Modelling and Measuring Trust in Human-Robot Collaboration

Erlantz Loizaga *, Leire Bastida, Sara Sillaurren, Ana Moya, Nerea Toledo

Posted Date: 26 December 2023

doi: 10.20944/preprints202312.1915.v1

Keywords: Human-Robot Collaboration (HRC); trust dimensions; trust dynamics; experimental process

Article

Modelling and Measuring Trust in Human-Robot Collaboration

Erlantz Loizaga [1,*], Leire Bastida [1], Sara Sillaurren [2], Ana Moya [1] and Nerea Toledo [3]

[1] TECNALIA Basque Research and Technology Alliance (BRTA), Derio Bizkaia, Spain; leire.bastida@tecnalia.com (L.B.); ana.moya@tecnalia.com (A.M.)
[2] TECNALIA Basque Research and Technology Alliance (BRTA), Miñano Vitoria, Spain; sara.sillaurren@tecnalia.com (S.S.)
[3] University of the Basque Country UPV/EHU, School of Engineering of Bilbao, Spain; nerea.toledo@ehu.eus (N.T.)
* Correspondence: erlantz.loizaga@tecnalia.com (E.L.)

Abstract: Recognizing trust as a pivotal element for success within Human-Robot Collaboration (HRC) environments, this article examines its nature, exploring the different dimensions of trust, analysing the factors affecting each of them, and proposing alternatives for trust measurement. To do so, we designed an experimental procedure involving 50 participants interacting with a modified 'Inspector game' while we monitor their brain, electrodermal, respiratory, and ocular activities. This procedure allowed us to map dispositional (static individual baseline) and learned (dynamic, based on prior interactions) dimensions of trust considering both demographic and psychophysiological aspects. Our findings challenge traditional assumptions regarding the dispositional dimension of trust and establish clear evidence that the first interactions are critical for the trust-building process and the temporal evolution of trust. By identifying more significant psychophysiological features for trust detection and underscoring the importance of individualized trust assessment, this research contributes to understanding the nature of trust in HRC. Such insights are crucial for enabling more seamless human-robot interaction in collaborative environments.

Keywords: Human-Robot Collaboration (HRC); trust dimensions; trust dynamics; experimental process

1. Introduction

Human-Robot Collaboration (HRC) has emerged as a critical area in the engineering and social sciences domain. This paper ventures into this dynamic domain with a keen focus on environments, where the collaboration pivots on execution of routine tasks involving manipulation of components on recycling disassembly lines as well as management and classification of various electronic devices. We particularly explore the performance of collaboration with robotic arms, capable of a spectrum of autonomous actions under the guidance of a computer program.

In any kind of collaboration, including HRC, trust has been identified as a significant factor that can either motivate or hinder cooperation, especially in scenarios characterized by incomplete or uncertain information. Despite the ubiquitous understanding of the concept of "trust", its definition has proven to be complex due to the range of fields it applies in and the individual context in which it is studied. Several perspectives contribute to the understanding of trust, depending on the theoretical focus and the specific field of study it is applied to. While interpersonal trust is the most studied, there is also an increasing focus on the trust between humans and technology, which is essential in HRC.

This paper aims to delve into significant research questions about the function of trust in HRC: 1) What are the fundamental factors influencing trust in HRC?, 2) Is it possible to identify demographic or contextual variables that affect the nature and dynamics of trust?, and 3) Can trust in HRC be measured using psychophysiological signals?

 doi:10.20944/preprints202312.1915.v1

The paper is organized as follows: Section 1 introduces the concept of trust in HRC, highlighting its general framework and outlining the major research questions this paper intends to explore. Section 2 provides a theoretical background, covering current research and explaining the importance of defining trust in HRC. Section 3 describes the Materials and Methods, with a detailed explanation of the experimental design and methodologies used in the data analysis. Section 4 presents the Results, providing the empirical findings form the experimental process. In Section 5, a comprehensive Discussion is included, thoroughly exploring the implications of the research findings for trust in HRC. Lastly, Section 6 offers the Conclusion and provides answers to the research questions.

2. Background)

2.1. Theoretical Foundations of Trust

Trust is a crucial determinant of effective collaboration, in both human-to-human and human-to-machine interactions. Consequently, studies on trust modelling and measurement span a variety of disciplines, including psychology, sociology, biology, neuroscience, economics, management, and computer science [1–10]. These two approaches (– modelling and measuring –) share common knowledge, but with differing purposes and considerations. Trust modelling aims to depict human trust behaviour, extrapolating individual responses to a universal level, whereas trust measurement seeks quantifiable involuntary body responses to varying trust-related stimuli.

The multidisciplinary nature of human trust research signifies that defining a unique modelling approach is complex. A variety of trust definitions catered to specific topics contribute to a consolidated and detailed definition of the concept (see Table 1 for more details).

Table 1. Multidisciplinary definitions of Trust [2].

Discipline	Meaning of Trust
Sociology	Subjective probability that another party will perform an action that will not hurt my interest under uncertainty and ignorance [1].
Philosophy	Risky action deriving from personal and moral relationships between two entities [3].
Economics	Expectation upon a risky action under uncertainty and ignorance based on the calculated incentives for the action [4].
Psychology	Cognitive learning process obtained from social experiences based on the consequences of trusting behaviours [5].
Organizational management	Willingness to take risk and being vulnerable to the relationship based on ability, integrity, and benevolence [6].
International relations	Belief that the other party is trustworthy with the willingness to reciprocate cooperation [7].
Automation	Attitude that one agent will achieve another agent's goal in a situation where imperfect knowledge is given with uncertainty and vulnerability [8].
Computing & Networking	Estimated subjective probability that an entity exhibits reliable behaviour for particular operation(s) under a situation with potential risks [9].

In general terms, trust is perceived as a relationship where a subject (trustor) interacts with an actor (trustee) under uncertain conditions to attain an anticipated goal. In this scenario, trust is manifested as the willingness of the trustor to take risks based on a subjective belief and a cognitive assessment of past experiences that a trustee will demonstrate reliable behaviour to optimize the trustor's interest under uncertain situations [2].

This definition emphasizes several issues concerning the nature of trust:

- A subjective belief: Trust perception heavily relies on individual interactions and the preconceived notion of the other's behaviour.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

- To optimize the trustor's interest: Profit or loss implications for both the trustor and trustee through their interactions reveal the influence of trust/distrust dynamics.
- Interaction under uncertain conditions: The trustor's actions rely on expected behaviours of the trustee to optimize the anticipated outcome, but it may yield suboptimal or even prejudicial results.
- Cognitive assessment of past experiences: Trust is dynamic in nature, initially influenced by preconceived subjective beliefs but evolving with ongoing interactions.

Authors have proposed varying dimensions of trust to explain different elements of trust development. Some differentiate between moralistic trust, based on previous beliefs about behaviour, and strategic trust, based on individual experiences [10]. Other approaches identify dispositional, situational and learned trust as distinct categories [11]. Trust is also described as a combination of individual trust (derived from own personal characteristics and conformed by logical trust and emotional trust) and relational trust, referenced to the dimensions of trust that rise from the relationship with other entities [2].

Even if different authors use different classifications, it is possible to map similarities between different approaches. Despite omitting minor particularities, Table 2 shows the convergence of these classifications. For convenience, we will use the nomenclature defined in [11] – **dispositional, situational and learned trust** – without loss of generality.

Table 2. A rough similarity map between trust dimensions among different authors.

Similarity map of trust dimensions according to different authors			
[10]	[11]	[12]	[2]
Moralistic	Dispositional	Phenomenon-based	Emotional
	Situational	Sentiment-based	Relational
Strategic	Learned	Judgement-based	Logical

In essence, trust relies on multiple, complex factors, encompassing both individual and relational aspects. Notwithstanding the diverse disciplinary perspectives and methodologies in researching trust, there is considerable consensus on the fundamental concept and dimensions of trust. However, understanding, modelling, and measuring trust, particularly in human-to-machine contexts, continue to pose considerable challenges.

Tackling these challenges requires an in-depth understanding of multifaceted trust dynamics. The primary challenge lies in encapsulating the complexities of human trust in a computational model, given its subjectivity and dynamic nature. Quantifying trust is another significant obstacle, as trust is an internal and deeply personal emotion. The novelty of trust in the human-robot collaboration domain implies a lack of historical data and testing methodologies to build the trust models upon. Furthermore, implementing these models in real-world scenarios is another challenge due to constraints related to resources, variability in responses and the need for instantaneous adaptation. Despite these hurdles, the potential rewards of successfully modelling and measuring trust in human-robot collaborations - including enhanced efficiency, increased user acceptance and improved safety - are immense.

2.2. Trust in Human-Robot Collaboration

within the field of Human-Robot Collaboration, trust plays a crucial role and is considered a significant determining factor. Various studies, including [13,14], have dedicated efforts to investigate and identify the factors that influence trust in this collaborative context. These factors have not only been structured within a single matrix but also classified based on their origins and dimensions of influence, which are instrumental in facilitating trust and designing experimental protocols.

Authors in [15] provide a series of controllable factors with correlation to trust:

- **Robot behaviour**: This factor relates to the necessity for robot companions to possess social skills and be capable of real-time adaptability, taking into account individual human preferences

 doi:10.20944/preprints202312.1915.v1

[16,17]. In the manufacturing domain, trust variation has been studied in correlation to changes in robot performance based on the human operator's muscular fatigue [18].

- **Reliability**: An experiential correlation between subjective and objective trust measures was demonstrated through a series of system failure diagnostic trials [19].
- **Level of automation**: Consistent with task difficulty and complexity and corresponding automation levels, alterations in operator trust levels were noted [20].
- **Proximity**: The physical or virtual presence of a robot significantly influences human perceptions and task execution efficiency [21].
- **Adaptability**: A robot teammate capable to emulate the behaviours and teamwork strategies observed in human teams has a positive influence in trustworthiness and performance [22].
- **Anthropomorphism**: With anthropomorphic interfaces, greater trust resilience was recorded [23].
- **Communication**: Trust levels fluctuated based on the transparency and detail encapsulated within robot-to-human communication [24,25].
- **Task type**: The task variability was recorded to influence interaction performance, preference, engagement, and satisfaction [26].

Less controllable dimensions of trust include dispositional trust that is influenced exclusively by human traits and the organizational factors linked to the Human-Robot team [13,14]. These factors exhibit limited flexibility as they depend directly on the individual or the organizational culture. On the other hand, situational trust is controllable, heavily dependent on various factors such as the characteristics of the task being developed, making it possible to manipulate it based on the experiment's objective [13,14].

Moreover, trust manifests through brainwave patterns and physiological signals, making their use in assessing trust crucial [27,28]. Biologically driven, these elements foster a more symbiotic interaction, allowing machines to adapt to human trust levels.

Notably, trust in HRC is dynamic and influenced by a myriad of factors. Understanding the various dimensions of trust and the controllable and uncontrollable factors encompassed allows for the creation of experimental protocols and strategies to enhance trust, a hypothesis evident in studies looking into the triad of operator, robot, and environment [14]. The importance of fostering and maintaining trust in the HRC domain, especially considering the complexity of trust in the ever-evolving landscape of Human-Robot interaction.

2.3. Trust measuring using different and combined psychophysiological signals

Studies on trust have traditionally been situated within the context of interpersonal relationships, primarily utilizing various questionnaires to evaluate levels of trust [29–31]. However, due to the arrival of automatic systems and the decreased cost of acquiring and analysing psychophysiological signals, focus has shifted towards examining these types of signals in response to specific stimuli in a bid to lower the subjectivity and potential biases associated with questionnaire-based approaches. Recent studies, like [9,32–34], have been centred on the usage of psychophysiological measurements in the study of human trust.

The choice of these psychophysiological sensors can differ, depending on which human biological systems (central and peripheral nervous systems) they are applied to. A common pattern has emerged from studies in which EEG is the most used signal to measure central nervous system activity, with fMRI closely behind it - the latter being more extensively used in the context of interpersonal trust [35]. Additionally, attempts have been made to study trust through EEG measurements which only look at event-related potentials (ERPs), but ERP has proven to be unsuitable for real-time trust level sensing during human-machine interaction due to difficulty in identifying triggers [33,36].

Similarly, sensors measuring signals from the peripheral nervous system, notably ECG (electrocardiography) and GSR (galvanic skin response), have been frequently used in assessing trust [35]. GSR, a classic psycho-physiological signal that captures arousal based on the conductivity of the skin's surface, not under conscious control but instead modulated by the sympathetic nervous system, has seen use in measuring stress, anxiety, and cognitive load [37]. Some research revealed

 doi:10.20944/preprints202312.1915.v1

that the net phasic component as well as the maximum value of phasic activity in GSR, might play a critical role in trust detection [33].

In contrast, the use of single signals most commonly involves only EEG, succeeded by fMRI [35]. However, some studies have proposed that combining different psychophysiological signals (like GSR, ECG, EEG, etc.) improves the depth, breadth, and temporal resolution of results [38,39]. Interestingly, pupillometry has been recently highlighted as a viable method for detecting human trust, revealing that trust levels may be influenced by changes in partners' pupil dilation [40–42].

In short, the current state of the art exhibits an increasing trend towards the use of psychophysiological signals emanating from both central, and peripheral nervous systems. Also, it showcases an interest in combining these signals to create more robust trust detection mechanisms with an improved breadth and depth of results.

3. Materials and Methods

The study of the dynamics that determine trust is a topic of great interest, although there is limited collected data available for analysis to draw conclusions. In our case, the objective of the experimentation is to design and implement a process that can identify in real time if there has been a breach of trust between the operator and their robot companion (cobot) and to understand the factors that cause these variations in trust and identify the biological reactions related to it.

To minimize the inclusion of excessive variables that could occlude the true nature of trust, we imply that the cobot is a non-humanoid robotic arm with limited interaction capabilities. Thus, the only interaction with the human counterpart is reduced to the effectivity and performance of tasks under its control. Additionally, in order to avoid contextual randomness, we decided that every participant should interact with the system in a similar environment, and thus, we determined that the experimental study will focus solely in the dispositional and learned dimensions of trust.

Consequently, in the context of this research, trust is defined as the predictability of the human that the robot will exhibit appropriate behaviour, following the established guidelines for the performance of its work in an orderly manner, without failure or errors that could be interpreted as dangerous to the person, the robot itself, or its work environment.

Our strategic planning for the experimental process aimed at achieving the following objectives:

- Systematically collect a diverse array of relevant psychophysiological signals, emphasizing signal cleanliness and minimizing signal randomness;
- Investigate the influence of human traits on various aspects of human-machine trust, specifically in the context of dispositional and learned trust dimensions;
- Examine the role of the system's capabilities, especially predictability and reliability, in shaping the evolutionary process of trust.

In the following subsections, the conceptual design of the experimentation, as well as the equipment and method used for this research, are described in detail.

3.1. Conceptual design of the experiment

Considering the objectives to be satisfied and to expose participants to different trust stimuli towards machines, this experimental stage has been designed as a game following a variant of the Prisoner's Dilemma, known as the Inspection Game. These types of games are mathematical models that represent a non-cooperative situation where an inspector must identify whether his counterpart adheres to the established guidelines or, on the contrary, shrinks work duties. In this case, the participants take the role of the inspector and their mission is to detect if their robotic counterpart is carrying out their job adequately.

Simplicity has been kept to a maximum during the design of this experiment, thus participants interact with a single screen that provides them the sensor feedback. The trust decision process is implemented via a single command bottom. Participants are exposed to 120 iterations during an approximate time of 30 minutes. However, the first 20 iterations are part of the learning and familiarization process and, thus, only 100 iterations per participant are considered in the experimental analysis.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

Unknown to the participants, two experimental models were performed. Both had the same trustworthiness (the virtual sensor provided a correct answer 75 out of the 100 iterations), but in one of the cases the machine worked perfectly during the first 20 iterations whereas the second model presented 50% success rate during the same first 20 iterations.

3.2. Experimental sample

The experimental sample must be selected according to the population in which it is desired to validate the experimentation. In this case, since the final objective is still oriented to an industrial work environment, the target population is very broad, since it covers all individuals of active working age.

To refine the profile of the desired sample, its implications within the broader research context must be considered. Specifically, the demographical distribution of the sample has a direct impact on the collection of human traits and their eventual influence in the dispositional dimension of trust.

Taking into account these considerations, three factors are determined with which to compose the selected sample: Gender (Male/Female), Age (Under and over 40 years old) and Role in the work team (Technical/Non-technical). The first two belong to the category of human traits, while the third falls into the scope of the work environment. Figure 1 shows the distribution of the participants in this experimental stage according to the indicated categories, revealing balanced sample, with the exception of the age segment.

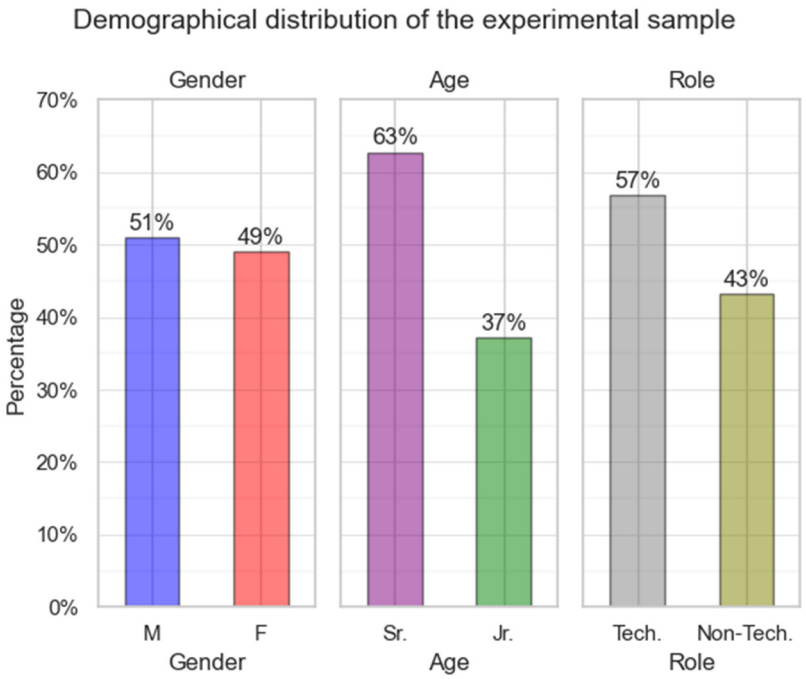

Figure 1. Categorical distribution of participants.

3.2. Equipment used

Given the specific objectives of this experimental phase, it may be prudent to consider acquiring a maximal number of psychophysiological signals possible to subsequently discriminate which ones are truly significant for the study of trust. However, this approach presents two significant challenges. Firstly, the acquisition of a multitude of signals for subsequent treatment without prior consideration could pose serious logistical problems for the experimental design and significant difficulties in their subsequent analysis. Secondly, the consideration of signals that, due to their acquisition methods, could be incompatible with real activity in industrial environments would not align with the approach of this research.

 doi:10.20944/preprints202312.1915.v1

After analysing the different bibliographic sources referring to previous studies, as well as the signals used during their execution and the acquisition systems used in each case, it is determined that the most appropriate signals to propose are those linked to:

- Brain activity (EEG): Rigid headband with twelve dry electrodes for reading of brain activity in anterior frontal regions AF [7-8], frontopolar Fp [1-2],frontal F [3-4], parietal P [3-4], parieto-occipital PO [7-8] and occipital O [1-2].
- Galvanic skin response (GSR): Electrodes positioned on the index and ring fingers non-dominant hand. In state excitement, glands sweat glands are activated, varying the electrical resistance of the skin. An applied low voltage current between both points allows detect these variations.
- Respiration (RSP): Elastomeric band located at the height of the diaphragm. Issue small electrical signals when varying its extension, so it is possible identify inhalation and exhalation cycles. They provide information about the frequency respiratory, tidal volume and characteristics of the respiratory cycle.
- Pupillometry (PLP): Glasses equipped with eye tracking sensors which enable the identification of the fixation point of gaze or eye movement refixation saccades. In addition to the direction of the gaze, they also provide information about the diameter of the pupil of each eye, which allows for the derivation of other parameters such as blinking frequency.

3.3. Experimental process

In order to ensure that the experimental phase aligns with the established objectives, it is imperative to define a set of guidelines during execution that endorse the accurate execution of the process. The tasks earmarked for experimentation are delineated below.

- **Participants reception**: Participants are briefed about the project and the experimentation, ensuring they are informed about the purpose, the physiological signals that will be collected, and the treatment they will receive. They are assured of data privacy through pseudo-anonymization and told of their right to opt-out anytime. Once they consent, they provide demographic data and complete a technology trust survey. They are then familiarized with the experimental setup and equipment to capture psychophysiological signals. Participants are instructed to minimize movement during the experiment for data quality control.
- **Biocalibration**: The biocalibration phase ensures the equipment is accurately tuned to individuals' varying physiological responses. This adjustment considers that without context, a specific value cannot conclusively indicate high or low intensity. This phase helps define the participant's normal thresholds in varied states of relaxation and excitement. After equipping the participants with the measuring gear, they perform tasks designed to both stimulate and soothe their signals, thereby minimizing uncontrolled disturbances.
- **Familiarization**: The familiarization stage aims at ensuring participants completely understand their tasks and possible implications during the experiment. In this phase, participants repetitively interact with the machine to understand its workings, ensuring they can easily express their trust or mistrust. Unlike the experimental stage, they're made aware of the sensor's performance, helping them form trust-based responses. This process also helps them become accustomed to the screens displaying crucial information during the interaction process.
- **Experimental process**: During the experiment, participants interact with the machine and gauge the sensor's trustworthiness. They are presented with a system state ("well lubricated" or "poorly lubricated") and the default action matches the system state. They only interact if choosing to disregard the sensor. They are then informed on the real machine state and the result of their decision. This cycle repeats a hundred times with varying patterns unknown to the participants.
- **Informal interview**: A brief interview follows the experimental phase for each participant to assess their experience, identify disruptions, and understand areas of future improvement. It is especially important for participants presenting anomalies in signal visualization or behaviour. It helps filter data from those negatively affected by conditions like discomfort with measuring instruments or misunderstanding their tasks. This interview also aids in understanding participants' perception of the system's reliability and identifying personal traits influencing their perception.

 doi:10.20944/preprints202312.1915.v1

Figure 2. Experimental phase.

4. Results

The following section comprises the experimental results, divided according to the research interests and main findings.

4.1. Factors affecting dispositional trust

We utilized information from a concise affinity survey during participants' orientation to identify potential factors influencing their natural inclination to trust in a Human-Robot environment. These data were collected before any interaction with the designed experiment, ensuring they remain unaffected by procedural biases in the experiment and reflect individuals' intrinsic perspectives on trust in robot collaboration.

Additionally, each participant provided demographic information, including gender, age, and the technological intensity of their work role, along with a self-assessment of their trust in robots. At this stage, we investigated potential relationships between these variables and the ex-ante self-assessed trust in robots.

Figure 3 presents descriptive charts (histogram, probability density function -PDF-, cumulative distribution function -CDF-, and boxplot chart) illustrating dispositional trust across three dimensions of interest: gender (male versus female), age (juniors -younger than 40 years old- versus seniors -older than 40 years old), and work profile (highly technical versus non-technical).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

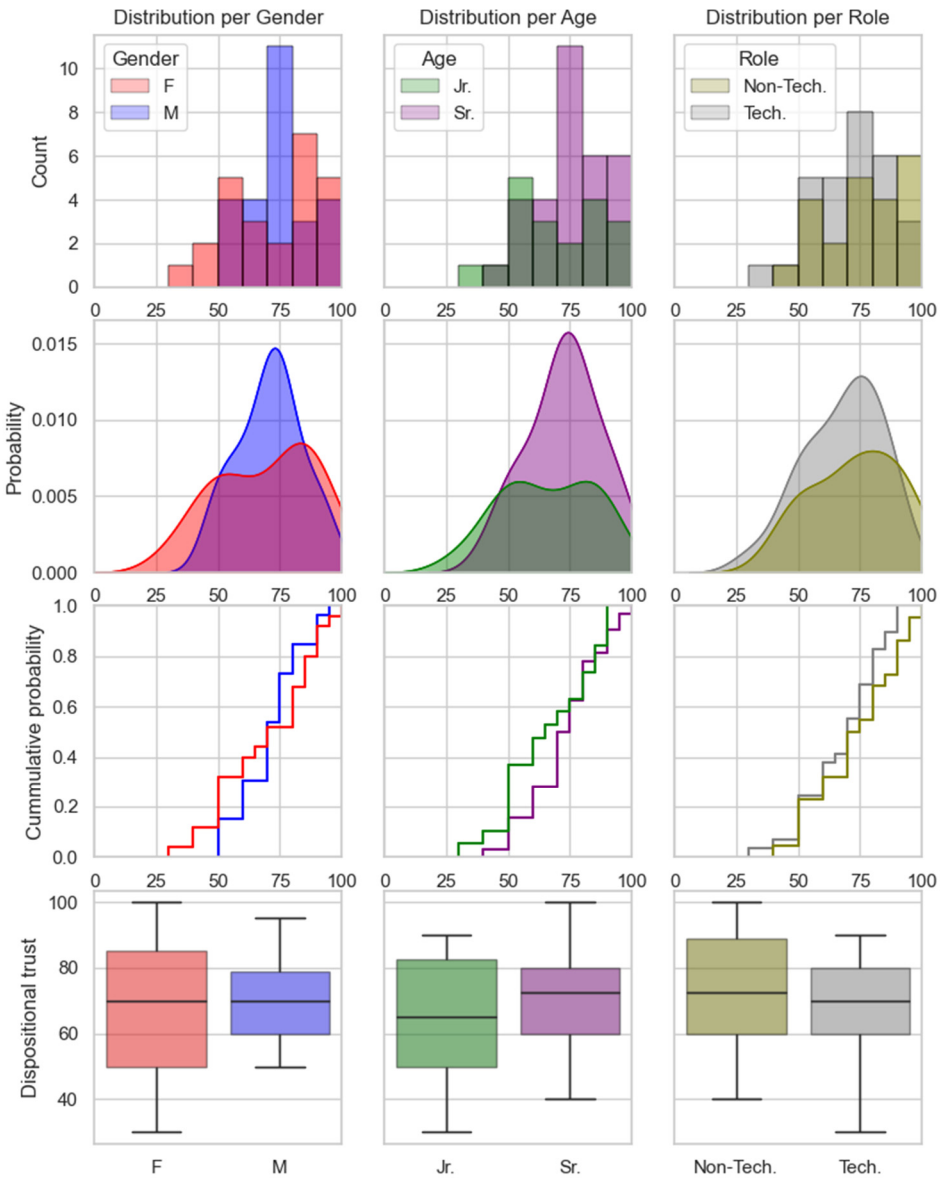

Figure 3. Dispositional trust distribution per different demographical factors.

Figure 3 highlights differences among the analyzed demographic groups. For instance, both the histogram and PDF reveal a noteworthy peak at the 70%-80% trust level for females, and a similar peak is observed for juniors. This pattern is absent in their respective counterparts (males and seniors). The boxplot chart also indicates that senior individuals exhibit a slightly superior mean trust level compared to their younger counterparts.

To validate these perceptions, and considering the non-normal distribution of data, we conducted a Mann-Whitney U-test to check for disparities in data distribution of the different gender, age and role collectives. Table 3 presents the obtained p-values from both single-tailed and double-tailed tests. The results indicate that, while differences may exist, demographics alone are not significant enough to explain such disparities.

Table 3. Results of the Mann-Whitney U-test for dispositional trust between different demographics.

Demographic	Segments	Statistic	Test	p-value
Gender	Female (X0) – Male (X1)	322.0	X0 <> X1	0.9621
			X0 > X1	0.5265
			X0 < X1	0.4810

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

Age	Junior (X0) – Senior (X1)	247.5	X0 <> X1	0.2713
			X0 > X1	0.8685
			X0 < X1	0.1357
Role	Non-Technical (X0) – Technical (X1)	372.0	X0 <> X1	0.3141
			X0 > X1	0.1570
			X0 < X1	0.8475

4.2. Factors affecting perceived trust

After the experimental session, we instructed the participants to assess the system's trustworthiness. It is important to note that, regardless the interaction model the participants experimented, the system behaved correctly 75% of the times, resulting in a consistent objective trust level across all cases which allowed us to compare results.

Once more, we examined the distribution of perceived trust and explored the influence of demographic factors on this variable, mirroring the analysis conducted for dispositional trust. However, given that these assessments were ex post, we also considered the impact of the interaction model to which participants were exposed. Figure 4 and Table 4 provide a summary of these results.

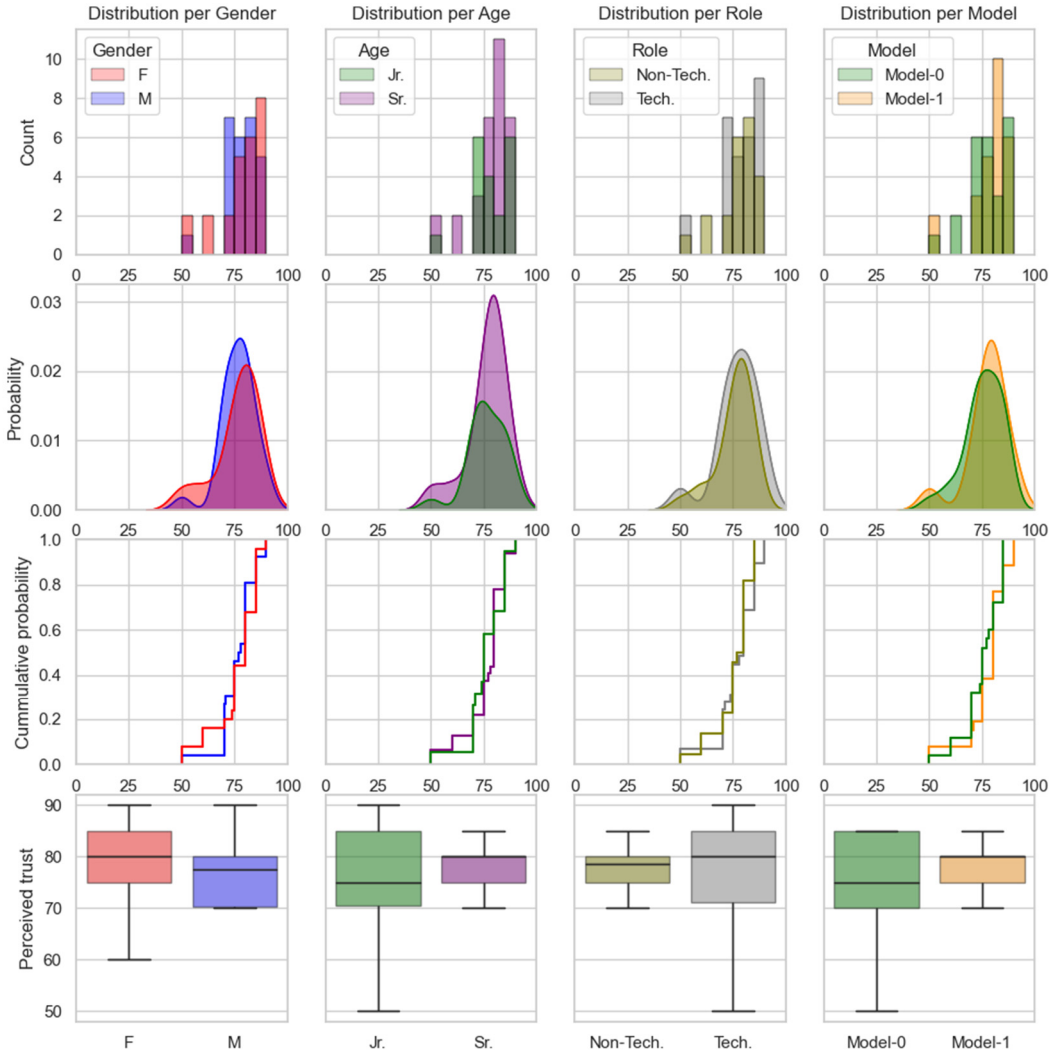

Figure 4. Perceived trust distribution per different demographical factors.

Similarly to dispositional trust, Figure 4 shows some possible differences regarding the demographical data distribution such as higher trust values for senior participants and subjects exposed to the experimental Model-1. However, the results of the conducted Mann-Whitney test

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

compiled in Table 4 show that these divergences are not significant enough to explain such disparities.

Table 4. Results of the Mann-Whitney U-test for perceived trust between different demographics.

Demographic	Segments	Statistic	Test	p-value
Gender	Female (X0) – Male (X1)	352.5	X0 <> X1	0.6050
			X0 > X1	0.3025
			X0 < X1	0.7041
Age	Junior (X0) – Senior (X1)	282.5	X0 <> X1	0.6775
			X0 > X1	0.6685
			X0 < X1	0.3388
Role	Non-Technical (X0) – Technical (X1)	293.0	X0 <> X1	0.6220
			X0 > X1	0.6956
			X0 < X1	0.3110
Exp. Model	Model-0 (X0) – Model-1 (X1)	270.5	X0 <> X1	0.3001
			X0 > X1	0.8539
			X0 < X1	0.1505

4.3. Influence of past iteraction in trust dynamics

The experimental process exposed participants to a series of interactions with an overall 75% reliability. Two experimental models were conceived to alter the sequence in which participants faced the test. Both models alternated between a fully reliable scenario and a chaotic sequence where only 50% of the readings were correct. Model-0 began with trustworthy iterations, while Model-1 presented chaotic iterations first. This setup allows us to investigate whether the order in which participants faced the test influences the propensity to trust.

To integrate the interaction results, we counted the times participants chose to trust the sensor's reading on each experimental stage and divided the result by the number of interactions on that stage, obtaining the trust rate for each stage. These results were aggregated according to the experimental model they interacted with. Figure 5 shows the distribution of trust rates among the different models and experimental stages, along with the mean trust rate for each case.

To verify the impact of past interactions on the conformation of trust, we decided to perform a statistical test on the trust rate distributions among the different experimental models for both the case of perfectly working sensors and for the randomly working sensor scenarios. Since Figure 5 shows that the trust rate distribution does not follow normality, we conducted the Mann-Whitney U-test to check the significance of the trust rate differences among experimental models. Table 5 shows the results of these tests, emphasizing those with statistical significance.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

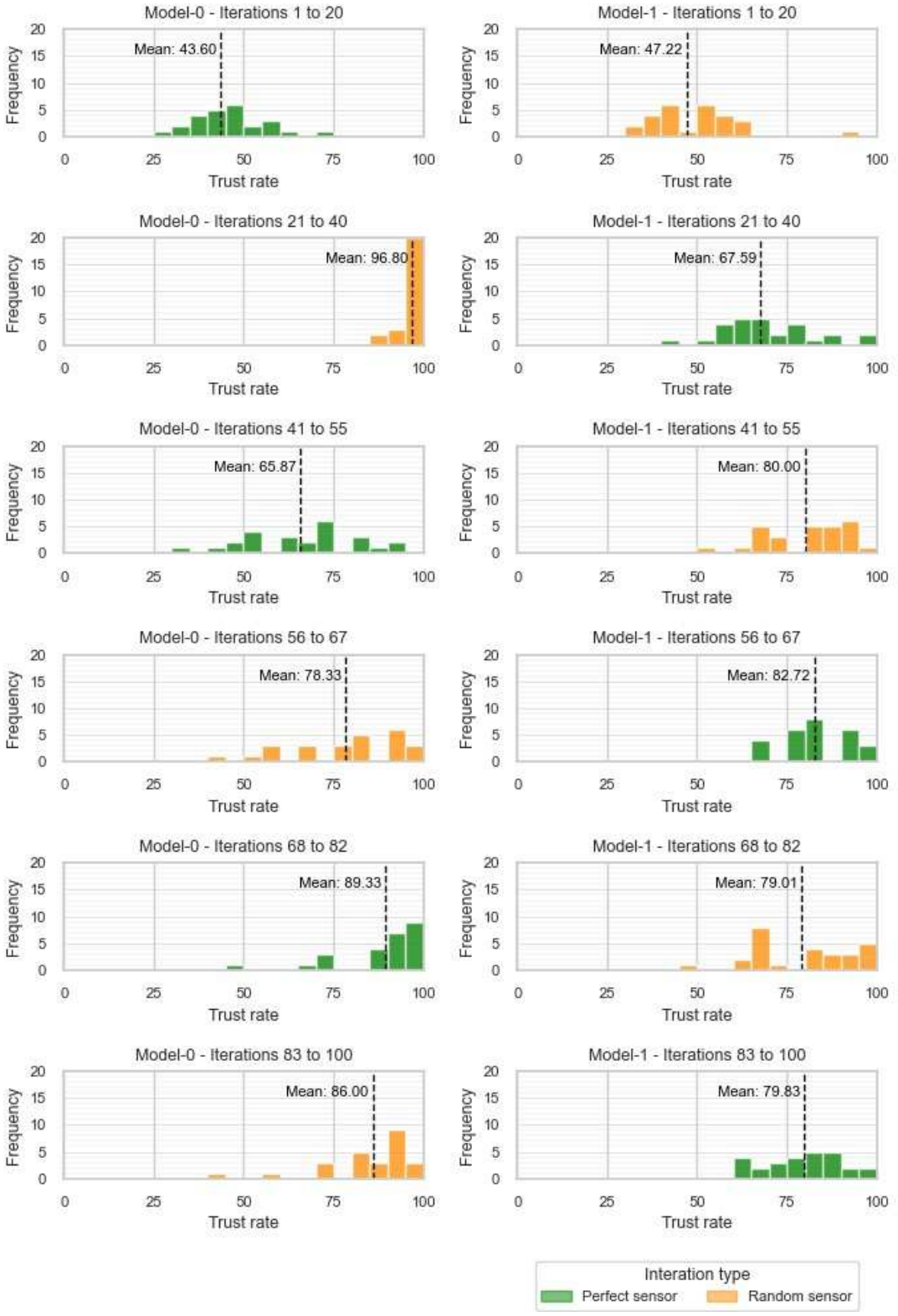

Figure 5. Trust rate distribution among experimental model and stages.

Table 5. Results of the Mann-Whitney U-test for trust dynamics between different models and experimental stages.

Stage	Segments	Statistic	Test	p-value
Correct sensor	Model-0 (X0) – Model-1 (X1)	2231.0	X0 <> X1	**(*) 0.0042**
			X0 > X1	0.9979

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

			X0 < X1	**(*) 0.0020**
Random sensor	Model-0 (X0) – Model-1 (X1)	4692.0	X0 <> X1	**(**) 3.840e-09**
			X0 > X1	**(**) 1.920e-09**
			X0 < X1	1.000

(*) Results presenting statistical significance.

(**) Results presenting very strong statistical significance

4.4. Universality of trust detection models

Deciphering trust in the intricate dynamics of human-robot requires many variables. The physiological signals captured during the experiment, including 12 detailed signals related to brain activity (EEG), skin conductance response (GSR), respiratory band extension (RSP), and pupil diameter variations (PLP), offer a variety of information to train a suitable model. However, there are some considerations to ponder to choose the most suitable approach. We analyzed three distinct yet complementary approaches in creating a trust detection model.

The General Approach involved constructing an expansive dataset comprising all participants' data and iterations. This method allowed us to leverage a broad dataset, providing a comprehensive overview of general trust trends. However, it introduced the challenge of potential information leaks between the training and test sets, as participant-specific information was included in both sets. Under this approach a single model covers all the trust detection need for every participant.

In the Leave-One-Out Approach, we addressed the risk of information leaks by training algorithms using the complete dataset, excluding participant-specific information. While this approach maintained a substantial amount of data, it sacrificed personalized information crucial for trust detection, leaving out precise the data that could contain the most valuable information. This approach created a specific classifier for each participant and the results where later aggregated.

The Individual Approach focused on the uniqueness of trust reactions. By exclusively utilizing each participant's individual dataset, this method created personalized algorithms for each individual. Despite having the smallest dataset for training, the Individual Approach allowed for the detection of specific trust nuances in each participant. As in the previous case, a specific model was created for every participant.

Table 6 presents key metrics, including minimum, mean, and maximum F1 scores, obtained with each training model, showcasing the strengths and nuances of each approach.

Table 6. F1 scores.

	Minimum	**Mean**	**Maximum**
General Approach	-	0.6172	-
Leave-One-Out	0.6098	0.6207	0.6326
Individual Approach	0.6363	0.7661	0.9219

4.5. Universality of signals used for trust detection

During the experiment we recorded a large variety of signals, including brain activity, electrodermal response, respiration, and pupillometry. However, whether these signals are valuable to detect general variations of trust is unclear. We seek to discern whether these signals exhibit consistent patterns applicable to detect trust variations in all the participants or whether they may be referred as key elements to detect trust variations on certain individuals.

To analyze this issue, we systematically computed the frequency with which each signal contributed to the general analysis approach and, concurrently, how frequently it featured in individualized models. Figure 6 succinctly encapsulates the outcomes, providing insights into the role of each signal in both general and individualized trust detection models.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 December 2023 doi:10.20944/preprints202312.1915.v1

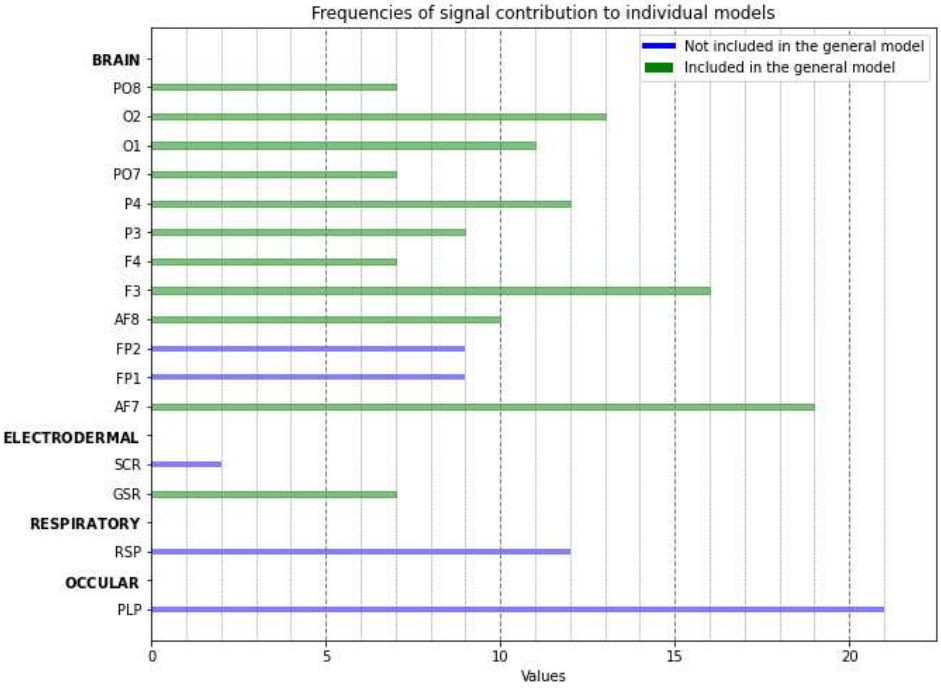

Figure 6. Summary of signal contribution to general and individual models.

5. Discussion

Throughout this article, we have presented the different dimensions of trust (dispositional, situational, and learned) and outlined the experimental process undertaken to determine the factors influencing them.

The literature consulted indicates a marked influence of personal factors on dispositional trust. Nevertheless, our empirical results contradicted conventional expectations by unveiling the absence of statistically significant correlations between these customary demographic indicators and the participants' dispositional trust in automated systems. Since our analysis focused on gender, age, and occupational role, we suggest to further analyze the potential influence of unexplored demographic aspects. Variables such as social status, educational background, and nationality, may wield a significant. Their potential significance to shape dispositional trust cannot be disregarded, and further studies are needed to analyze their impact.

To analyze the evolution and dynamics of learned trust, we designed two experimental models with similar iterations arranged in different orders, and thus, allowing the detection of variations in trust evolution dynamics. As indicated by statistical tests, the behavior of both groups is disparate. Specifically, participants who start with a series of iterations exhibiting random behavior (participants interacting with Model-1) experience a higher dynamic variation compared to their counterparts (participants interacting with Model-0), thus exhibiting higher trust levels when the system functions properly and lower trust levels when the system's performance is erratic. In this context, it could be argued that the initial iterations with the system significantly influence the trajectory of future iterations. Drawing an analogy with classical mechanics, it can be asserted that the initial iterations are critical in determining the level of "trust inertia" individuals have towards automated systems and their future interactions.

Regarding the third component of trust, situational trust, its influence is not directly covered by the experimentation and analysis conducted in this research. However, a small portion of it is reflected in the link between post-trust perception and the interaction model to which each participant has been subjected. This is justified by considering that this combination incorporates elements external to the individual into a static value of trust and, therefore, cannot correspond to dispositional (individual) trust or learned (dynamic) trust. The analyses performed highlight a low

 doi:10.20944/preprints202312.1915.v1

significance in the link between these elements. This is mainly because this factor is not the main focus of the study, and thus, the experimental process is not designed to emphasize its influence.

Regarding mathematical models aimed at trust detection, we designed a common model for all participants using the generalist approach. The Leave-One-Out approach is also suitable, in general terms, for any individual and, in this regard, allows the same universalization as the general model. These models offer a moderate performance, achieving an F1 score close to 0.61-0.62. On the other hand, individualized models show much higher performance than these approaches, achieving F1 values that surpass those of the previous models in all cases, reaching an average F1 score of 0.76 and even reaching occasional values of 0.92 on this score. This indicates that, compared to generalist models, individualized models offer better performance in detecting trust violation and recovery situations. However, the implementation of this approach faces other challenges, such as the need to train independent algorithms for each subject.

The influence of the different psychophysiological signals in the proposed prediction models must also be considered. Variables such as pupil dilation or changes in breathing patterns play a fundamental role in implementing personalized trust detection models, buts lack significance in the generalist model. Conversely, signals like the brain activity in the parietal-occipital lobe (electrodes PO7 and PO8) contribute to the generalist model but has very little influence on personalized models. Although the cause of these phenomena has not been studied in detail, it can be inferred that signals included in the generalist model exhibit similar behavior in most individuals, but these changes are not necessarily the most sensitive to variations in trust situations.

Having outlined the scope of the results obtained, it is worth mentioning some points that remain inconclusive in this research and could be addressed in future work.

6. Conclusions

To conclude our study, we need to review the core research questions that guided our research regarding the nature of trust in Human-Robot Collaboration.

First, we aimed to identify the fundamental factors affecting trust in HRC environments. We addressed this issue with an extensive literature review that revealed three distinct dimensions of trust: dispositional, situational, and learned. This review also stressed several demographic aspects that influence the dispositional dimension of trust.

The second question, aimed to identify specific variables that could influence the previously disclosed dimensions of trust. To answer this issue, we created a specific experimental process that allowed us to identify potential factors affecting dispositional and learned trust. The literature suggested a marked influence of personal factors such as gender, age, and occupational roles on dispositional trust. However, our empirical findings challenged conventional assumptions, revealing no statistically significant correlations between these traditional demographic markers and participants' dispositional trust in automated systems. This unexpected result prompted us to delve further into broader demographic factors—social status, educational background, and nationality—which were not explicitly considered in our initial study. On the other hand, the study proved that the first iterations between humans and robots play a crucial role the evolution of trust dynamics. Presumably, this points out that the learned dimension of trust is more sensible and susceptible to change than the other dimensions.

The third and final research question aimed to identify crucial human signals for measuring trust in HRC and understand their contributions to the development of trust detection models. Following previous works detailed in the literature review, our exploration into psychophysiological signals encompassed brain (EEG), electrodermal (GSR), respiratory (RSP), and ocular (PLP) activities. We focused on three different approaches, varying from a generic to individualized trust models. Results revelled the difficulty to extrapolate a general model of trust. On one hand, the individualized models worked better than the general model, and, on the other hand, several individually significant psychophysiological signals showed very particular responses and, thus, resulted irrelevant in the general model. This issue emphasizes the very complex and personal nature of trust.

In conclusion, our study effectively addresses the research questions, shedding light on the intricate interplay of factors influencing trust, the temporal dynamics of trust evolution, and the optimal human signals for trust measurement in HRC. We are confident that this research will empower the design of future reliable, robust and trusted Human-Robot collaborative environments.

Author Contributions: Conceptualization, E.L., L.B., A.M., S.S. and N.T; methodology, E.L.; software, E.L.; validation, L.B., A.M. and S.S.; formal analysis, E.L.; investigation, L.B., A.M. and S.S.; resources, S.S.; data curation, E.L.; writing—original draft preparation, E.L., L.B., A.M., S.S. ; writing—review and editing, L.B., S.S and N.T..; visualization, A.M.; supervision, L.B. and S.S.; project administration, S.S.; funding acquisition, L.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from the European Union's Horizon 2020 research and Innovation Programme under grant agreement No. 820742. The results obtained in this work reflect only the authors views and not the ones of the European Commission; the Commission is not responsible for any use that may be made of the information they contain.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the HR-RECYCLER's Ethics Manager (EM) from VRIJE UNIVERSITEIT BRUSSEL (December 2018).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study and the participation to the study was voluntary.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The authors would like to thank the participants who took part in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. D. Gambetta, Can we trust, Trust: Making and breaking cooperative relations, vol. 13, pp. 213–237, 2000.
2. J. Cho, K. Chan y S. Adali, A Survey on Trust Modeling, ACM Comput. Surv., vol. 48, no. 2, pp. 28:1–28:40, 2015.
3. B. Lahno y O. lagerspetz, Trust. The tacit demand, Ethical Theory and Moral Practice, vol. 2, no. 4, pp. 433–435, 1999.
4. H. S. James, The trust paradox: a survey of economic inquiries into the nature of trust and trustworthiness, Journal of Economic Behavior & Organization, vol. 47, no. 3, pp. 291–307, 2002.
5. J. B. Rotter, Interpersonal trust, trustworthiness, and gullibility, American psychologist, vol. 35, no. 1, p. 1, 1980.
6. F. D. Schoorman, R. C. Mayer y J. H. Davis, An integrative model of organizational trust: Past, present, and future, 2007.
7. H. Kydd, Trust and mistrust in international relations, Princeton University Press, 2007.
8. J. Lee y K. A. See, Trust in automation: Designing for appropriate reliance, Human factors, vol. 46, no. 1, pp. 50–80, 2004.
9. J. Cho, A. Swami y R. Chen, A survey on trust management for mobile ad hoc networks, IEEE Communications Surveys & Tutorials, vol. 13, no. 4, pp. 562–583, 2010.
10. E. Uslaner, The moral foundations of trust, Cambridge University Press, 2002.
11. K. Hoff y M. Bashir, Trust in automation: Integrating empirical evidence on factors that influence trust, Human factors, vol. 57, no. 3, pp. 407–434, 2015.
12. D. M. Romano, The nature of trust: conceptual and operational clarification, 2003.
13. M. Bashir y K. Hoff, Trust in automation:Integrating empirical evidence on factors that influence trust, Human factors, 57(3):407–434, 2015.
14. K. Schaefer, The perception and measurement of human-robot trust, 2013.
15. P. A. Hancock, D. Billings, K. Schaefer, J. Chen, E. De Visser y R. Parasuraman, A meta-analysis of factors affecting trust in human-robot interaction, Human factors, vol. 53, no. 5, pp. 517–527, 2011.
16. K. Dautenhahn, Socially intelligent robots: dimensions of human–robot interaction, Philosophical transactions of the royal society B: Biological sciences, vol. 362, no. 1480, pp. 679–704, 2007.
17. K. Dautenhahn, Robots we like to live with? A developmental perspective on a personalized, life-long robot companion, International Workshop on Robot and Human Interactive Communication, 2004.

18. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil y Y. Wang, Modeling and control of trust in human-robot collaborative manufacturing, obust Intelligence and Trust in Autonomous Systems, Springer, pp. 115–141, 2016.
19. A. Wiegmann, A. Rich y H. Zhang, Automated diagnostic aids: The effects of aid reliability on users' trust and reliance, Theoretical Issues in Ergonomics Science, vol. 2, no. 4, pp. 352–367, 2001.
20. N. Moray, T. Inagaki y M. Itoh, Adaptive automation, trust, and self-confidence in fault management of time-critical tasks, Journal of experimental psychology: Applied, vol. 6, no. 1, p. 44, 2000.
21. W. A. Bainbridge, J. Hart, E. Kim y B. Scassellati, The effect of presence on human-robot interaction, IEEE International Symposium on Robot and Human Interactive Communication, pp. 701–706, 2008.
22. J. Shah, J. Wiken, B. Williams y C. Breazeal, Improved human-robot team performance using Chaski, A human-inspired plan execution system, International Conference on Human-Robot Interaction, pp. 29–36, 2011.
23. J. de Visser, S. Monfort, R. McKendrick, M. Smith, P. McKnight, F. Krueger y R. Parasuraman, Almost human: Anthropomorphism increases trust resilience in cognitive agents, Journal of Experimental Psychology: Applied, vol. 22, no. 3, p. 331, 2016.
24. K. Akash, T. Reid y N. Jain, Improving Human-Machine Collaboration Through Transparency-based Feedback–Part II: Control Design and Synthesis, IFAC-PapersOnLine, vol. 51, no. 34, pp. 322–328, 2019.
25. K. Fischer, H. Weigelin y L. Bodenhagen, Increasing trust in human–robot medical interactions: effects of transparency and adaptability, Paladyn, Journal of Behavioral Robotics, vol. 9, no. 1, pp. 95–109, 2018.
26. Li, P. Rau y Y. Li, A cross-cultural study: Effect of robot appearance and task, International Journal of Social Robotics, vol. 2, no. 2, pp. 175–186, 2010.
27. J. Winston, B. Strange, J. O'Doherty y R. Dolan, Automatic and intentional brain responses during evaluation of trustworthiness of faces, Nature neuroscience, 5(3):277–283, 2002.
28. J. X. a. E. Montague, Working with an invisible active user: Understanding, Interacting with computers, 25(5):375–385, 2013.
29. K. Leichtenstern, N. Bee, E. André y U. Berkmüller, Physiological measurement of trust-related behavior in trust-neutral and trust-critical situations, International Conference on Trust Management, pages 165–172, Springer, 2011.
30. T. Nomuraand y S. Takagi, Exploring effects of educational backgrounds and gender in human-robot interaction, International Conference on user science and engineering, pp. 24–29, 2011.
31. S. Soroka, J. Helliwell y R. Johnston, Measuring and modelling trust, Diversity,social capital and the welfar estate, pp. 279-303, 2003.
32. Ajenaghughrure, S. Sousa, I. Kosunen y D. Lamas, Predictive model to assess user trust: a psycho-physiological approach, 2019.
33. K. Akash, W. Hu, N. Jain y T. Reid, A Classification Model for Sensing Human Trust in Machines Using EEG and GSR, ACM Transactions on Interactive Intelligent Systems 8, pp. 1-20, 2018.
34. W. Hu, K. Akash, N. Jain y T. Reid, Real-time sensing of trust in human-machine interactions, IFAC-PapersOnLine, 49,pp. 48–53, 2016.
35. Ajenaghughrure, S. Sousa y D. Lamas, Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used, Multimodal Technol. Interact., 4, 63, 2020.
36. C. Boudreau, M. McCubbins y S. Coulson, Knowing when to trust others: An ERP study of decision making after receiving information from unknown people, Social cognitive and affective neuroscience, vol. 4, no. 1, pp. 23–34, 2008.
37. S. C. Jacob, R. Friedman, J. Parker, G. Tofler, A. Jimenez, J. Muller, H. Benson y P. Stone, Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research, American heart journal, vol. 128, no. 6, pp. 1170–1177, 1994.
38. X. J. Montague, Understanding active and passive users: The effects of an active user using normal, hard and unreliable technologies on user assessment of trust in technology and co-user, Applied Ergonomics, 2011.
39. E. Montague, J. Xu y E. Chiou, Shared Experiences of Technology and Trust: An Experimental Study of Physiological Compliance Between Active and Passive Users in Technology-Mediated Collaborative Encounters, IEEE Transactions on Human-Machine Systems, vol. 44, no. 5, pp. 614-624, 2014.
40. M. Kret, A. Fischer y C. De Dreu, Pupil mimicry correlates with trust in in-group partners with dilating pupils, Psychological science, vol. 26, no. 9, pp. 1401–1410, 2015.

41. G. M. a. K. Lohan, Using Pupil Diameter to Measure Cognitive Load, arXiv preprint arXiv:1812.07653, 2018.
42. M. Ahmad, J. Bernotat, K. Lohan y F. Eyssel, Trust and Cognitive Load During Human-Robot Interaction, arXiv preprint arXiv:1909.05160, 2019.