Do we trust social robots?

Lorenzo COMINELLI[1] Francesco FERI[2] Roberto GAROFALO[1]

Caterina GIANNETTI[1,3] Miguel A.MELENDEZ-JIMENEZ[4] Alberto GRECO[1]

Mimma NARDELLI[1] Enzo Pasquale SCILINGO[1] Oliver KIRKCHAMP[5]

21st September 2020

Abstract. Understanding human trust in machine partners has become an imperative following the widespread use of intelligent machines in a variety of applications and contexts. The aim of this paper is to study experimentally whether human-beings trust a social robot - i.e. a *human-like* robot that embodies emotional states, empathy and non-verbal communication - differently than other types of agents. To do so, we adapt the well-known economic trust-game proposed by Charness and Dufwenberg (2006) to assess whether receiving a promise from a robot increases human-trust in it. We find that receiving a promise from the robot increases the trust of the human in it, but only for individuals who perceived the robot very similar to a human-being. Importantly, we could replicate a similar pattern in choices when we replaced the humanoid counterpart with a real human but not when it was replaced by a computer-box. We additionally find that human participants' psychophysiological reaction is stronger when confronted with the humanoid.

Introduction

Trust is considered as a social glue that connects people and promotes collective goals. It is normally defined as the "intention to accept vulnerability based on the positive expectations or beliefs regarding the intentions or behaviour of other people in general" [1]. As a consequence, behavioral science has always been interested in trust, and more particularly in its influence on decision making [2, 3]. In parallel, trust also is relevant if we want to build social artificial agents that interact alongside with

[1]Department of Information Engineering and Center E. Piaggio. University of Pisa. Italy
[2]Department of Economics. Royal Holloway University of London, United Kingdom.
[3]*Corresponding author:* Department of Economics and Management. University of Pisa. Email: caterina.giannetti@unipi.it
[4]Department of Economic Theory and Economic History. University of Malaga
[5]Chair of Behavioural and Experimental Economics. University of Jena

people (e.g. robo-advisors, co-working robots, assistive robots, etc.) and take responsible roles in our society [4, 5]. A lesson learned from previous research and inter-disciplinary evidence (e.g. economics, neuroeconomics, psychology) is that (general) trust is deeply rooted in social experiences, being more a matter of culture than genetics [1], and highly affected by the emotional states of the individuals [6, 7, 8]. Indeed, emotions have been proven to play a fundamental role in the decision making process in general[9], as confirmed among other neuroscientists, by Damasio and colleagues in their studies [10, 11, 12, 13].

This stream of research thus suggests that trust and emotions are highly intertwined in the decision-making process in human-human interactions [14, 15, 16, 17], and may act as reasonable drivers in human-robot interactions as well [18]. It has been shown, for example, that not binding communications (i.e. cheap talk) is beneficial not only among humans but also to achieve higher cooperation when interacting with a machine (e.g [19]). In particular, a simple conversation with a robot changes individual behaviour towards the artificial agent [4, 20]. Very similar behavioural responses can be observed in children [4]. More in general, increasing the anthropomorphic features and the human social skills of a technology (e.g. by adding a name or a human voice to an autonomous vehicle) increases the individual willingness to accept and trust the technology itself (e.g. [21, 22, 13]).

Nonetheless, while the importance of emotions in driving the choice of a human to trust another human has been highly studied, less evidence is available when the decision to trust involves the interaction between artificial agents and humans ([23], [7, 21]). Moreover, we know that trust is highly culturally based, and that the appearance of the robot (especially its human-likeness, see [24]) affects the emotions perceived by its interlocutors. Therefore, studies on human-robot interactions and trust should always be repeated with different robot players having different aesthetics.

On that premise, the present study investigates how trust in a social robot is affected by its human likeness (both in terms of aesthetics and speech content), while taking into account the emotional states of the players during the interaction through physiological signal processing. The objectives are two-fold. On the one side, we can gain insights on how human-likeness interacts with emotions to instill people's trust in artificial agents, comparing it with that in human parterns so as to asses the differences (if any).[1] On the other side, we can gain a better understanding on how to design machines - both in terms of appearance and (e.g. communication) behaviour - in a way that help facilitate a fruitful inter-

[1] *Integral emotions* are emotions arising from the choice at hand and strongly shapes, and possibly bias, decision making. For example, a person who feels anxious about the potential outcome of a risky choice may choose the safer option.[9] On the other hand, *incidental emotions* are by definition unrelated to the outcomes under considerations although may still cause alterations in the choice process.

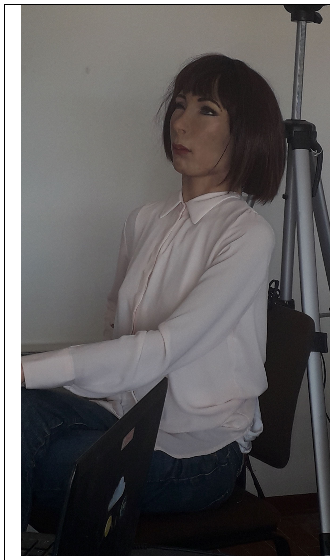
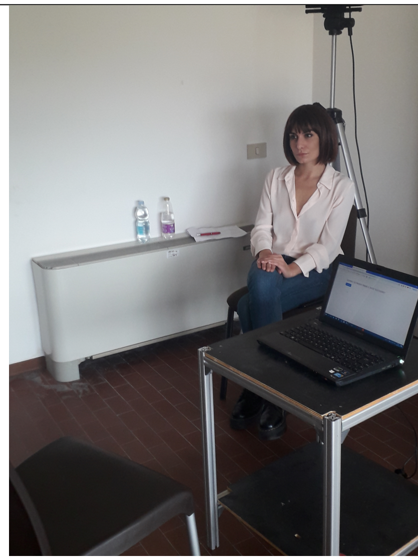
action with humans. To this end, we present a series of experimental treatments based on a modified version of a well-known game used in behavioral economics to study trust among humans: the trust game as proposed by Berg and colleagues and adapted by Charness and Dufwenberg[25, 26], In this game, the outcome of the interaction depends on whether the first mover (the trustor) decides or not to trust the second mover (the trustee). If the first mover decides to trust the counterpart by remaining in the game, the second mover has to decide between a choice that does not benefit the trustor but it is more benificial for himself (i.e. provides him with the highest payoff) and a choice that benefits the trustor but provides him with a lower payoff. If the first mover decides not trust, both players get a lower outside payoff. In other words, there is a conflict of interest between the two players when remaing in the game, but both would be better off if a mutual relationship is established (i.e. the first player remains in the game). A peculiar characteristic of this game is that prior to the trustor's choice of remaining in the game, the trustee is given the opportunity to send him a non-binding (i.e. cheap-talk) message. We rely on this game as it has been specifically conceived to assess whether receiving a message containing a promise from the opponent increases individual trust in him (her).

In our experiment the role of the trustor is always played by a (human) experimental subject while the role of the trustee is played by three different types of players: a humanoid robot with high human-likeness (*F*ACE, Fig. 1), a human counter-part (*Human*, Fig. 1) or a computer-box machine (*Computer-Box*, Fig. 1). In all cases, we compare the trustors' choices when the trustee sends a generic message - not including any type of promise (i.e. an '*empty*' message) - with the trustors' choices when the trustee sends instead a message containing a promise. Specifically, to generate the messages from the robot, we rely on real sentences that occurred between human participants in the experiment of Charness and Dufwenberg[25], and were therein classified either as empty or promising. Finally, to monitor the emotional states of our participants, in all sessions we analyzed two of the most widely used automatic nervous system correlates, such as pulse rate variability and electrodermal activity, which are well known to contain information about affective state of a subject.[27]

1 Experimental design

In the experiment we use the trust game proposed by Charness and Dufwenberg [25], which is depicted in Figure (2). There are two players: A (the trustor) and B (the trustee). Player-A chooses between two options, *In* and *Out*. If Player-A chooses *Out*, the game ends and each player wins 5 Euro. If Player-A chooses *In,* then Player-B has to choose between two options, *Roll* or *Don't Roll*. If he chooses *Don't Roll*,

Figure 1: THREE TYPES OF PLAYER-B

FACE	Human	Computer-box
		

then he wins 14 Euro while Player-A earns 0. If he chooses *Roll*, Player-A wins 0 Euro with probability 1/6 and 12 Euro with probability 5/6, while Player-B wins 10 Euro in any case.

From an economic point of view, for Player-B it is better if Player-A chooses *In*, while for Player-A choosing *In* is convenient only if B chooses *Roll*. The main charateristic of this game is that when Player-A wins zero Euro, it is not possible for Player-A to infer with certainty whether Player-B has chosen either *Roll* or *Don't Roll*. This game thus reflects (as many other experiments in economics) real-world situations where it is not possible to perfectly observe the behaviour of a partner that can be delegated to make relevant payoff decisions. In this experiment, the type of Player-B (i.e., the trustee) changes across treatments, while Player-A is always a human participant. In particular, the role of Player-B is played by either a humanoid (FACE), a computer-box or a human. Regarding the message Player-B sends to Player-A, it can be of two kinds:: a message containing a promise to roll the dice *(promising)*, and a generic message *(empty)*. In particular, we select messages from the original study of Charness and Dufwenberg[25] (as available on their Supplementary material in the online Appendix). To further check whether the length of messages affects individual choices, for each type of message (i.e. promising and empty), we specifically select two short (less than 10 seconds) and two long (more than 10 seconds) messages. Thus, we have a 3x2x2 design. Treatments are illustrated in Table (1), and an English translation of the instructions is available at the end of the paper. In FACE treatments, the role of Player-B is played by FACE, i.e. a hyper-realistic humanoid robot with the aesthetic of a woman (see Figure 1)

Figure 2: THE GAME

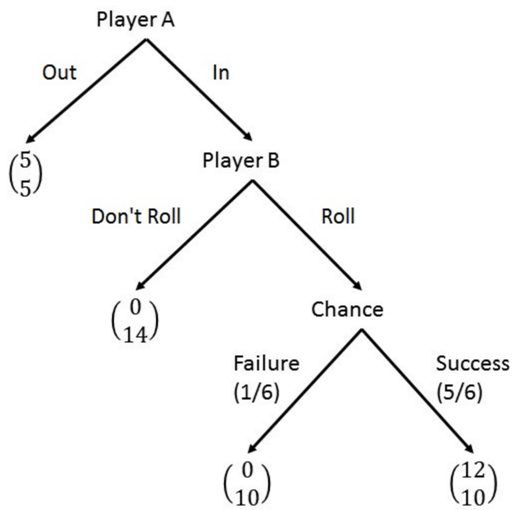


Table 1: TREATMENTS

This table classifies the number of observations collected in our study according to the type of counterpart the human participants confront with (i.e. Computer-box, Human, and Humanoid) and the type of sentence they have to listen to (i.e. cointaing a promise or not, either a short or long sentence).

	Empty			Promising			Grand Total
	Short	Long	Total	Short	Long	Total	
Computer- box	12	19	31	20	13	33	64
Human	16	10	26	14	8	22	48
Humanoid (FACE)	15	10	25	16	9	25	50
Total	43	39	82	50	30	80	162

that due to its perceptive, reasoning and expressive capabilities, constitutes a sophisticated observation platform to study what happens when human and machine establish empathic links ([28]). However, although it has been shown that computer agents can use the expression of emotion to influence human perceptions of trustworthiness, we do not rely on FACE's ability of showing emotional information through facial expressions in order to isolate only the effect of human-likeness and promise in influencing the emotional state of our partecipants, as well as their choices. In the *Computer-Box* treatments, the role of Player-B is played by a light-emitting audio-box reproducing the same audio-sentences and taking decisions in the same way as in FACE . Importantly, both in *FACE* and *Computer-Box* treatments, the artificial agent has its own cognitive system with its perception analysis and architecture, i.e. the so-called Social Emotional Artificial Intelligence (SEAI).[2] This framework allows the social scenario to be acquired and to influence the parameters which correspond to the 'mood' of the artificial agent (see [29]). Specifically, in this experiment, due to SEAI, the artificial agent benefits from its own artificial emotions for choosing whether to *Roll* or *Don't Roll* (see the Appendix for more information about how the robot takes a decision). More importantly, the participants in this experiment are aware that the artificial agent (like the human counterpart) is able to take its decision autonomously, i.e. not randomly but following its own behavioural rules, and therefore the results of game interaction is not determined by chance only.

In the *Human* treatments, the role of Player-B is played by the same professional actress who gave her voice for recording FACE/Computer-Box' audios. The actress is free to autonomously decide her choices in the game, i.e. *Roll* or *Don't Roll*, being paid accordingly, but she has no room to decide which sentences to state that have to be exactly the same ones, and in the same identical order, as the ones pronounced in *FACE* and *Computer-Box*. Moreover, the actress is instructed to avoid any facial expressions during the interaction with a participant, and has to wear FACE's hairs and dresses. Similarly, she has to follow the same experimental procedure as in the Computer-Box and FACE treatments (see the Appendix for details).

To investigate the psychophysiological state of Player-A while taking the decision, in all sessions the participants wear a wearable device on their left wrist (a sensorized bracelet called 'Empatica'[3]) for the real-time collection of physiological data, such as PRV and EDA. XXX The processing of these two signals allows us to characterize the ANS activity of Player-A and infer about his (her) psychophysiological states. In particular, two indexes were computed to quantify the sympathetic nervous system activity (i.e. the EDAsymp index) and the sympthovagal balance (i.e. EDAHFnu index). In Appendix we describe in details how we computed these two indices.

At the end of the experiment, participants has to fill in a questionnaire asking information about how they perceive Player-B, as well as information about their individual characteristics, such as age, gender, and field of studies. In particular, as Nitsch and Glassen,[20] participants has to rate on 7-likert scale how much they perceive Player-B as a human (i.e. the human-likeness, where 0 means non-human at all and 7 means totally human) and how much they perceive Player-B as a machine (i.e. the machine-likeness) . We also ask participants to rate how much they believe their behaviour has affected Player-B's choiceand to make a guess about Player-B's choice (Roll/Don't roll). Finally, we elicite their

[2]The only exception being the actuation control (i.e. commands to induce movement and facial expressions), which is obviously different.

[3]`https://www.empatica.com/`

Table 2: TYPE OF MESSAGES

TYPES	# PHRASES	# SECONDS	PHRASES
Empty	2	<10	- *'Good luck!'* - *'Please choose IN, so we both earn more money.'*
Empty	2	>10	- *'If you stay IN, the chances of the die coming up other than 1 are 5 in 6 – pretty good. Otherwise, should you choose OUT we'd both be stuck at 5 Euro.'* - *'Good luck on your decision. Choose whatever. If you choose "out", you get only 5 Euro more. If you choose "In" you can get 12 Euro instead of only 5 Euro. 7 Euro more is a lot of money!'*
Promising	2	<10	- *'I will roll the dice'* - *'Choose In and I will Roll. You have my word.'*
Promising	2	>10	- *'Choose in, I will roll dice, you are 5/6 likely to get 2,3,4,5, or 6 and win 12 Euro. This way both of us will win something.'* - *'Choose in and I will roll. That way, we'll both get extra money.'*

This table reports 8 sentences that occured between human participants in the study of Charness and Dufwenberg (2006) and were selected in our study. 4 out of 8 sentences were classified as short (i.e. they last less than 10 seconds) and empty (i.e. they did not contain any type of promise to roll the dice).

technological affinity by an ATI scale as in Franke and coauthors[30] and measure their individual risk ambiguity with an INTRA tests (see [31]).

The experiment has been conducted from the end of July till October 2019, and 162 randomly invited participants out of a pool of more than 1500 students coming from all departments of the University of Pisa (91 students were female and 72 male with no substantial difference across treatments).

2 Results

We start analyzing how participants rated the different types of player-B as a human and a machine, as well as their technological affinity. In Table (3) we report the average of these variables by type of Player-B. Note that in the following, we denote with p_p the one-sided p-value for a test for proportions, with $\underline{p_t}$ the one-sided p-value for a t-Student test, and with p_{perm} the one-sided p-value for a test with 500 data permutations). If we compare how much individuals rated Player-B as a human, we observe that *Human* is ranked higher than *Face* (mean diff=1.49, p_t=0.000), and *Face* is ranked higher than *Computer-box* (mean diff=0.87, p_t=0.007). Moreover, if we look at how participants assessed Player-B as a machine, we consistently find that *Face* ranked higher than *Human* (mean diff=2.03, p_t=0.000) and lower than

Table 3: PARTICIPANTS' PERCEPTION AND TECHNOLOGICAL AFFINITY

For each type of player-B, this table reports the average values of variables measuring on a scale from 0 to 7 human-likeness, machine-likeness and tecnological affinity (ATI scale as in [30]).

	Human-likeness	Machine-likeness	ATI
Human	4.96	3.60	4.84
FACE	3.46	5.64	5.08
Computer-Box	2.59	5.93	4.98
Total	3.56	5.15	4.97

Computer-box (although not significantly). It is important to remark that we ask our participants to give the same rating also to the human (actress) counterpart as her behaviour is not entirely natural, as she has to avoid any additional interactions as well as any facial expression during the game. We do not find any significant difference in technological affinity between participants in the different treatments.

The main results are summarized in Table (4), which reports the relative frequencies of choice 'In' made by participants (acting as Player-A) by treatments and human-likeness. Specifically, for each type of Player-B, we categorize the level of human-likeness as Low when the participant rating is in the lower side of the distribution on the 7-likert scale), and High otherwise. Note that we pool the data regarding the length of the message, since it does not significantly affect the decisions to play 'In' in any scenario.

We first compare the results according to the type of Player-B. We note that the frequency of choice 'In' is significantly lower when player-B is a Human than when player B is either FACE (0.60 vs 0.80, mean diff=-0.20, p_p=0.030, $p_{perm} = 0.016$) or a Computer-box (0.77, mean diff=-0.17, p_p=0.066, p_{perm} =0.016). There is no significant difference between FACE and Computer-box.

Regarding the effect of receiving a promise (vs. receiving an empty message), we do not find any significant effect on the frequency of choice 'In' looking at each type of player-B separately. However if we distinguish by human-likeness, we find significant effects of receiving a promise. Specifically, when Player-B is Human and human-likeness is high, the frequency of choice 'In' is significantly higher when a promise is received (0.86 vs 0.53, mean diff=0.33, p_p=0.030, p_{perm} =0.018). A similar, but only weakly significant, effect is found when Player-B is FACE and human-likeness is high (1 vs 0.85, mean diff=0.15, p_p=0.097, p_{perm} =0.000).

We now delve into the effects of human-likeness for each type of Player-B. To begin with, we observe that if participants assigned a high human-likeness to Player-B, the probability of choosing 'In' is significantly higher than those who assigned it a low human-likeness when Player-B is either FACE (0.91 vs 0.70, mean diff=0.21,p_p=0.033, p_{perm} =0.010) or Human (0.69 vs 0.47, mean diff=0.22, p_p=0.067, p_{perm} =0.032). There is no significant difference when Player-B is a Computer-box. Furthermore, if

Table 4: Relative frequencies of ‘choice in’ by treatment and human-likeness

		Human-likeness		Total
		Low	High	
FACE	Empty	0.67 [12]	0.85 [13]	0.76 [25]
	Promising	0.73 [15]	1 [10]	0.84 [25]
	Total	0.70 [27]	0.91 [23]	0.80 [50]
Human	Empty	0.55 [11]	0.53 [15]	0.54 [26]
	Promising	0.37 [8]	0.86 [14]	0.68 [22]
	Total	0.47 [19]	0.69 [29]	0.60 [48]
Computer-Box	Empty	0.71 [21]	0.80 [10]	0.74 [31]
	Promising	0.79 [19]	0.79 [14]	0.79 [33]
	Total	0.75 [40]	0.79 [24]	0.77 [64]

This table reports the relative frequencies of (i.e. the share of participants) choosing ‘IN’ for each treatment by human-likeness. Human-likeness is Low when the participant rating is in the lower side of the distribution on the 7-likert scale, and High otherwise. The number of observations are in squared brackets.

we further distinghuish between the group of participants who received a promise from those who received an empty message, we observe that, when Player-B is FACE, the effect of higher human-likeness is significant only among those who received a promise (1 vs 0.73, mean diff = 0.27, p_p=0.037, p_{perm} =0.000). A similar result is observed when Player-B is Human (0.86 vs 0.37, mean diff= 0.49, p_p=0.010, p_{perm} =0.002). Overall, we can conclude that the choice to trust FACE is significantly related to the way a participant perceived it as a human. If a participant recognises FACE very similar to a human being, the probability that he will choose ‘In’ increases. We find that this effect is mainly driven by those participants who received a promise.

If we attend to the emotional reaction of the participants (using the two indices EDAsymp and EDAHFnu computed by the physiological data recorded during the experiment, see Tab 5), we find a significantly higher reaction when Player-B is FACE that when Player-B is either Computer-box (0.724 vs -0.211, mean diff$_{EDAsymp}$ = 0.935, p_t=0.016, p_{perm}=0.008; 2.837 vs -0.107, mean diff$_{EDAHFnu}$ =2.944, p_t=0.053, p_{perm}=0.050) or Human (0.724 vs -0.186, mean diff$_{EDAsymp}$ = 0.909, p_t=0.056; , p_{perm}=0.074; 2.837 vs 0.747, mean diff$_{EDAHFnu}$ =3.584, p_t=0.063, p_{perm}=0.068). Furthermore, when Player-B is FACE, we find that subjects who rated Player-B high in human-likeness are more likely to experience a stronger emotional reaction than participants who rated it low (1.731 vs -0.129, mean diff$_{EDAsymp}$=-1.859, p_t=0.017

Table 5: PHYSIOLOGICAL DATA: EDASYMP AND EDAHF_NU

Index	Human-likeness	*Box*	*Human*	*FACE*
EDASymp	LOW	-0.144	-0.288	-0.129
		[28]	[9]	[26]
	HIGH	-0.327	-0.128	1.731
		[16]	[16]	[22]
	Total	-0.211	-0.186	0.724
EDAHFnu	LOW	-0.175	-2.173	0.275
		[28]	[9]	[26]
	HIGH	0.012	0.055	5.865
		[16]	[16]	[22]
	TOTAL	-0.107	-0.747	2.837

The EDAsymp index quantifies the activity of the sympathetic nervous system, while the EDAHFnu index quantifies the sympthovagal balance. A full description is available in the Appendix. Human-likeness is Low when the participant rating is in the lower side of the distribution on the 7-likert scale, and High otherwise. The number of observations are in squared brackets.

, p_{perm} =0.000; 5.865 vs 0.275 $_{EDAHFnu}$=-5.590, p_t=0.009 , p_{perm} =0.000;). We do not find a similar effect when Player-B is Human or Computer-box. Finally, we note that the psychophysiological reaction of subjects rating FACE high in human-likeness is significantly higher than that experienced by subjects interacting either with Computer-box or Human, regardless of the rating of human-likeness. Regarding the relationship between the emotional reaction of participants and their choices, we do not find any significant correlation using the two indices EDAsymp and EDAHFnu. However, if we split our participants into two groups according to whether they express a stronger (or weaker) psychophysiological reaction that the median level of the distribution of EDAsymp, we can observe that those who experienced a stronger reaction are also less likely to choose IN in both Computer (0.636 vs 0.909, mean diff=0.273, and p_p=0.015) and Human (0.462 vs 0.750, diff=0.288, and p_p=0.070 see Table 6).

Finally to study the interaction between human-likeness and psychophysiological reaction of our participants we conduct a probit analysis for the probability of playing 'In' using as a set of regressors player human-likeness and EDAsymp dummy, along with a dummy for each treatments. Results are report in Figure (3). As this figure highlights increasing the psychophysiological reaction from a low one to a high one reduces the probability of playing 'IN'. However, increasing the level of human-likeness counterbalance this negative effect, especially in Face and in Computer-box.

3 Discussion and conclusion

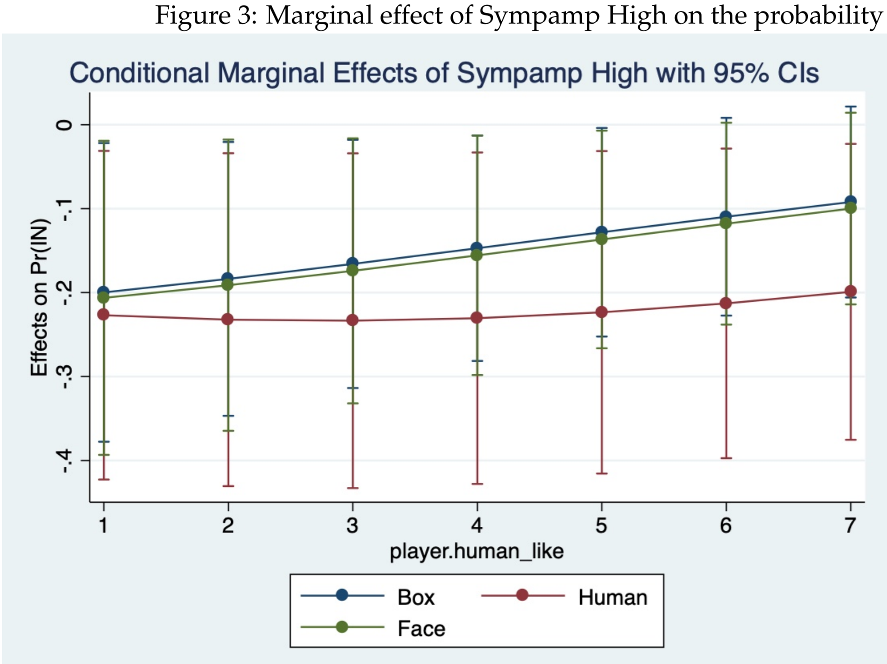
In our experiment participants were confronted with a counterpart which differed in the degree of human-likeness: a light-emitting computer-box, a female humanoid and a human female (which resembled the humanoid). The participants needed to decide - after listening to a message from the coun-

Table 6: RELATIVE FREQUENCIES OF 'CHOICE IN'BY PHYSIOLOGICAL STATE AND HUMAN-LIKENESS

	Human-likeness	EDASymp		Total
		High	Low	
FACE	High	0.916 [12]	0.900 [10]	0.909 [22]
	Low	0.667 [12]	0.714 [14]	0.692 [26]
	Total	0.792 [24]	0.792 [24]	0.792 [48]
Computer-Box	High	0.667 [7]	0.857 [9]	0.750 [16]
	Low	0.616 [15]	0.933 [13]	0.786 [28]
	Total	0.636 [22]	0.909 [22]	0.770 [44]
Human	High	0.500 [8]	0.875 [8]	0.686 [16]
	Low	0.400 [5]	0.500 [4]	0.444 [9]
	Total	0.462 [13]	0.750 [12]	0.600 [25]

Each cell represents the frequencies of choice 'In' within each category. An individual is classified in EDAsymp High whenever is above the median level of the EDAsymp distribution, and EDAsymp Low otherwise. Human-likeness is Low when the participant rating is in the lower side of the distribution on the 7-likert scale, and High otherwise. The number of observations are in squared brackets.

Figure 3: Marginal effect of Sympamp High on the probability of playing 'In'



terpart, containing in half of the cases a promise - whether to trust or not their opponent in the game. We find evidence that a human receveing a promise from a humanoid has more trust in it only when he (or she) perceived the artificial agent very similar to a human-being. Indeed, if we replace the social robot by a human we find a similar pattern. However, replacing it by the computer -box the effect of receiving a promise disappears. We also find that participants experienced a stronger psychophysiological reaction when confronted with a humanoid, especially if it appears to them very close to human. Moreover, we observe that those participants expressing stronger psychophysiological reaction are less likely to trust the counterpart when this is either a computer-box or a human (i.e. choose more often the safer option).

Taken all together, these results suggest that human-likeness and (integral) emotions play both an important role in the decision to trust the counterpart, possibly in interaction with each other. However, some remarks are in order. While in this experiment we can fully control for the degree of human-likeness by varying it across treatments, we have less control of the type of emotions experienced by our subjects. Although physiological measures such electrodermal activity (EDA) have been used over 100 years for representing emotional arousal, and most scholars accept a physiological component in the definition of emotions, it is not possible to directly match the physiological state of a participant with a direct type of emotion (e.g. fear or anxiety). In addition, as the literature on emotion arousal highlights there might be individuals exhibiting different physiological responses to the same emotional state[32]. Therefore, our results can only suggest a greater or a weaker 'emotional arousal' without giving any insights on the type of emotions proved by our participants.

Nevertheless, the vast psychological literature on emotions and decision-making offers us an interesting framework to interpret our results. In particular, recent evidence from laboratory experiments is mostly consistent with the Appraisal-Tendency Framework according to which emotions change individuals' appraisal of a situation, thereby affecting individual choices[9, 33]. Importantly, in that framing, emotions of the same valence (such as fear and anger) can exert opposing influences on choices. Thus, what matters is whether an emotion (either positive or negative, strong or weak) by leading to a more cautious appraisal of the situation reduces the feeling of control, e.g. thereby reducing the willingness to take risks. Therefore, even if we are not able to disentangle among different types of emotions, we can reasonably assert that in our framework, whenever the experience of a stronger emotional arousal lead a participant to a more cautious appraisal of the counterpart, we observe a more careful assessment of the situation and a lower willingness to take risk and trust the counterpart. This interpretation of our results is also consistent with previous research showing that participants with ventromedial prefrontal cortex (a key area of the brain for integrating and integrating emotion and cognition) repeatedly select a riskier financial option over a safer one, even to the point of bankruptcy, despite their understanding of the suboptimality of their choices. In particolar, their psysiological measure of skin response suggests that they did not experience the emotional signals (i.e. the somatic markers) that lead normal decision makers to fear high risks[9].

Overall, these results strongly support the efforts in developing technologies enhancing the humanity of social robots, both in terms of human appearance and communication behaviour. Indeed, if from one-side it is not possible to control for human emotions, our results - in line with recent studies [21, 22] - suggest that increasing the human-likeness of an artificial agent increases sensibly the likelihood that a human counterpart will trust it. At the same time, the analysis we conducted opens an interesting question about the role of specific emotions, also over the longer time-horizons, that we are not able to fully disentangle in our simple one shot-game.

To conclude, we see several directions for future interdisciplinary research. The first one is to explore different types of human-robot interactions, for example prisoner dilemma games, coordination games or repeated interactions (e.g. by replicating the analysis of Crandall and co-authors with a social robot [19]). The second direction of research is on the side of the social robot. It would be very interesting to introduce - within standard experiments in economics - the behavior of people interacting with a robot that can also additionally adapt its facial expression, as well as the mode of communication, to the perceived emotions of the human counterpart.

References

[1] Lange, P. A. M. V. (2015) Generalized Trust: Four Lessons From Genetics and Culture. Current Directions in Psychological Science, **24**(1), 71–76. (document)

[2] Fehr, E. (2009) On the economics and biology of trust. Journal of the european economic association, **7**(2-3), 235–266. (document)

[3] Langevoort, D. C. (1996) Selling hope, selling risk: some lessons for law from behavioral economics about stockbrokers and sophisticated customers. Cal L. Rev., **84**, 627. (document)

[4] Nishio, S., Ogawa, K., Kanakogi, Y., Itakura, S., and Ishiguro, H. (2018) Do robot appearance and speech affect people,Äôs attitude? Evaluation through the ultimatum game. Geminoid Studies: Science and Technologies for Humanlike Teleoperated Androids, pp. 263–277. (document)

[5] Picard, R. W. (2004) Toward Machines with Emotional Intelligence.. In ICINCO (Invited Speakers) Citeseer pp. 29–30. (document)

[6] Engelmann, J. B., Meyer, F., Ruff, C. C., and Fehr, E. (2018) The neural circuitry of emotion-induced distortions of trust. BioRxiv, p. 129130. (document)

[7] Schniter, E., Shields, T. W., and Sznycer, D. (2018) Trust in Humans and Robots: Economically Similar but Emotionally Different. (document)

[8] Jung, E.-S., Dong, S.-Y., and Lee, S.-Y. (2019) Neural Correlates of Variations in Human Trust in Human-like Machines during Non-reciprocal Interactions. Scientific reports, **9**(1), 1–10. (document)

[9] Lerner, J. S., Li, Y., Valdesolo, P., and Kassam, K. S. (2015) Emotion and decision making. Annual review of psychology, **66**. (document), 1, 3

[10] Damasio, A. R. (1996) The somatic marker hypothesis and the possible functions of the prefrontal cortex. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, **351**(1346), 1413–1420. (document)

[11] Damasio, A. R. (2001) Descartes error revisited. Journal of the History of the Neurosciences, **10**(2), 192–194. (document)

[12] Vaa, T. (2001) Driver behavior models and monitoring of risk: Damasio and the role of emotions. In International Conference: Traffic Safety on Three ContinentsPTRC Education and Research Services Limited Number VTI Konferens 18A. (document)

[13] Fox, A. S., Lapate, R. C., Shackman, A. J., and Davidson, R. J. (2018) The nature of emotion: fundamental questions, Oxford University Press, . (document)

[14] Arkin, R. C., Ulam, P., and Wagner, A. R. (2011) Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. Proceedings of the IEEE, **100**(3), 571–589. (document)

[15] Tortosa, M. I., Strizhko, T., Capizzi, M., and Ruz, M. (2013) Interpersonal effects of emotion in a multi-round Trust Game.. Psicologica: International Journal of Methodology and Experimental Psychology, **34**(2), 179–198. (document)

[16] Campellone, T. R. and Kring, A. M. (2013) Who do you trust? The impact of facial emotion and behaviour on decision making. Cognition & emotion, **27**(4), 603–620. (document)

[17] Engelmann, J. B., Hare, T. A., Fox, A. S., Lapate, R. C., Shackman, A. J., and Davidson, R. J. (2018) Emotions can bias decision-making processes by promoting specific behavioral tendencies. (document)

[18] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., and Parasuraman, R. (2011) A meta-analysis of factors affecting trust in human-robot interaction. Human factors, **53**(5), 517–527. (document)

[19] Crandall, J. W., Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., Rahwan, I., et al. (2018) Cooperating with machines. Nature communications, **9**(1), 1–12. (document), 3

[20] Nitsch, V. and Glassen, T. (2015) Investigating the effects of robot behavior and attitude towards technology on social human-robot interactions. In 2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) IEEE pp. 535–540. (document), 1

[21] Waytz, A., Heafner, J., and Epley, N. (2014) The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. Journal of Experimental Social Psychology, **52**, 113–117. (document), 3

[22] Nass, C. and Moon, Y. (2000) Machines and mindlessness: Social responses to computers. Journal of social issues, **56**(1), 81–103. (document), 3

[23] March, C. (2019) The behavioral economics of artificial intelligence: Lessons from experiments with computer players. (document)

[24] Mori, M. (2017) The uncanny valley: The original essay by Masahiro Mori. IEEE Robots &,. (document)

[25] Charness, G. and Dufwenberg, M. (2006) Promises and partnership. Econometrica, **74**(6), 1579–1601. (document), 1

[26] Berg, J., Dickhaut, J., and McCabe, K. (1995) Trust, reciprocity, and social history. Games and economic behavior, **10**(1), 122–142. (document)

[27] Tao, J. and Tan, T. (2005) Affective computing: A review. In International Conference on Affective computing and intelligent interaction Springer pp. 981–995. (document)

[28] Mazzei, D., Billeci, L., Armato, A., Lazzeri, N., Cisternino, A., Pioggia, G., Igliozzi, R., Muratori, F., Ahluwalia, A., and De Rossi, D. (2010) The face of autism. In 19th International Symposium in Robot and Human Interactive Communication IEEE pp. 791–796. 1

[29] Cominelli, L., Mazzei, D., and De Rossi, D. E. (2018) SEAI: Social emotional artificial intelligence based on Damasio,Äôs theory of mind. Frontiers in Robotics and AI, **5**, 6. 1, 4.1

[30] Franke, T., Attig, C., and Wessel, D. (2019) A personal resource for technology interaction: development and validation of the affinity for technology interaction (ATI) scale. International Journal of Human–Computer Interaction, **35**(6), 456–467. 1, 3

[31] Rieger, M. O., Wang, M., and Hens, T. (2015) Risk preferences around the world. Management Science, **61**(3), 637–648. 1

[32] Picard, R. W. (2000) Affective computing, MIT press, . 3

[33] Meier, A. N., Emotions, risk attitudes, and patience. Technical report, SOEPpapers on Multidisciplinary Panel Data Research (2019). 3

[34] Lazzeri, N., Mazzei, D., Cominelli, L., Cisternino, A., and De Rossi, D. E. (2018) Designing the mind of a social robot. Applied Sciences, **8**(2), 302. 4.1

[35] Bosse, T., Jonker, C. M., and Treur, J. (2008) Formalisation of Damasio,Äôs theory of emotion, feeling and core consciousness. Consciousness and cognition, **17**(1), 94–113. 4.1

[36] Zaraki, A., Pieroni, M., De Rossi, D., Mazzei, D., Garofalo, R., Cominelli, L., and Dehkordi, M. B. (2016) Design and evaluation of a unique social perception system for human–robot interaction. IEEE Transactions on Cognitive and Developmental Systems, **9**(4), 341–355. 4.1

[37] Cominelli, L., Mazzei, D., Carbonaro, N., Garofalo, R., Zaraki, A., Tognetti, A., and De Rossi, D. (2016) A Preliminary Framework for a Social Robot ,ÄúSixth Sense,Äù. In Conference on Biomimetic and Biohybrid Systems Springer pp. 58–70. 4.1

[38] Mazzei, D., Cominelli, L., Lazzeri, N., Zaraki, A., and De Rossi, D. (2014) I-clips brain: A hybrid cognitive system for social robots. In Conference on Biomimetic and Biohybrid Systems Springer, Cham pp. 213–224. 4.1

[39] Giarratano, J. C. and Riley, G. (1998) Expert systems, PWS publishing co., . 4.1

[40] Russell, J. A. (1980) A circumplex model of affect.. Journal of personality and social psychology, **39**(6), 1161. 4.1

[41] Cominelli, L., Mazzei, D., Pieroni, M., Zaraki, A., Garofalo, R., and De Rossi, D. (2015) Damasio,Äôs somatic marker for social robotics: preliminary implementation and test. In Conference on Biomimetic and Biohybrid Systems Springer pp. 316–328. 4.1

[42] Mazzei, D., Lazzeri, N., Hanson, D., and De Rossi, D. (2012) Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars. In 2012 4th IEEE RAS & EMBS International Conference on biomedical robotics and biomechatronics (BioRob) IEEE pp. 195–200. 4.1

[43] Kreibig, S. D. (2010) Autonomic nervous system activity in emotion: A review. Biological psychology, **84**(3), 394–421. 4.4

[44] Vernet-Maury, E., Deschaumes-Molinaro, C., Delhomme, G., and Dittmar, A. (1993) Autonomic nervous system activity and mental workload. International Journal of Psychophysiology, **14**(2), 153–154. 4.4

[45] Greco, A., Valenza, G., Bicchi, A., Bianchi, M., and Scilingo, E. P. (2019) Assessment of muscle fatigue during isometric contraction using autonomic nervous system correlates. Biomedical Signal Processing and Control, **51**, 42–49. 4.4

[46] Greco, A., Valenza, G., and Scilingo, E. P. (2016) Advances in Electrodermal activity processing with applications for mental health, Springer, . 4.4.1

[47] Greco, A., Valenza, G., Lanata, A., Scilingo, E. P., and Citi, L. (2015) cvxEDA: A convex optimization approach to electrodermal activity processing. IEEE Transactions on Biomedical Engineering, **63**(4), 797–804. 4.4.1

[48] Posada-Quintero, H. F., Florian, J. P., Orjuela-Cañón, A. D., Aljama-Corrales, T., Charleston-Villalobos, S., and Chon, K. H. (2016) Power spectral density analysis of electrodermal activity for sympathetic function assessment. Annals of biomedical engineering, **44**(10), 3124–3135. 4.4.1

[49] Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., and Suri, J. S. (2006) Heart rate variability: a review. Medical and biological engineering and computing, **44**(12), 1031–1051. 4.4.2

[50] Takens, F. (1981) Detecting strange attractors in turbulence. In Dynamical systems and turbulence, Warwick 1980 pp. 366–381 Springer. 4.4.2

[51] Chen, W., Wang, Z., Xie, H., and Yu, W. (2007) Characterization of surface EMG signal based on fuzzy entropy. IEEE Transactions on neural systems and rehabilitation engineering, **15**(2), 266–272. 4.4.2

[52] Azami, H., Rostaghi, M., Abásolo, D., and Escudero, J. (2017) Refined composite multiscale dispersion entropy and its application to biomedical signals. IEEE Transactions on Biomedical Engineering, **64**(12), 2872–2879. 4.4.2

[53] Nardelli, M., Scilingo, E. P., and Valenza, G. (2019) Multichannel Complexity Index (MCI) for a multi-organ physiological complexity assessment. Physica A: Statistical Mechanics and its Applications, **530**, 121543. 4.4.2

[54] Li, P., Liu, C., Li, K., Zheng, D., Liu, C., and Hou, Y. (2015) Assessing the complexity of short-term heartbeat interval series by distribution entropy. Medical & biological engineering & computing, **53**(1), 77–87. 4.4.2

[55] Karmakar, C., Udhayakumar, R. K., and Palaniswami, M. (2015) Distribution entropy (disten): a complexity measure to detect arrhythmia from short length rr interval time series. In 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) IEEE pp. 5207–5210. 4.4.2

[56] Nardelli, M., Greco, A., Danzi, O. P., Perlini, C., Tedeschi, F., Scilingo, E. P., Del Piccolo, L., and Valenza, G. (2019) Cardiovascular assessment of supportive doctor-patient communication using multi-scale and multi-lag analysis of heartbeat dynamics. Medical & biological engineering & computing, **57**(1), 123–134. 4.4.2

[57] Tulppo, M. P., Makikallio, T., Takala, T., Seppanen, T., and Huikuri, H. V. (1996) Quantitative beat-to-beat analysis of heart rate dynamics during exercise. American journal of physiology-heart and circulatory physiology, **271**(1), H244–H252. 4.4.2

[58] Nardelli, M., Greco, A., Bolea, J., Valenza, G., Scilingo, E. P., and Bailón, R. (2017) Reliability of lagged poincaré plot parameters in ultrashort heart rate variability series: Application on affective sounds. IEEE journal of biomedical and health informatics, **22**(3), 741–749. 4.4.2

[59] Strigo, I. A. and Craig, A. D. (2016) Interoception, homeostatic emotions and sympathovagal balance. Philosophical Transactions of the Royal Society B: Biological Sciences, **371**(1708), 20160010. 4.4.3

[60] Sleight, P. and Bernardi, L. (1998) Sympathovagal balance. Circulation, **98**(23), 2640–2640. 4.4.3

[61] Ghiasi, S., Greco, A., Nardelli, M., Catrambonel, V., Barbieri, R., Scilingo, E. P., and Valenza, G. (2018) A New Sympathovagal Balance Index from Electrodermal Activity and Instantaneous Vagal Dynamics: A Preliminary Cold Pressor Study. In 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) IEEE pp. 3068–3071. 4.4.3

4 Methods

4.1 The FACE Robot and the SEAI Cognitive System

The FACE robot (Facial Automaton for Conveying Emotions) is a humanoid with hyper-realistic adult female aesthetics, specifically designed for social robotics [34]. It is composed with a passive body on the top of which a Hanson Robotics' head has been mounted. The head is designed to host 32 servomotors that guide the neck of the robot, its eyes, mouth, and facial expression. The face of the ginoid is made of Frubber[4], a registered material with skin-like mechanical and aesthetical features. This hardware is controlled by SEAI (Social Emotional Artificial Intelligence), a distributed control architecture made of perception, cognitive and actuation systems, that endow the robot with expressive and communicative capabilities [29], including also the possibility to emulate verbal communication following prerecorded audio files[5]. SEAI is a bio-inspired architecture based on neuroscientific theories of mind. In particular, it has been inspired by the findings of Antonio Damasio and it is consistent with the computational

[4] `https://patents.google.com/patent/US7113848?oq=frubber`

[5] The audio files used for the experiment have been recorded using the voice of a professional actress, the same who interpreted the role of Player-B in the interactions with the real person; the sentences were the Italian translation of the sentences between the Charness trust game players.

formalization made by [35]. In its development, the influence of emotions in the decision-making process has been of primary importance. The perception part of the system is the Scene Analyzer, an audiovisual perception system conceived to analyze a social environment using the robot sensors and to extract meaningful social cues from these available data. Features that can be extracted from a human interlocutor are, e.g., the three dimensional postition of 25 joint coordinates, their speaking probability, meaningful postures and gestures, estimated facial expressions, age and gender [36]. This Social Perception System has already been successfully integrated with the acquisition of physiological parameters (i.e., elctrodermal activity, respiration rate and heart rate variability) in past experiments (see [37]). All the environmental information anlayzed by the perception system of the robot is then processed by its cognitive system, i.e., the I-CLIPS Brain [38], a rule-based expert system written in CLIPS language [39]. The knowledge base of the expert system is written by means of IF-THIS-THEN-THAT rules, where each rule contains a set of actions that will be executed if several conditions about the upcoming factual information are satisfied. Thanks to these rules it is possible to design the behavior of the humanoid. For example, a particular expression gathered in its intelocutor can lead to the trigger of a sentence or a facial expression performed by the robot, but also to the modification of the robot's internal values. In fact, SEAI includes emotional internal values (i.e., valence and arousal), which combination describes an emotional state, here defined as *mood*. This method of representing emotion is based on the well-known Russell's Circumplex Model of Affect [40]. In the case of the robot, mood is not necessarily externalised by perceivable movements, rather it is implied in biasing the chaining of the rules, and so, the decision tree of the robot. Emotion biasing decision in this cognitive system has been previously tested (see [41]). The instructions coming from the cognitive block about the emotion to be expressed through facial expression (v,a values), the sentence to say, and the point to look at, are merged and continuosly executed thanks to the actuation system, which translate them in movements performed by the motors that drive the face, the mouth and the neck of the ginoid [42]. Furthermore, the SEAI architecure is completely modular and portable, all the blocks composing the framework are stand-alone applications that process a limited set of information. These modules are distributed in a local net of computers that communicate by means of the YARP middleware [6]. This implies that each module can be activated or deactivated, and that the perception and cognitive systems can be used also without controlling the FACE Robot. As a result, we were able to use exactly the same rules engine in the computer box case, simply disabling the actuation part of the system that control the robot, and using instead the bluetooth speaker, presented as a smart computer box, actually running the same perception and actuation system

[6] https://www.yarp.it/

of the robot. This led to a very close and controlled comparison.

4.2 How the robot takes a decision, the Rules Engine

In this experiment, the robot (as well as the computer box) decides whether to *Roll* or *Don't Roll* according to its emotional state and following its decision rules. In particular, a positive mood in SEAI (i.e., an emotional state with positive valence) will lead the robot to be collaborative with the human player and play *Roll*; while a negative mood in SEAI (i.e., an emotional state with negative valence) will lead the robot to play *Don't Roll* (see Figure 5). The decision is taken at the end of the interaction with Player-A, when the subject goes out of the room, and so out of the field of view of the robot.

If in the moment in which the robot has to take a decision, it is in a qualitatively neutral mood (v=0, regardless the arousal), the decision will be taken randomly (50%). Participants' behavior during all the time spent alone in the room with the robot, once observed by the Scene Analyzer and processed in SEAI, act as an input modifying the parameters of the robot which correspond to its 'mood', thus in turn affecting its course of action (i.e., its final decision). However, in this experiment, at each interaction with a new participant the robot always resetted its internal values at the «neutral emotional state» (which corresponds to $v = 0, a = 0$ in the graph). In conclusion, thanks to SEAI the robot was completely autonomous, by means of the rules everything was pre-programmed and automatized, starting from the rules that use perceived social cues to modulate the emotional state of the robot, to other rules determining which sentence it has to say, when to start and to end a treatment, and the storage of all the data acquired with timestamps in a structured dataset. The complete code of the rules engine is available in appendix A.

4.3 Experimental procedure

Each participant arrives in the laboratory and enter a room in which (s)he read and sign the consent to participate in the study. The participantthen sits in front of a computer screen where (s)he can read autonomously the experiment instructions and fill in some preliminary questions about their attitudes towards the technology. Once the time dedicated to this part has expired, the participant is lead by the experimenter to another room where the robot is located. The participant seats on chair, always located at the same distance from the robot, and when is ready to start the experiment has to rise his hand. At this point, the robot welcomes the participant with a standard sentence ('Nice to meet you! Let's start') to then state one random sentence out of 8 (according to the treatment, see again Table 1). The robot then tell the participant a standard final sentence, inviting the participant to enter his(her) choice in the computer in front of the participant. The robot cannot observe though the choice the participant has made. To conclude the experiment, the participant has to return to the initial room, to complete an exit questionnaire about the interaction of the robot, and receive the final payment.

4.4 Description and analysis of Physio data

Pulse rate variability (PRV) and electrodermal activity (EDA) signals are directly modulated by the autonomic nervous system (ANS) activity and, therefore, are considered ideal non-invasive physiological signals to investigate the ANS dynamics. Indeed, the ANS plays a crucial role in the processing of the emotional response, mental fatigue and workload [43, 44, 45]-

4.4.1 EDA processing

The EDA signal measures the activity of eccrine sweat glands on the hand surface. Since sweat glands are directly innervated by the sympathetic branch of the ANS (and in particular the sudomotor nerve), the EDA analysis is considered one of the best ways to monitor the sympathetic activity. EDA is considered as the superposition of two main components, phasic and tonic, which differ for their time scales and relationships with the external stimuli [46]. In this study, we adopted the well-known cvxEDA model [47] to decompose the EDA signal and extract informative and effective features form both the phasic and tonic signals.

Specifically, EDA algorithm based on Bayesian estimation and convex optimization provides a decomposition of the EDA robust to noise, and enables the estimation of the neural bursts of the sudomotor nerve activity (SMNA), providing a window on the sympathetic nerve activity.

After the application of the cvxEDA model, we extracted some features in order to quantify the activity of the sympathetic nervous system. Particularly, we calculated the frequency of the SMNA peaks and the sum of all amplitudes within each window (EDA_AmpSum), whereas, from the slow-varying tonic component, we computed the mean value (MeanTonic). Moreover, we estimated the power spectrum within the frequency range of 0.045 and 0.25Hz (EDAsymp), which has been demonstrated to be strongly correlated to the sympathetic nervous system activity [48].

4.4.2 ECG processing

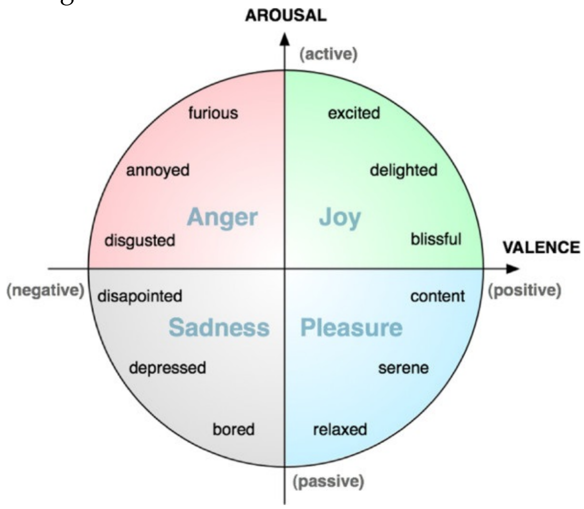
The interbeat interval series (IBI) (were acquired throughout the entire experiment for each participant. Two sessions of twenty seconds of movement-artifact-free IBI series were extracted from each recording: the first localized during the experiment instruction reading, and the second during the period when the participant was in front of the robot/actress/box.

A total amount of eighteen features was extracted from IBI series, in the time and frequency domains [49], and applying nonlinear methods taken from the phase space reconstruction theory [50]. Considering the time-domain, the following four features were calculated from each IBI series lasting twenty seconds [49]: the mean value of IBI segments and their standard deviation (IBI mean and IBI std), the root mean square of successive IBI interval differences (RMSSD), and the relative number of successive IBI sample pairs that differ more than 50 msec, expressed as a percentage of the total number of IBIs (pNN50). PRV signals were computed from IBI series using a sampling frequency of 4 Hz.

Frequency domain analysis consisted in the extraction of eight features from the Power Spectral Density (PSD) related to each PRV signal [49]. Two main spectral bands were considered: low frequency (LF) band (ranging between 0.04 and 0.15 Hz), and high frequency (HF) band (from 0.15 to 0.4 Hz). The following features were calculated: the power values in LF and HF band (LF power and HF power), the power in LF band and HF band normalized to the sum of LF and HF power (LF nu and HF nu), the power in LF band and HF band expressed as percentage of the total power (LF % and HF%), and the ratio between LF power and HF power (LF/HF).

Two entropy algorithms were implemented by using the IBI series, i.e., Fuzzy entropy (FuzzyEn) [51, 52, 53] and Distribution entropy (DistEn) [54, 55, 56]. The first was used to investigate the irregularity of IBI series and the second to quantify spatial complexity of the related attractors in the phase space. Furthermore, five features were extracted to quantify the shape of Poincaré map obtained plotting the lagged IBU interval series, IBI_{n+1}, against the series IBI_n. Three geometrical quantifiers were calculated, according to the ellipse-fitting technique [57, 58]: the standard deviation of the points calculated along the direction perpendicular to the line-of-identity $IBI_{n+1} = IBI_n$ (SD1), the standard deviation of the points along the line-of-identity $IBI_{n+1} = IBI_n$ (SD2), the ratio between SD1 and SD2 (SD12). Other two Poincaré Plot quantifiers were used to minimize the loss of information by accounting also for the points lying outside the ellipse: the mean (M_d) and the standard deviation (S_d) of the euclidean distances calculated between each Poincaré Plot point and the centroid [56].

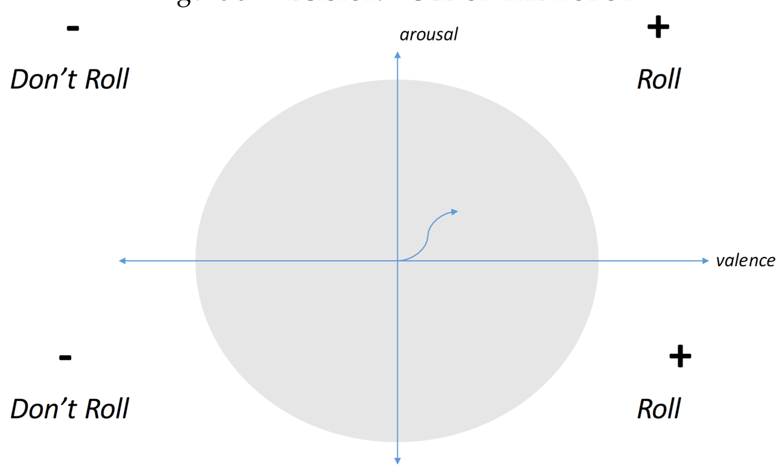
Figure 4: EMOTIONAL STATE OF THE ROBOT



4.4.3 New index from the sympathovagal assessment

Emotions regulation process modulates the sympathovagal balance [59, 60], which is considered a reliable marker of the human affective state. Previous studies have suggested that LF power spectrum can provide a quantitative marker of the sympathetic outflow and have used the LF/HF ratio as a correlate of the sympathovagal balance. However, the LF power is now regarded as a measure of both sympathetic and vagal tone, leading to ambiguities and possible inconsistent conclusions on the use of the LF/HF ratio as sympathovagal marker. In this study, we employed novel indexes of the sympathovagal dynamics based on the combination of the information extracted from the EDA and PRV signal [61]. Indeed, while EDAsymp reliably characterizes the sympathetic activity, there are several cardiovascular features in the time, frequency and nonlinear that reliably quantify the parasympathetic outflow: HF, HFnu, RMSSD, HF%, and SD1. Accordingly, we have estimated the sympathovagal activity combining the EDAsymp with each of the features characterizing the parasympathetic activity building five sympathovagal markers: EDAsymp/HF [61], EDAsymp/HFnu, EDAsymp/RMSSD, EDAsymp/HF%, and EDAsymp/SD1.

Figure 5: DECISION RULE OF THE ROBOT



INSTRUCTIONS: English translation from Italian

Welcome! This experiment will last about 30 minutes. You will receive 5 Euro for your participation. Based upon the choices you will take in the experiment; you can earn additional money. We now ask you to turn off your mobile phone and to read the instructions carefully.

The aim of this experiment is to study how people take decisions. In particular, this experiment wants to study how people take decision when interacting with a human-like robot.

Should you have any doubt, please do not hesitate to ask clarifications to the experimenter.

The data related to this experiment will be saved and analyzed anonymously. No video will be recorded.

In this experiment you will play with FACE i.e. a social robot which is able to prove and express its emotions. [with a computer-box which is given a system of social perception]. FACE [The Computer box] is also able to take its decisions autonomously, following its own behavioral rules. In this game, FACE [The Computer box] is programmed to choose autonomously between two actions: **ROLL** and **DON'T ROLL** a six-faces dice.

[In this experiment you will play with Deborah. Deborah can choose autonomously between two actions: ROLL and DON'T ROLL a six-faces dice.]

YOUR CHOICE

You will have to choose between two options: whether to play **IN** or **OUT**.

Should you choose **OUT,** both you and FACE [Computer box] [Deborah] will

earn 5 Euro each.

Should you choose **IN, FACE** [Computer box] [Deborah] can then choose between the two options: **ROLL** and **DON'T ROLL** the six-faces dice. In the event FACE [Computer box] [Deborah] choosing **DON'T ROLL,** you will receive 0 Euro and FACE [Computer box] [Deborah] will earn 14 Euro. In the event FACE [Computer box] [Deborah] choosing **ROLL,** FACE [Computer box] [Deborah] will always earn 10 Euro

561 while you earning depends on the results of dice roll. If the result of the dice roll is a number between 2 and 6 you will earn 12 Euro, otherwise if the result of the dice roll is the number 1 you will receive 0 Euro.

It is important to notice that FACE [Computer box] [Deborah] will not know whether you opted either IN or OUT when has to reach a decision. It is also important to notice that the money earned by FACE will remain to FACE itself [will remain to the lab (e.g. maintenance)], and used for its necessity (e.g. maintenance)

The payments are summarized in the table below.

	Dice roll	**You earning**	**FACE's** [Computer box] [Deborah] **earning**
If you choose **OUT**	-	5 Euro	5 Euro
If you choose **IN** FACE choose **DON'T ROLL**	-	0 Euro	14 Euro
If you choose **IN** FACE choose **ROLL**	Result: 1	0 Euro	10 Euro
	Results: 2,3,4,5,6	12 Euro	10 Euro

Now you have 5 minutes to read these instructions alone and ask clarifications questions to the experiment. Once you have finished reading, the experiment will bring you to another room where FACE [Computer box] [Deborah] is. You will have to seat on the chair in front of face, and in order to begin the experiment you need to raise your right hand. At the point, you will hear a message from FACE [Computer box] [Deborah]. You will then enter your choice in the computer close to you.

Once you have done, we will wait for you to come back again to this room, to fill in a final questionnaire and receive your payment.