquitous computing. UbiComp 2007. Lecture Notes in Computer Science, vol 4717. Springer, Berlin, Heidelberg, pp 145–162. https://doi.org/10.1007/978-3-540-74853-3_9
41. Takayama L, Go J (2012) Mixing metaphors in mobile remote presence. In: Poltrock S, Simone C, Grudin J, Mark G, Riedl J (eds) Proceedings of the ACM 2012 conference on computer supported cooperative work, Seattle, WA, 11–15 Feb 2012. ACM, New York, pp 495–504
42. Tapus A, Matarić MJ (2008) Socially assistive robots: the link between personality, empathy, physiological signals, and task performance. In: Lisetti C, Hudlicka E, Horswill I, Velásquez JD (eds) Papers of the AAAI spring symposium: emotion, personality, and social behavior. AAAI, Menlo Park, pp 133–140
43. Torrey C, Fussell SR, Kiesler S (2013) How a robot should give advice. In: Kuzuoka H, Evers V, Imai M, Forlizzi J (eds) Proceedings of the 8th ACM/IEEE international conference on human–robot interaction (HRI '13), Tokyo, Japan, 3–6 March 2013. IEEE, Piscataway, pp 275–282. https://doi.org/10.1109/HRI.2013.6483599
44. Van de Schoot R, Depaoli S (2014) Bayesian analyses: where to start and what to report. Eur Health Psychol 16(2):75–84
45. Van Vugt HC, Hoorn JF, Konijn EA (2009) Interactive engagement with embodied agents: an empirically validated framework. Comput Animat Virtual Worlds 20:195–204
46. Van Vugt HC, Hoorn JF, Konijn EA, De Bie Dimitriadou A (2006) Affective affordances: improving interface character engagement through interaction. Int J Hum Comput Stud 64(9):874–888. https://doi.org/10.1016/j.ijhcs.2006.04.008
47. Winter SD, Hoorn JF (2016) Robot vs. human doctor. Technical report. Vrije Universiteit, Amsterdam
48. Woods SN, Walters ML, Koay KL, Dautenhahn K (2006a) Comparing human robot interaction scenarios using live and video based methods: towards a novel methodological approach. In: Sabanovic A (ed) Proceedings of the 9th IEEE international workshop on advanced motion control (AMC '06), Istanbul, Turkey, 27–29 March 2006. IEEE, Piscataway, pp 750–755
49. Woods SN, Walters ML, Koay KL, Dautenhahn K (2006b) Methodological issues in HRI: a comparison of live and video-based methods in robot to human approach direction trials. In: Dautenhahn K (ed) Proceedings of 15th IEEE international symposium on robot and human interactive communication (RO-MAN '06), Hatfield, UK, 06–08 Sept 2006. IEEE, Piscataway, pp 51–58

50. Zhou S, Bickmore T, Paasche-Orlow M, Jack B (2014) Agent-user concordance and satisfaction with a virtual hospital discharge nurse. In: Bickmore T, Marsella S, Sidner C (eds) International conference on intelligent virtual agents (IVA 2014), Boston, MA, 27–29 Aug 2014. Lecture notes in computer science, vol 8637. Springer, New York, pp 528–541. https://doi.org/10.1007/978-3-319-09767-1_63

Johan F. Hoorn graduated with two transdisciplinary Ph.D.-theses: his first Ph.D. was in Literature and Psychology (VU University Amsterdam, 1997), while he obtained his second Ph.D. degree in Computer Science (VU University Amsterdam, 2006). He worked at Utrecht University, Tilburg University, and in four different schools at VU University Amsterdam (Humanities, Science, Life Sciences, and Social Science). Johan was an adjunct professor of the Hong Kong Polytechnic University and the former director of the Center for Advanced Media Research Amsterdam at VU University. Johan was the principal investigator and project leader of SELEMCA, a multi-million, multi-stakeholder research and design project in social robotics for the care domain, granted by the Ministry of Education, Culture, and Science. Currently, he is the founding father and chairman of the Social Robotics Pop-up Lab, a research and design facility for the public at large (www.robopop.nl).

Sonja D. Winter graduated with an M.Sc. in Developmental Psychology from Utrecht University (2013). She is currently pursuing her Ph.D. in Quantitative Psychology at the University of California, Merced. Her research focuses on assessing the potential benefits of Bayesian estimation, specifically in the context of Structural Equation Modeling.