Bridging Multilevel Time Scales in HRI: An Analysis Framework

THIBAULT ASSELBORN, CHILI Lab, EPFL, Switzerland
KSHITIJ SHARMA, NTNU, Trondheim, Norway
WAFA JOHAL, CHILI Lab and BIOROB Lab, EPFL, Switzerland
PIERRE DILLENBOURG, CHILI Lab, EPFL, Switzerland

In this article, we present a multi-level time scales framework for the analysis of human-robot interaction (HRI). Such a framework allows HRI scientists to model the inter-relation between measures and factors of an experiment. Our final goal with the introduction of this framework is to unify scientific practice in the HRI community for better reproducibility. Our new approach transposes Newell's framework of human actions to model human-robot interaction. Measures from the interaction are sorted into categories (time scales) corresponding to the temporal constraints proposed by Newell. According to this sorting, a bottom-up or top-down analysis can then be performed to correlate variables which allows a better understanding and explanation of the interaction. The utilization of our method within two experimental use cases is then presented. The first one, a child-robot interaction, involves two robots and one child playing a memory game. The second is based on an analysis of the PInSoRo dataset, involving 30 child-robot pairs in a freeplay interaction. Finally, we introduce clear guidelines to re-use the framework.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**; • **Computer systems organization** → *Robotics*;

Additional Key Words and Phrases: HRI analysis framework, guideline for variables sorting, multi-level time scale, Newell's time scale, research reproducibility

ACM Reference format:
Thibault Asselborn, Kshitij Sharma, Wafa Johal, and Pierre Dillenbourg. 2019. Bridging Multilevel Time Scales in HRI: An Analysis Framework. *ACM Trans. Hum.-Robot Interact.* 8, 3, Article 17 (August 2019), 24 pages.
https://doi.org/10.1145/3338809

1 INTRODUCTION

In the field of social human-robot interaction (HRI) there are many different metrics that are used to analyze the quality of the interaction that can be found in just as many studies [1]. Even though many of these metrics are not comparable between studies, we observe that the HRI research community is starting to seek reproducibility, which manifests as a consensus that has begun to appear concerning common measures that can be used across a wide range of studies [2, 3].

Authors' addresses: T. Asselborn, W. Johal, and P. Dillenbourg, EPFL-CHILI, RLC D1 740 (Rolex Learning Center), Station 20, CH-1015 Lausanne, Switzerland; emails: {thibault.asselborn, wafa.johal, pierre.dillenbourg}@epfl.ch; K. Sharma, Department of computer science, Norwegian University of Science and Technology, NO-7491 Trondheim, Norway; email: kshitij.sharma@ntnu.no.

2573-9522/2019/08-ART17
https://doi.org/10.1145/3338809

In social HRI, the evaluation of the quality of an interaction is complex because it is very task dependent. For example, a companion robot would have as a goal to be more socially accepted and engaging while the goal of a robot used for educational purposes would probably be to maximize the learning gain of children. Besides, Human-Robot Interaction is multimodal, hence, an extensive number of HRI studies use multiple metrics (and sometimes their inter-dependencies) in order to analyze the interaction. For instance, several authors tried to find correlations between non-verbal behaviors or personality questionnaires and engagement [4–6]. Even though most HRI studies report the use of several metrics, the analysis is often limited to the correlations or the causalities between one condition and one metric or between two metrics without a rigorous methodology. The scope of our work is to identify and design a clear scientific method for the analysis of human-robot interaction.

We propose a framework based on Newell's time scale [7] and the work of Anderson [8] (see Section 3) to order variables with the objective of better analyzing and understanding how multimodal interactions take place between the robot and the user, rather than just analyzing the effect of a factor on several independent metrics. One goal could also be to use objective measures of behavior as a proxy for higher level subjective judgments. The big question is whether these are disconnected or linked. For example, we could ask: Is there a link between engagement and body movements?

We propose a systematic method that will build a model of the interaction including the (statistical) interactions effects. In line with the open science and open data trend, our framework should allow better reproducibility and better integration of results from various studies. This interaction model will allow researchers to have a global picture of the interaction as a whole but also to focus on particular relations between variables. The main contribution we propose is a framework to analyze multi-modal data at different temporal granularities in social HRI scenarios. To do so, the following steps will be followed:

(1) we present the theoretical aspects of the framework;
(2) we present two case experiments with three case studies as an example to show how to analyze the data using the proposed framework; and
(3) we present guidelines to apply the framework to other use cases.

2 EXPERIMENTAL MEASURES FOR HRI

Experimental designs and measures in HRI are becoming a field of research on their own, as proven by many summer schools, workshops and special sessions dedicated to the domain that focuses on the experimentation and evaluation of interactive robotic systems.[1] This phenomenon is driven by the lack of common benchmarks and standardized metrics to evaluate robotic systems and the quality of the interactions with the users. Indeed, user experience is central to validate the credibility and acceptability of a system. However, long-term social human-robot interaction is difficult to set up for a complex robotic system. Measuring the interaction quality in HRI, especially for social HRI, is essential but measures used are often very context dependent. From an epistemological point of view, as a new field of research, HRI has to develop strong metrics in order to guarantee its reproducibility and secure the findings of the domain. Of course, depending on the type of evaluation (online survey, large-scale experiment, case study, or longitudinal analyses) the metric used in HRI can vary. But some methodologies can be applied to ensure a common ground of knowledge.

[1]We can cite for the summer 2015: SMART-LABEX on Computational Social and Behavioural Sciences, IROS2015 Open forum on evaluation of results, replication of experiments, and benchmarking in robotics research, IEEE RAS Summer School on Experimental Methodology, Performance Evaluation and Benchmarking in Robotics, IROS2015 Workshop on Designing and Evaluating Social Robots for Public Settings, ...

Table 1. Common Techniques Used Based on [10, 11]

	Measures	Aim	Timing	Benefits	Drawbacks	Examples
Questionnaires	Likert-like series of items	Make the user express their feelings	Pre/post experiment	Easy to integrate	Adaption to answer experiments hypothesis might be difficult	NARS [12], RAS [13], GodSpeed [14], IoS, COIRS [15], …
Interviews	Open questions	Make the user express its feeling	Pre/post experiment	Specific to experiment	Difficult to integrate and time-consuming	
Task-related	Performances	Measure the influence of the system on the user's performances	Pre/post experiment	Specific to experiment, brings quantitative measures	Might not provide social insight	Response time, scores, …
Observation	Behavior	Observation of behavioral changes	During experiment	Specific to experiment, brings quantitative measures	Time-consuming	Occurrence or duration of behavior, …
Physiological measures	User's body reaction	Measure physiological reaction	During experiment	Allow to explore the user's subconscious reactions	Might be invasive and hard to interpret	Heartbeat, pupil size, eye tracking, …

According to Bethel [9], there are five primary methods of evaluation used in HRI: self-assessment, interviews, observational or behavioral measures, psychophysiology measures, and task performance metrics. In line with this work, Weiss [10] proposed the Usability, Social acceptance User experience and Societal impact (USUS) evaluation framework. The USUS framework gives methodological guidelines according to the research objectives that are aimed to measure usability, social acceptance, user experience, and societal impact. The USUS evaluation framework also provides indicators for each of the research questions and the associated methods of evaluation (i.e., expert evaluation, user studies, questionnaires, physiological measures, focus groups, and interviews). Be they in laboratory, field study, or Wizard-of-Oz experiments, the community often proposes scenario-based experimental protocols. Bethel and Murphy [9] recommends using at least three forms of evaluation in order to have reliable results for an experiment. We propose to group these categories and to give some examples of measures used in HRI.

Based on the classification from Johal [11], Weiss et al. [10], and Bethel and Murphy [9], we summarize the most common metrics used in social HRI in Table 1. This table presents several techniques used in the HRI community to asses the impact or perceived impact of a robot's interaction on the user. This impact can vary in terms of nature as it can have a social or a performance impact. Strategies of measures would also depend on the public that is involved in the experiment (children, elderly, persons with dementia, etc.). Indeed, techniques such as self-assessment are not well suited for young children or people with dementia.

There exist several classical **questionnaires** used in HRI usually administrated before or after the experiment; they aim at assessing the user's perception of the system or interaction in terms of social acceptance, utility, and societal impact [10]. According to Bethel and Murphy [9], however, it can be difficult to see attitude change through questionnaires since the self-reporting is done post-interaction. Also, self-reporting and interviews do not allow one to assess subconscious attitudinal changes and can be less accurate with children [16, 17]. The impact of the robot's attitude can be subtle and hence other types of measure are necessary.

Task-related measures are more and more common in HRI [9], as they allow one to assess the usability of the robotic system when the user is accomplishing a task. In robotics for education, for instance, this method is the most frequently used. It allows one to evaluate the students' performance according to the level of assistance of a robot [18, 19]. [20] proposed a series of task-related metrics such as effectiveness in the task, the neglect tolerance, the robot attention demand, the free time the user has, and so on. These metrics are accurate in measuring the usability and ease of use of a robotic system in a task but do not fit measure in the social aspect of human-robot interaction such as social preferences of the user. Studies usually report scores, errors, and the time of responses as collected data. Task-performance evaluations are well suited for robots in a teacher or coach role. However, it does not provide information about social bound (attachment) of the user with the robot.

Several measures have been developed to measure attitude change or physiological effects of the robot's interaction with humans in order to investigate the user experience. These measures are either observed and manually annotated from video or audio recordings, for instance, or automatically computed from the collected data. This data deals often with social signal processing that aims to inform about the emotions and the social relationship of individuals. Signals commonly used are heterogeneous and take into account behavioral cues from voice, posture, gaze, interpersonal distance, gestures, and so on. Hence, some social signals have been used to measure the quality of interaction with robots. Non-verbal communication signals are a part of these signals and have received great interest in research these past years.

Observational measures can be very accurate but are fastidious to obtain if the number of participants is high. It requires annotation guidelines and often several annotators. These measures are often used in case studies or long-term interactions when the number of session or participants is restrained. Also, if the aim is to make realtime recognition of attitude, for instance, observational data has to be made computational so that the robot can be autonomous.

Behavioral measures, such as engagement [21–23] or proxemics as a proxy for immediacy [24–26], can be computed via computer vision techniques. Some studies have focused on measuring attention of the user during the interaction. For instance, [27] uses gaze to assess attention in human-robot interaction. A lot of studies have worked on gaze and facial expression to determine engagement.

Physiological measures, such as heart rate, body temperature, skin conductance, or eye gaze [28–31], are collected and computationally treated. These measures are interesting as they reveal biological responses but are often noisy and difficult to interpret.

Finally, most of the works have combined different modalities to measure the quality of interaction [32–34], by combining contextual information, questionnaire, and autonomized behavioral or physiological measures. However, the analyses reported in these works are often limited to evaluating the correlation between dependent and independent variables regardless of the temporal granularity of the metrics used.

3 USING NEWELL'S TIME SCALE TO MODEL INTERACTIONS

As explained before, our aim in this article is to present a new framework based on Newell's time scales. Newell and Card [7] argued that there exist different time scales of human actions and proposed a logarithmic scale to identify human actions based on their duration (in seconds). Table 2 shows the complete scale of human actions presented by Newell and Card in [7]. Newell and Card argued that there are two distinct visions of HCI processes. The first vision saw HCI as an evaluative, experimental and predictive field, while the second vision saw it as a more psychological and explanatory field. They further argued that neither of the visions were sufficient to make contributions to HCI based on four main arguments:

Table 2. Newell's Time Scale of Human Action

Time scale (secs)	Human action	Newell's band	Associated theories
10^7 (months)	Technology	Social	Social and Organizational
10^6 (weeks)	Design	Social	Social and Organizational
10^5 (days)	Task	Social	Social and Organizational
10^4 (hours)	Task	Rational	Bounded rationality
10^3 (10 minute)	Task	Rational	Bounded rationality
10^2 (minutes)	Task	Rational	Bounded rationality
10^1 (10 seconds)	Unit Task	Cognitive	Psychology
10^0 (second)	Operations	Cognitive	Psychology
10^{-1} (100 ms)	Deliberate act	Cognitive	Psychology
10^{-2} (10 ms)	Neural Circuit	Biological	Neural and biochemical
10^{-3} (1 ms)	Neuron	Biological	Neural and biochemical
10^{-4} (μs)	Organelle	Biological	Neural and biochemical

(1) Emphasis is put on low-level features and actions such as key strokes, whereas the main problems involve multiple task organization and dialog with the system. Hence, there is a need to understand the compatibility of such interactions (low level) with the design of the entire system.
(2) The study is limited to the duration of the interaction, that is to say, we only consider the data collected during the interactive task and tend to neglect other aspects such as visual field, language, probability of errors, and preferences of users. If one does not consider these aspects, the contributions are too limited to yield theoretical forms.
(3) Most often, by the time research is carried out, technology becomes obsolete. This makes it difficult to make a theoretical contribution.
(4) The gap between the experiment and the application is often too wide.

To overcome the aforementioned problems, Newell and Card proposed the time scales of human action. This scale is divided into four different bands, based on the theories they support or are supported by.

— The first band that is roughly below a few milliseconds, "biological band" is governed by natural laws, neurological signals, or other biochemical signals. Theories from Physics, Chemistry, and Biology provide the corresponding analysis framework.
— The second band is for "psychological" theories dealing with the short-term tasks, such as symbolic processing (e.g., memorization) or mental mechanics (e.g, turn taking game) that usually take from a few fractions of a second to a few seconds. Most of the analytic frameworks are driven by the laws of information processing architectures.
— The third band is "bounded rationality," where the short-term and goal-oriented tasks take place, such as, skill acquisition or debugging a program, for example, which could take from a few minutes to a few days. In this band, the means to ends theory largely governs the analysis, that is to say, the human actions are bounded by the knowledge and the computational capabilities of a user to address a given problem/task.
— The fourth band is for "social and organizational" theories, dealing with culture, development, and education; these changes usually take weeks to decades in terms of duration to take effect. Human interactions plays a strong role in this band of actions and the Statistical laws and aggregate effects drive most of the analysis.

Furthermore, the scales can also be differentiated on the basis of human actions, such as technology change (months), design (weeks), goal-oriented tasks (a few minutes to days), portions of the task or unit task (a few tens of seconds), individual operations (a few seconds), deliberate act (a few hundreds of milliseconds), and biological signals (less than 10 milliseconds).

This model proposes to solve the basic problems mentioned above in the following ways:

(1) Overcoming the emphasis on the low level: the time duration of HCI is proposed in the cognitive and rational bands. Using memory, reasoning, and encodings (cognitive band) the user performs multiple tasks and acquires/practices certain skills or creates the long-term memory (rational band). Newell and Card proposed that connecting these two bands would be essential to understand what helps users and how.
(2) Overcoming limited scope: the proposed framework based on the different time scales allows one to combine a wide range of physio-psychological phenomena in a single framework, which can later be used to model the cognition (learning, skill acquisition) in terms of perception, operations, and motor constraints. The framework also provides a way to model the low-level perceptual effects (contrasts, color effects, motion) happening with large time durations, such as semantics and pragmatics of screen space during the whole task. Moreover, the framework also allows us to model the cognitive skill acquisition at two different levels: long-term memory (declarative, what) and skill acquisition (procedural, how). These two happen at different time scales in the human behavior.
(3) Overcoming obsoleteness: Newell and Card argue that human-computer interaction happens using a few devices used by the masses. By modeling the fast-paced behavior to understand the slow changing technology and design, one can start imposing similar constraints on the different interfaces.
(4) Overcoming application gap: Newell and Card proposed to look for the "ripe" (for science) domains to apply the results from experiments such as intelligent systems that interact with the user with a narrow subject matter, systems which can monitor the user and provide guidelines to them by analyzing their behavioral patterns at the different time scales.

One of the criticisms of Newell's time scales, Bruer [35] argued that the link between education and neuronal activities is a "bridge too far." Bruer suggested a link between cognitive psychology and lower bands of Newell and another link between the upper bands and cognitive psychology. In other terms, Bruer suggested to use cognitive psychology as an anchor between the different bands of human actions. However, one of the main loopholes of this criticism, as described by Anderson [8], is what Bruer considers as cognitive psychology that resides at the upper bounds of "biological band," the whole "cognitive band," and lower bounds of "rational band." Hence, in a large effect, the argument of Bruer [35] is based on the plausibility of bridging the lower level of human actions (biological signals) to the social level constructs. In his contribution, Anderson [8] provides a framework about how it could be done.

Anderson [8] proposed four kinds of bridges to connect the different time scales suggested by Newell [7]. These bridges are identified based on how many orders of magnitude, of human action duration, they connect. For example, a "long bridge" connects three orders of magnitude (e.g., task in lower bound of rational band to cognitive operations). A "longer bridge" connects four order of magnitudes or more (e.g., task in lower bound of rational band to deliberate acts). Similarly, a "short bridge" connects two orders of magnitude (e.g., task in lower bound of rational band to unit tasks). Using these different kinds of bridges, Anderson showed that it is plausible to relate the higher-level constructs to the lower-level data sources. It is important to notice that analyzing variables connected by a bridge with high orders of magnitude (e.g., longer bridge with variables

separated by six orders of magnitude) might bring a situation where the criticism raised by Bruer makes sense. It is then more cautious to constrain the analysis to variables separated by a relatively small order of magnitude.

A question one might raise could be about the need for such model (typically based on psychological theories) in a human-robot interaction context. The answer could be based on the argument that Newell and Card [7] make regarding the role of psychological theories in human-computer interaction. Newell and Card [7] quote an article from 1984 Human Factor Society Bulletin:

> *"Many computer system designers appear to have no knowledge of human factors, are not aware that the human-computer interface is vital to their systems, or that a substantial human-factors database exists to help them."* - [36]

This could be said for the designers and developers of robotics systems these days, however in a less strict manner than what Muckler [36] indicated three decades ago for the computer system designers. Nowadays there is little knowledge of human factors that is taken into account while designing robotic systems. However, this cannot be said about analytical frameworks in HRI. In light of this observation, we propose to use Newell's time scales to analyze variables defined at the different temporal granularities. These variables could be defined based on a given theoretical aspect related to the given human action band.

In human-computer interaction (HCI) situations, Newell's human action bands have been used by a multitude of researchers to inform system design and analysis. For example, Nielsen and Molich [37] used the action bands to provide different heuristics for small-scale evaluation of user interfaces. In their book "Usability Engineering," Rosson and Carroll [38] cite action bands to support their claim about theory-oriented predictive models; while Carroll [39] uses the action bands to inform the scenario-based design practices in HCI. Lewis et al. [40] used the action bands to inform the design and testing of "off-the-shelf" interfaces. In recent times, Newell's bands have been used to study the interplay between designers and users [41], evaluating web-interfaces [42, 43], driver training systems [44, 45], electrical safety interfaces [46], and safety interfaces [47].

We propose that the results from a small bridge are easier to understand than a long bridge, hence the preference for choosing a long/short/longer bridge should depend on who (robot or human) is adapting in the interaction. For example, a robot can adapt to high-level signals (e.g., trust) by learning low-level signals (e.g., motor signals, gaze); but a human might need a higher level feedback (in the cognitive band of Newell and Card [7]) to adapt. Therefore, the choice of a bridge is of utmost importance in this framework. In the use case presented in the next section, we will show with two examples what could influence this choice.

4 USE CASE 1: THE MEMORY GAME

The memory game is a simple game involving two or more players. These players are facing a set of matching pairs that are randomly disposed face-hidden. Alternatively, players turn over two cards in an attempt to find and collect a matching pair. At the end of the game (when all the pairs are collected), the winner is the one that has collected the highest number of pairs.

The memory game was chosen due to the simplicity of its rules, its short duration, as well as the fact that it is a competitive game that pushes the engagement of the child in the task.

It was decided to limit the number of cards to 16 (predisposed in a 4 by 4 grid) due to the young age of the children participating in this experiment. The game was implemented on a tablet running Android.

Simple artificial intelligence was implemented for the robots' strategy. The robots would forget the cards flipped during a given round R_i according to a probability that depended on the number

Fig. 1. A and B: The two Nao robots playing the memory game against the child. One of them stays static while the other performs adaptor movements during idle moments. C: cameras are facing child used to extract various features from the child's face. D: The tablet implementing the memory game.

of rounds played between the current round R and R_i. The probability to forget a card on R_i is given by the formula $P(i) = 1 - 1/(R_i - R)$ (with i the index of the round).

4.1 Experimental Setup

The experiment was conducted in a school in France where 20 children (12 girls, and 8 boys) aged 5 years old participated. Children played the memory game against two identical NAO robots placed in front of them (see Figure 1). In addition to being physically identical (to remove any bias, the two robots were of same color and gender), the robots behaved the same way. This meant that they acted the same way when returning a card, exhibited similar behaviors when winning and losing and employed the same artificial intelligence–based decision-making process throughout the game. The head movements were also implemented on both robots as a lot of studies showed their importance on human perception [48, 49]. The only difference was that while one robot (called static robot) was completely static during idle moments, the other (called adaptor robot) would display specific movements, called adaptor movements, during this period.

4.2 Adaptor Movements

Non-verbal movements were categorized into five groups [50]: (1) *emblems*, (2) *illustrators*, (3) *affect display*, (4) *regulators*, and (5) *adaptors*. The first four are communicative movements used with a semantic goal as opposed to the last one. Adaptor movements are postural or non-verbal movements that are often performed during idle moments. In robotics, it is commonly admitted that these micro-movements can help the interaction appear more credible [51–53].

A database from Asselborn et al. [51] containing more than 60 animations that mimic adaptor movement was used. These animations were sorted on three different levels (namely, *low*, *medium*, and *high*) according to their intensity (see Section 4.4).

During a game, every child played three rounds, each time with a randomly assigned level of adaptor movement.

4.3 Case Studies

In the present experiment, we collected data from a multitude of sources such as the children's performance, their perceived anthropomorphism and proficiency toward robots, the game state, children's attention and emotions, and robot's animation type and animation intensity (see Section 4.4 for more details). To provide an example that shows *how to use Newell's action scales to analyze data from an HRI scenario* using the data acquired, we will present two case studies with different streams of data that belong to the different levels of human (or in our case, robot) actions.

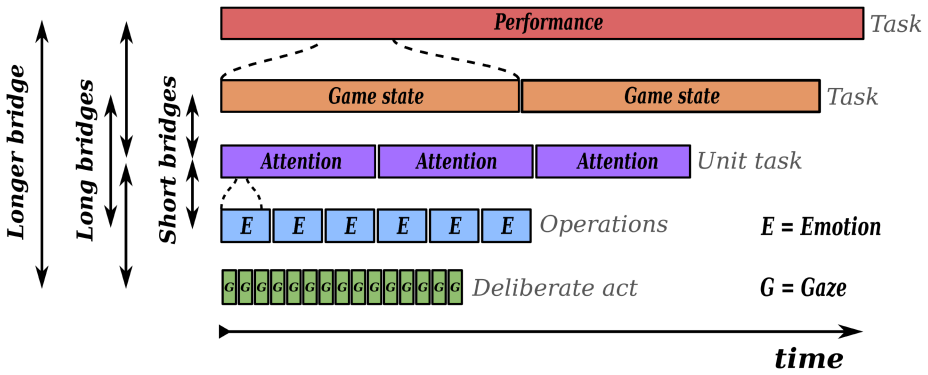

Fig. 2. Measures for the "memory" case study sorted according to Newell's time scale. Short bridges connect variables located at juxtaposed levels while long bridges connect variables that are not directly juxtaposed (connecting three orders of magnitude). Longer bridges connect four orders of magnitude or more.

(1) Case 1: the first set of variables that we analyzed consisted of the "game state," "children's attention," and their "emotions." Their respective placement in Newell's time scales and the bridges used to connect them are shown in Figure 2. We will use two short bridges (game state–attention and attention–emotion) and one long bridge (game state–attention–emotion). The details of the measures are given in Section 4.4.
(2) Case 2: the second set of variables that we analyzed consisted of children's "performance," "attention," and "gaze." Their respective placement in Newell's time scales and the bridges used to connect them are also shown in Figure 2. We will use two bridges (performance–attention and attention–gaze), and one longer bridge (performance–attention–gaze) in this case. The details of the measures are provided in Section 4.4.

4.4 Measures

Game Id. Every child was asked to play three rounds, every time with a robot displaying a different level of adaptor movement.

Perceived Anthropomorphism and Proficiency. Items from the GodSpeed questionnaire [54] were adapted to measure the anthropomorphic differences between the two robots. Children had to compare the two robots on four different metrics: **humanity**, **friendliness**, **proficiency**, and **attention** through simple questions like "Which robot was the most friendly?" and had to answer on a 5-point Likert scale (for sure Robot 1, probably Robot 1, same, probably Robot 2, for sure Robot 2). The questionnaire used can be found online.[2] The questions were asked at the end of every game (so three times per experiment, each time for a different adaptor level (see Section 4.2)). The four metrics used can be found in the original godspeed questionnaire [54] and were chosen due to our interest in the fluidity of the interaction (humanity and friendliness metrics) as well as the consequences of these movements on the cognitive appearance of the robots (proficiency and attention).

Performance. The number of matching pairs collected at the end of the memory game by the child was used as a measure of performance. A median cut was done on the average number of cards collected during a game by the children. If during a game, the child collected more cards than this value, his/her performance was labeled as **high**, otherwise, it was labeled as **low**. As the

[2] https://github.com/asselbor/Questionnaire.

Table 3. Weighting Factor Assigned to the Different Joints of Nao to Compute the Intensity I of the Different Animations

Ankle	Elbow	Hand	Head	Hip	Knee	Wrist
3	2	1	3	4	4	1

children played three games, three measures of performance are available for the analysis for each child.

Game State. The number of cards collected was also used during this analysis. For every child, 24 logs of this type was recorded as the children played three games where eight pairs of cards needed to be discovered.

Child's Attention. The child's attention was also measured. Per period of 10 seconds, the child was labeled as **attentive** or **non-attentive**. This updating period of 10 seconds in the measure of attention appears reasonable in the light of the literature that is interested in the duration of children's attention span [55, 56]. To measure their attention from the videos, the following procedure was applied by two coders: videos were cut in segments of 10 seconds and treated independently. 25% of these video segments were annotated by both and used to compute the inter-rater reliability for these segments, which was found to be 84%. The remaining segments (75% of the videos) were split into two groups. The first group was annotated by coder #1 and the second by coder #2.

Animation's Type and Intensity. Different types of animations were launched throughout the game. Both robots launched animations expressing an emotion when losing or winning, animations when returning the cards, and animations when thinking. Finally, the adaptor robot launched animations at regular intervals displaying adaptor movements. All these animations might have a different impact on the child depending on its **type** (*happy animations sad animations*, *returning card animation*, *thinking animation*, and *adaptor animation*) but also on its intensity level (low, medium, high) (see Section 4.2). That is why, the intensity of each single animation was logged. On average, an animation with a different type and intensity was launched every 5s.

In order to measure this intensity, the product of three variables was used: the angle swept by joints during the animation, the angular speed of these joints, as well as the duration of the animation. In addition, as some joints have a smaller impact on the motion (hands, fingers) than others (knees, shoulders), a weighting factor was also introduced (see Table 3) to take that into account in the intensity calculation. This is based on a method commonly done to compute the quantity of motion [57]. Concretely, the following equation was used to compute the intensity I of a given animation:

$$I = t_m * \sum_{i=1}^{n} c_i * \alpha_i * v_i \tag{1}$$

with:

— c = weighting factor depending on the specific joint concerned (see Table 3),
— t_m = Total duration of animation,
— α = Angle swept during animation,
— v = Mean joint velocity while moving,
— i = The joint id starting from the first one to the last one n.

Child's Emotions. A camera facing the child (see Figure 1) was used to extract the emotions from the face of children. To do so, we used the Facial Action Units (FAUs) [58] that have become a

Table 4. Metrics Measured During the "Memory" Interaction Sorted According to Newell's Time Scale

Metric	Time scale	Newell's time scale
Game ID	3 times	Task
Questionnaire	3 times	Task
Performance	3 times	Task
Game state	24 times∗	Task
Child's attention	every 10s	Unit task
Animation type	every 5s	Unit task
Animation intensity	every 5s	Operations
Emotion	every 1.5s	Operations
Gaze	every 0.5s	Deliberate act

∗The child plays 3 games where 8 pairs of cards need to be discovered.

standard to systematically categorize the physical expression of emotions [59]. The work done by [60] was used to extract in real time the action units from the face of the child. Finally, in order to make the link between the FAUs and the emotions (namely, **Happiness**, **Sadness**, **Surprise**, **Fear**, and **Anger**), the model presented by [61] was used. The logs concerning these emotions were updated every 1.5s.

Child's Gaze. A second camera facing the child (see Figure 1) was used for gaze tracking and head pose estimation. This information was used to extract the localization in space where the child was looking and especially whether he/she was looking at the robot displaying adaptor movement, the static robot, or the tablet. The FeatureFace library[3] (based on Openface [60]) was used for this purpose. The logs concerning this visual attention were updated every 0.5s.

A summary of all the measures and their association on their Newell's time scale can be found in Table 4.

4.5 Results and Implications

In this section, we present the results from the two case studies mentioned in Section 4.3.

4.5.1 Case 1: Game State, Attention, and Emotions. The first set of measures considered comprises the game state, children's attention, and emotions. Their respective positions on Newell's action scale is shown in Figure 2.

Short Bridge 1: Game State and Attention. In order to find a relation between attention and game state, a logistic model was created according to the following equation:

$$log\left(odds\left(\frac{Attention}{No_attention}\right)\right) = \alpha \cdot \sum_{n=1}^{6} \beta_{1,2*n} \cdot Number\_of\_cards\_collected(=2*n).$$

The child's attention seems to fluctuate with the game state (number of cards collected during the game). Intuitively, we could think that the child's attention would decrease with the progress of the game. We can see in Figure 3 that the reality is more complex. At the beginning of the game, the level of attention decreases (between two and four cards collected) probably because of the novelty effect that vanishes. Then the level of attention seems to fluctuate (as can be seen in Table 5) until 12 cards have been collected in the game. Finally, the level of attention decreases

[3] https://github.com/asselbor/features_face.

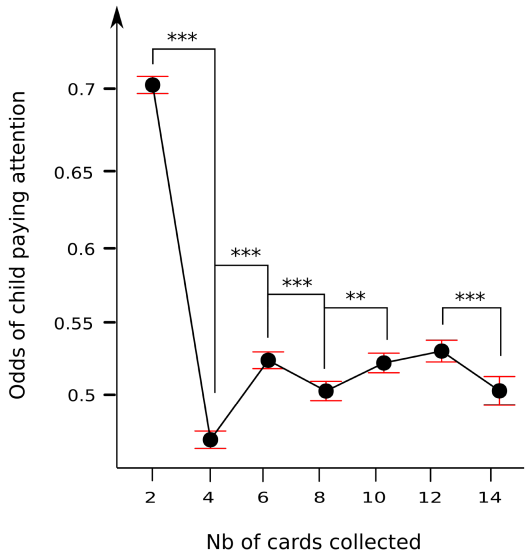

Fig. 3. Odds of a child paying attention given the number of cards collected in the game; the "*" represents the significance level of the difference between the two consecutive numbers of cards collected. p values: "*"<0.05; "**"<0.01; "***"<0.001.

Table 5. Logistic Model for Attention Levels ("Attention" and "No Attention") Given the Number of Cards Collected in the Game

Number of cards collected	Estimate	Variance of estimate	p value
Intercept	α: 0.01	0.02	0.58
2	$\beta_{1,1}$: 0.11	0.02	0.0001
4	$\beta_{1,2}$: 0.07	0.02	0.98
6	$\beta_{1,3}$: −0.005	0.02	0.0002
8	$\beta_{1,4}$: 0.08	0.02	0.0001
10	$\beta_{1,5}$: −0.13	0.02	0.0001
12	$\beta_{1,6}$: 0.84	0.02	0.0001

again probably because the child does not have to be attentive anymore because he/she certainly already has the knowledge of the cards' positions.

Short Bridge 2: Attention and Emotion. In order to find relations between attention and emotions, a logistic model was created according to the following equation.

$$log\left(odds\left(\frac{Attention}{No_attention}\right)\right) = \alpha + \beta_1 \cdot joy + \beta_2 \cdot sadness + \beta_3 \cdot surprise + \beta_4 \cdot anger.$$

We can see in Table 6 that the child's attention level appears to be positively correlated with surprise and anger, and negatively correlated with fear, sadness, and joy. One plausible explanation concerning the positive correlation between surprise and attention is that the surprise can result in a saliency effect [62–64] that increases with the level of attention. Indeed, the more you are focused on the game (high level of attention) the more something (e.g., a robot moving) can be a source of surprise.

Table 6. Logistic Model to Predict the Attention Level ("Attention" or "No Attention") Given the Emotions

Variable	Estimate	Variance of Estimate	p value
intercept	α: −0.30	0.01	<0.0001
joy	β_0: −0.13	0.004	<0.0001
sadness	β_1: −0.33	0.007	<0.0001
surprise	β_2: 0.57	0.008	<0.0001
anger	β_3: 0.41	0.007	<0.0001
fear	β_4: −0.38	0.01	<0.0001

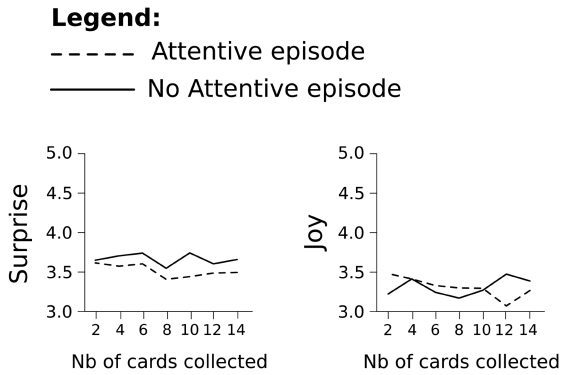

Fig. 4. Interaction effect of the state and attention episodes ("attention" and "no attention") for the surprise and joy emotions.

To explain the positive correlation of anger with attention, we should keep in mind that the majority of time, children were losing in the game. One of our hypotheses is that the more the child is focused on the game, the more they will care about the result and the fact that the robot is winning most of the rounds therefore generates anger. The same explanation can be made for joy, which can be interpreted as the opposite of anger.

Concerning the negative correlation between sadness and the level of attention, one possible explanation could be the following. When the child is losing, two paths are open to them. The first outcome is that they become angry (which means that they are focused on the game as shown above) or become sad. Alternatively, the child has already given up on the game, therefore justifying the low levels of attention.

Finally, we can explain the negative correlation of fear with attention by the fact that if the child is afraid, it means that their attention is not only focused on the game but also on things (e.g., robots) that produced this emotion.

Of course, it appears clear that additional research/data would need to be conducted/collected in order to support the assumptions described above.

Long Bridge: Game State, Attention, and Emotion. Figure 4 shows the interaction effect for emotions given the two other variables, namely, the child's attention episodes ("attention" and "no attention") and the game state (number of cards collected). An interaction effect can be observed for joy (see Table 7). For example, at the beginning of the game (until the number of collected cards reaches 10), when the child has an attentive episode, they are more joyful than when they have a non-attentive one; this relation inverts once the number of collected cards is above 10 (as seen

Table 7. ANOVA Results: Different Attention Levels ("Attention" and "No Attention") and State for Emotions

Emotion		State	Attention	State::Attention
Joy	F	38.68	57.93	813.74
	p value	<0.001	<0.001	<0.001
Surprise	F	501.52	2,026.48	266.08
	p value	<0.001	<0.001	<0.001

Table 8. One-Way ANOVA Results for Attention Levels Given the Performance, without Assuming Equal Variances

variable	df1	df2	F	p
attention	1	25.36	2.72	0.11
no attention	1	23.01	4.39	0.04

in Figure 4). One possible explanation could be that toward the end of the game, during attentive episodes (when they are focused on the game), they realize the fact that they are losing (taking into account that kids lost the vast majority of the time) leading to a decrease in joy. On the other hand, joy increases during a non-attentive episode as a possible result of relief.

Furthermore, concerning the surprise (see Figure 4), we can see that the intensity of this emotion decreases with the number of cards collected and after a while (when eight cards are collected) it remains nearly constant for attentive episodes but fluctuates for non-attentive ones. These results could be explained by the fact that surprise occurs more easily during non-attentive episodes (if the user is being attentive, then they are only surprised at the beginning, with the novelty effect).

4.5.2 Case 2: Performance, Attention, and Gaze. The second set of measures considered comprises the children's performance, attention, and emotions. Their respective positions on Newell's action scale is shown in Figure 2.

Long Bridge 1: Performance and Attention. The first relation analyzed is between the attention and the performance of the children through a one-way ANOVA. Table 8 shows no difference for attention across the different performance levels, whereas there is a significant difference for the absence of attention across the different performance levels. As can be seen in Figure 5 (right), not paying attention will possibly lead to low performance. To resume, while paying attention might not ensure high performance, not paying attention, on the other hand, guarantees low performance. Since the task was a memory game, where the intermediate steps for winning the game consisted in finding two matching tiles and remembering their positions on the grid, it was necessary for the kids to be attentive and remain focused in order to win and avoid giving the robots the upper hand.

Long Bridge 2: Attention and Gaze. We observe no significant relationship between the different targets (where the kids were supposed to look) and the attention levels ("Attention" and "No attention"). One plausible explanation for the absence of any relation between these two variables could be the fact that looking at a certain object (or area in the field of view) does not always imply that the person (in our case, kids) is paying attention [65].

Longer Bridge: Performance, Attention, and Gaze. A logistic model was created according to the following to study the interaction effect of gaze and attention on the performance levels ("high"

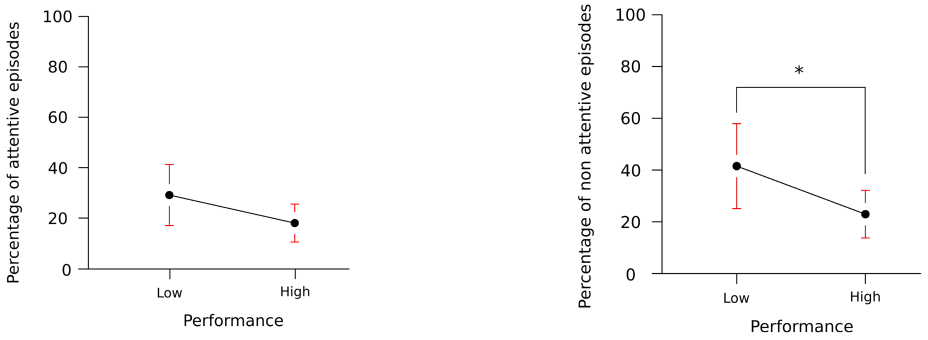

Fig. 5. Left: Percentage of "attentive" episodes for the two levels of performance. Right: Percentage of "non-attentive" episodes for the two levels of performance. The red bars show the 95% confidence interval.

Table 9. Logistic Regression Model for the Performance Levels ("High" or "Low") Given the Attention Levels and the Gaze Targets

Attentional targets	Estimate	Error	p value
intercept	α: −1.78	0.08	0.0001
Static_robot	$\beta_{1,1}$: 0.07	0.12	0.53
tablet	$\beta_{1,2}$: 0.23	0.10	0.03
No_attention	$\beta_{2,1}$: 0.04	0.10	0.69
Static_robot::No_attention	$\beta_{3,1}$: −0.14	0.16	0.39
Tablet::No_attention	$\beta_{3,2}$: −0.39	0.14	0.005

and "low"; see Section 4.4 for details). By observing Table 9, we can see that if the child is looking at the tablet and not paying attention the probability of winning the game (having a high performance) significantly decreases. This confirms and consolidates the results from the "long bridge 1" (Section 4.5.2).

$$log\left(odds\left(\frac{High\ Performance}{Low\ Performance}\right)\right) = \alpha + \beta_{1,1} \cdot gaze\_target(= static\_robot) + \beta_{1,2} \cdot gaze\_target(= tablet) + \beta_{2,1} \cdot attention(= no\_attention) + \beta_{3,1} \cdot gaze\_target(= static\_robot) : attention(= no\_attention) + \beta_{3,2} \cdot gaze\_target(= tablet) : attention(= no\_attention).$$

4.5.3 Implications of Using the Framework. In the two case studies, we presented typical examples of using the proposed framework. Case 1 analyzed two short bridges and one long bridge. Provided that, the two short bridges give insight into two relations: (1) game state and attention and (2) attention and emotions. Thanks to the analysis of the long bridge, we could test our hypothesis about the interaction of these three variables. Finding out the relations between the attention levels and game state changes based on the emotions would have not been possible without the long bridge. This supports our initial recommendation about the choice of long/short bridge. In

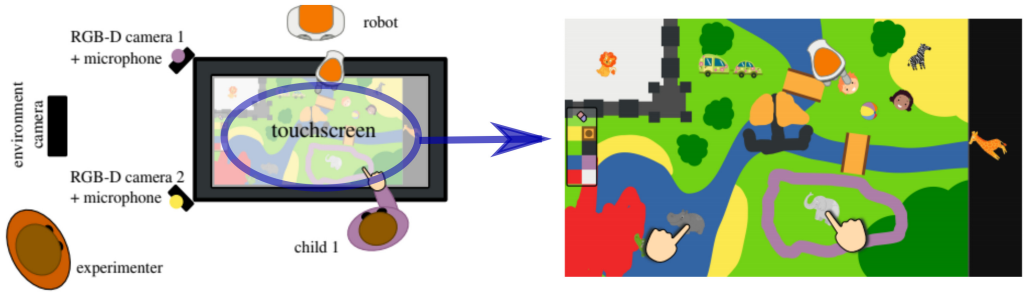

Fig. 6. Left: The free-play social interaction sandbox used in the Pinsoro study. One child and one robot interacted in a free-play situation, by drawing and manipulating items on a touchscreen. Right: Example of a possible game situation. These images directly come from [66].

this case, if we would have considered only two short bridges, we would have obtained a result that did not correctly explain the relationship between the variables.

Case 2 analyzed two long bridges and a longer bridge. The results from the long bridge 2 did not reveal any relation between the gaze and attention levels. Similar to Case 1, implementing a longer bridge provided a deeper understanding of the multilevel relation between performance, attention, and gaze.

In a nutshell, the framework based on the works of Newell and Card [7] and Anderson [8], inspires researchers to hypothesize the relations between the multilevel temporal variables that are separated by several orders of magnitude (gaze and performance are separated by three orders of magnitude in time scales), which are not always intuitive. Thus, using this framework might provide additional (sometimes non/counter-intuitive) insight into the data collected from an experiment.

5 CASE STUDY 2: PINROSO

The Pinsoro dataset [66] comprises more than 45 hours of hand-coded recordings of social interactions between 45 child-child pairs and 30 child-robot pairs. Only the child-robot pairs are considered in this study. To build this dataset, Lemaignan et al. captured a rich set of behavioral patterns occurring in natural social interactions between a child and a NAO robot. The interaction is based on free (participants are not directed to perform any particular task beyond playing) and playful interactions in a "sandbox environment" taking place on a large horizontal touchscreen (see Figure 6).

The following variables were extracted during each interaction.

5.1 Measures

Task Engagement. Concerning the task engagement, Lemaignan et al. distinguished four categories: "no play," "goal oriented," "aimless," "adult seeking." A clear distinction between "on-task" and "off-task" behaviors was done by the annotators. "On-task" behaviors are called "goal oriented": they encompassed considered, planned actions (that might be social or not). "Off task" called "aimless" behaviors encompassed opposite behaviors: "being silly, chatting about unrelated matters, having a good laugh, and so on." These data were updated every 10 seconds and fits the unit task in Newel's time scale.

Emotion. In the same way as in the memory game case study, a camera facing the child was used to extract emotions from the child's faces. To do so, the FAUs were once again used [58] to

Table 10. Metrics Measured During the "Pinsoro" Interaction Sorted According to Newell's Time Scale

Metric	Time scale	Newell's time scale
Task engagement	every 10s	Unit task
Emotion	every 1.5s	Operations
Motion Intensity	0.5s	Deliberate act

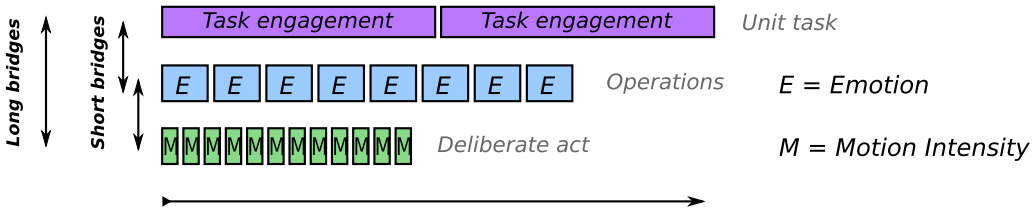

Fig. 7. Measures for the "Pinsoro" case study sorted according to Newell's time scale. Short bridges are connecting variables located at juxtaposed levels while long bridges are connecting variables that are not directly juxtaposed (connecting three orders of magnitude).

extract emotions (namely, **Happiness**, **Sadness**, **Surprise**, **Fear**, and **Anger**) (see Section 4.4 for more details). The logs concerning these emotions were updated every 1.5s.

Motion Intensity. The motion intensity is defined as the mean of the optical flow computed from the camera facing the child (as seen in Figure 6). The logs concerning the motion intensity were averaged to get a measure each 0.5s.

A summary of all the measures and their association on Newell's time scale can be found in Table 10.

5.2 Case 1: Task Engagement, Emotions, and Motion Intensity

The set of measures considered comprises the task engagement, motion intensity, as well as emotions. Their respective position on Newell's time scale is presented in Figure 7.

Short Bridge 1: Task Engagement and Emotion. In order to find the relation between the task engagement (*On play* vs *Off play*) and the child's emotion, a logistic model was created according to the following equation:

$$log\left(odds\left(\frac{OnPlay}{OffPlay}\right)\right) = \alpha + \beta_1 \cdot joy + \beta_2 \cdot anger + \beta_3 \cdot sadness + \beta_4 \cdot surprise + \beta_5 \cdot fear.$$

As can be seen in Table 11, *On play* task engagement seems to be correlated with the level of surprise the child is expressing. Intuitively, this can be explained in the same way as in the memory case study: the more the child is focused on the game (*OnPlay* engagement), the more something (e.g., the robot doing an action within the game) can be a source of surprise for the child.

Short Bridge 2: Emotion and Motion Intensity. Linear regression models were created to predict the emotion based on the intensity of motion (one different model per emotion). The formula can be described as follows:

$$Emotion = \alpha \cdot Motionintensity + \varepsilon.$$

Table 11. Details of the Logistic Model for Task Engagement (On Play vs Off Play) as Modeled by the Emotions

	Estimate	Error	p value
Intercept	α: −3.24	2.00	0.10
Joy	β_1: −8.25	6.64	0.21
Anger	β_2: 3.54	7.32	0.62
Sadness	β_3: 5.36	8.76	0.54
Surprise	β_4: 23.04	10.73	**0.03**
Fear	β_5: −11.41	11.54	0.32

Table 12. Details of the Linear Regression Models for the Emotions as Modeled by the Motion Intensity

Emotion	Estimate (α)	Intercept (ε)	p value
Joy	−0.0069	0.0394	**<0.001**
Surprise	0.0097	0.1249	**<0.001**
Sadness	0.0088	0.1607	**<0.001**
Anger	−0.0002	0.1896	0.83
Fear	0.0082	0.1964	**<0.001**

As can be seen in Table 12, all emotions except anger are significantly correlated with the level of motion intensity. This result appears to be in line with a lot of research showing the link existing between emotion and motion [67, 68].

Long Bridge: Task Engagement, Emotions, and Motion Intensity. In order to find the relation between the task engagement (*On play* vs *Off play*), the child's emotions and the intensity of motion, a logistic model with interaction was created according to the following equation:

$$log\left(odds\left(\frac{OnPlay}{OffPlay}\right)\right) = \alpha + \sum \beta_i * emotion_i + \delta * motionIntensity + \sum \gamma_i * emotion_i * motionIntensity.$$

As can be seen in Table 13, the interaction between the level of emotion concerning the sadness and the child's intensity of motion is marginally significant. In other words, and as can be seen in Figure 8, the less the child is feeling sad, the more a high motion intensity becomes a good predictor of task engagement (*On-task* engagement).

5.3 Implication of Using the Framework

In this case study, we analyzed two short bridges and one long bridge. The two short bridges focused on the relations between (1) task engagement and emotion and (2) emotion and motion intensity. The analysis of the long bridge allowed us to extract the interaction between these three variables. In particular, finding out that the relation between the task engagement and the sadness emotion changed based on the value of the motion intensity was only possible thanks to the analysis of the long bridge. As seen, a different interpretation could have been drawn if the analysis of these three variables was done only with the two short bridges.

6 DISCUSSION

In this section, we present the framework and explain how someone willing to run an HRI experiment could use the multi-level time scale representation in order to better understand the

Table 13. Details of the Logistic Model for Task Engagement (On Play vs Off Play) as Modeled by the Emotions

	Estimate	Error	p value
Intercept	α: 29.23	18.30	0.11
Joy	β_1: −152.82	100.75	0.12
Anger	β_2: 52.24	51.18	0.30
Sadness	β_3: −172.30	104.47	**0.09**
Surprise	β_4: −88.37	75.94	0.24
Fear	β_5: 95.73	82.69	0.24
Motion intensity	δ: −1.83	1.03	**0.07**
Joy:Motion intensity	γ_1: 8.18	5.64	0.14
Anger:Motion intensity	γ_2: −3.25	3.13	0.30
Sadness:Motion intensity	γ_3: 10.05	5.88	**0.08**
Surprise:Motion intensity	γ_4: 5.76	4.18	0.16
Fear:Motion intensity	γ_5: −5.48	4.41	0.21

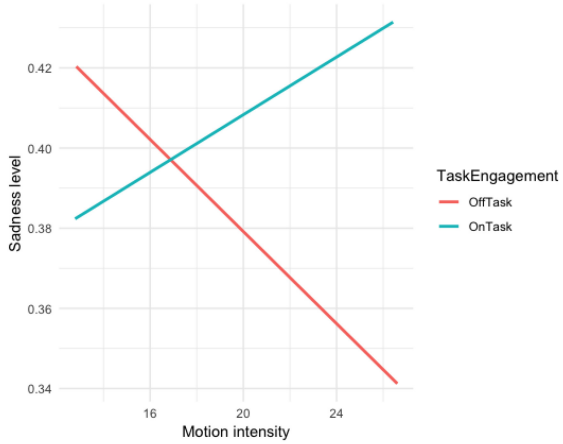

Fig. 8. Interaction effect of Sadness levels and motion intensity for the different levels of task engagement.

experimental variable inter-dependencies. The proposed framework could also allow the HRI community to unify the way results are presented and to ensure better re-usability between studies.

Below, we explain the three steps of the framework to guide the experimenter.

Record Multi-Modal Data. Quite often in HRI studies, experimenters acquire multi-modal data. Being, interactions logs, gaze data, emotion estimation, speech, or data from other types of sensors, these data should have an associated timestamp to allow multi-level time scales analysis.

Fit Your Measures into the Time Scale. Depending on the way the recordings are used to compute features, the same raw data can fall into various time scales. Figure 9 shows an example of how a given data source (log of facial expression) can be used to define variables to be interpreted in function of the time scale used [69]. Table 14 shows examples of metrics that can be extracted from different time scales.

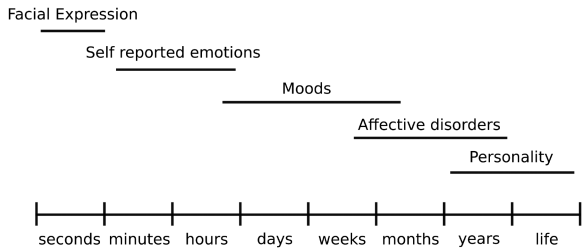

Fig. 9. Different interpretation of facial expression logs, depending on the time scale used.

Table 14. Levels for HRI Time Scale

Time scale	HRI time scale	Example of metric
>session duration	Social	Acceptance, Trust, Skills
3–15min	Task	Answers to a quiz, a questionnaire, making a cocktail, task performances
≈1min	Unit task	Robot's plan state, pour cocktail in cup, achieve a sub-goal
sec	Operations	Move a cube, draw a letter, robot gives instructions
<sec	Deliberate act	Gaze, touch sensor activation

In the use cases presented above, the video recordings allowed the computation of emotions and gaze, for instance. Finally, it is important to notice that a minimum of three variables disposed at different levels on Newell's time scale is necessary for the analysis with the presented framework.

Build Short, Long, and Longer Bridges. The experimental hypothesis can serve as the basis to check for correlations between variables, which is the method classically used in experimental data analysis. However, with the framework presented in this article, the user will also be able to specifically check for correlations between variables at different time scales. Indeed, the last step results in analyzing interactions between measures at different time scales by exploring their possible link with the help of bridges (short bridge, long bridge, and longer bridge).

7 CONCLUSION

The multi-level time scales framework for analysis of human-robot interaction presented in this article aims to better understand how the multi-modal interaction takes place between the robot(s) and the human. As seen in the case studies presented, some relations sometimes appear to be counter-intuitive at first but can explained and better understood using this analysis framework.

However, the proposed framework presents some limitations. First of all, as all variables need to be timestamped and pre-processed (features computed according to meaningful timestamps), preparing and running such kind of multi-level analyses might take more time and more skills than a regular "traditional" analysis. Moreover, the framework will be only valid with at least three variables disposed at different levels of Newell's time scale. Finally, it is important to notice that the framework was tested in the specific case of one user interacting with several robots. The model of the interaction might thus be more complicated when being used for experiments involving several users at the same time. In such cases, however, we believe that one can simply consider using the same proposed method to analyze the interaction between humans.

Since we have a multi-level comprehension of the relations among the variables using this framework, we can use this understanding to provide more relevant feedback to the robot(s). Therefore, this framework might also allow us to more easily bridge toward autonomous robots by categorizing actions and interactions' information in a real time manner. In addition, the model

can also be used to check for eventual bias in an experiment (e.g., if two variables have no meaningful relation, then it may indicate that there is an experimental design bias). Such a framework also allows researchers to build a model of the interaction which can then improve the theoretical knowledge of the interactions between human(s) and robot(s).

Finally, the proposed framework can also help researchers better structure their experiments by providing a common ground and guidelines showing how to choose and sort the variables describing their interaction. In addition, it will allow them to better reuse previous research with the access of the explicit time frequencies of the variables extracted during the interaction.

REFERENCES

[1] Robin R. Murphy and Debra Schreckenghost. 2013. Survey of metrics for human-robot interaction. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI'13)*. IEEE, 197–198.

[2] Aaron Steinfeld, Terrence Fong, David Kaber, Michael Lewis, Jean Scholtz, Alan Schultz, and Michael Goodrich. 2006. Common metrics for human-robot interaction. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 33–40.

[3] Dimitrios Chrysostomou, Paolo Barattini, Johan Kildal, Yue Wang, Jacopo Fo, Kerstin Dautenhahn, Francois Ferland, Adriana Tapus, and Gurvinder S. Virk. 2017. ReHRI'17 - Towards reproducible HRI experiments: Scientific endeavors, benchmarking and standardization. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. ACM, New York, NY, 421–422. DOI: http://dx.doi.org/10.1145/3029798.3029800

[4] Ginevra Castellano, Iolanda Leite, Andre Pereira, Carlos Martinho, Ana Paiva, and Peter W. McOwan. 2012. Detecting engagement in HRI: An exploration of social and task-based context. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT'12) and 2012 International Conference on Social Computing (SocialCom'12)*. IEEE, 421–428.

[5] Salvatore M. Anzalone, Sofiane Boucenna, Serena Ivaldi, and Mohamed Chetouani. 2015. Evaluating the engagement with social robots. *International Journal of Social Robotics* 7, 4 (2015), 465–478.

[6] Jennifer Goetz and Sara Kiesler. 2002. Cooperation with a robotic assistant. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 578–579.

[7] Allen Newell and Stuart K. Card. 1985. The prospects for psychological science in human-computer interaction. *Human-Computer Interaction* 1, 3 (1985), 209–242.

[8] John R. Anderson. 2002. Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science* 26, 1 (2002), 85–112.

[9] Cindy L. Bethel and Robin R. Murphy. 2010. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics* 2, 4 (July 2010), 347–359. DOI: http://dx.doi.org/10.1007/s12369-010-0064-9

[10] Astrid Weiss, Regina Bernhaupt, and Manfred Tscheligi. 2011. The USUS evaluation framework for user-centered HRI. In *New Frontiers in Human-Robot Interaction*. 89–110.

[11] Wafa Johal. 2015. *Companion Robots Behaving with Style: Towards Plasticity in Social Human-Robot Interaction*. Thesis. Université Grenoble Alpes. https://tel.archives-ouvertes.fr/tel-01679314.

[12] Dag Sverre Syrdal, Kerstin Dautenhahn, K. Koay, and M. L. Walters. 2009. The negative attitudes towards robots scale and reactions to robot behavior in a live human-robot interaction study. In *Adaptive and Emergent Behaviour and Complex Systems*. 109–115. http://uhra.herts.ac.uk/handle/2299/9641.

[13] Tatsuya Nomura, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. 2008. Prediction of human behavior in human–robot interaction using psychological scales for anxiety and negative attitudes toward robots. *IEEE Transactions on Robotics* 24, 2 (April 2008), 442–451. DOI: http://dx.doi.org/10.1109/TRO.2007.914004

[14] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. 2008. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. *Metrics for HRI Workshop, Technical Report* (2008). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.2966&rep=rep1&type=pdf#page=43file:///Users/johal/Library/ApplicationSupport/MendeleyDesktop/Downloaded/Bartneck,Kulic,Croft-2005-Measuringtheanthropomorphism,animacy,likeability.

[15] David Robert and Victor Van Den Bergh. 2014. Children's openness to interacting with a robot scale (COIRS). (2014).

[16] Robert J. Fisher. 1993. Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research* 20, 2 (1993), 303–315. http://www.jstor.org/stable/2489277.

[17] Tony Belpaeme, Paul Baxter, Joachim de Greeff, James Kennedy, Robin Read, Rosemarijn Looije, Mark Neerincx, Ilaria Baroni, and Mattia Coti Zelati. 2013. Child-robot interaction: Perspectives and challenges. In *Social Robotics*, Guido Herrmann, Martin J. Pearson, Alexander Lenz, Paul Bremner, Adam Spiers, and Ute Leonards (Eds.). Springer International Publishing, Cham, 452–459.

[18] Daniel Leyzberg, Samuel Spaulding, and Brian Scassellati. 2014. Personalizing robot tutors to individuals' learning differences. *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction (HRI'14)*, 423–430. DOI: http://dx.doi.org/10.1145/2559636.2559671

[19] Deanna Hood, Séverin Lemaignan, and Pierre Dillenbourg. 2015. The CoWriter project: Teaching a robot how to write. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts*. ACM, 269–269.

[20] D. R. Olsen and M. A. Goodrich. 2003. Metrics for evaluating human-robot interactions. *Proceedings of PERMIS* 2003, 4 (2003), 507–527. DOI: http://dx.doi.org/10.1016/j.intcom.2005.10.004

[21] Ginevra Castellano, André Pereira, Iolanda Leite, Ana Paiva, and Peter W. Mcowan. 2009. Detecting user engagement with a robot companion using task and social interaction-based features interaction scenario. *Interfaces* (2009), 119. DOI: http://dx.doi.org/10.1145/1647314.1647336

[22] Iolanda Leite, Marissa McCoy, Daniel Ullman, Nicole Salomons, and Brian Scassellati. 2015. Comparing models of disengagement in individual and group interactions. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 99–105.

[23] Jyotirmay Sanghvi, Ginevra Castellano, Iolanda Leite, André Pereira, Peter W McOwan, and Ana Paiva. 2011. Automatic analysis of affective postures and body motion to detect engagement with a game companion. In *Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction (HRI'11)*. IEEE, 305–311.

[24] Ross Mead, Amin Atrash, and Maja J. Matarić. 2011. Proxemic feature recognition for interactive robots: Automating metrics from the social sciences. (2011), 52–61. file:///Users/johal/Library/ApplicationSupport/MendeleyDesktop/Downloaded/Mead,Atrash,Matari?-2011-ProxemicFeatureRecognitionforInteractiveRobotsAutomatingMetricsfromtheSocialSciences.pdf.

[25] Ross Mead, Amin Atrash, and Maja J. Matarić. 2013. Automated proxemic feature extraction and behavior recognition: Applications in human-robot interaction. *International Journal of Social Robotics* 5, 3 (May 2013), 367–378. DOI: http://dx.doi.org/10.1007/s12369-013-0189-8

[26] Dominique Vaufreydaz, Wafa Johal, and Claudine Combe. 2016. Starting engagement detection towards a companion robot using multimodal features. *Robotics and Autonomous Systems* 75, PA (2016), 4–16. DOI: http://dx.doi.org/10.1016/j.robot.2015.01.004

[27] Maria Staudte and Matthew W. Crocker. 2011. Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition* 120, 2 (2011), 268–291. DOI: http://dx.doi.org/10.1016/j.cognition.2011.05.005

[28] Pramila Rani and Nilanjan Sarkar. 2003. Operator engagement detection for robot behavior adaptation. *Advanced Robotic* (2003), 1–12. file:///Users/johal/Library/ApplicationSupport/MendeleyDesktop/Downloaded/Rani,Sarkar-2003-OperatorEngagementDetectionforRobotBehaviorAdaptation.pdf.

[29] D. Kulic and E. Croft. 2005. Anxiety detection during human-robot interaction. In *Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 616–621. DOI: http://dx.doi.org/10.1109/IROS.2005.1545012

[30] Amirhossein H. Memar and Ehsan T. Esfahani. 2018. Physiological measures for human performance analysis in human-robot teamwork: Case of tele-exploration. *IEEE Access* 6 (2018), 3694–3705.

[31] Maria Staudte and Matthew W. Crocker. 2009. Visual attention in spoken human-robot interaction. In *Proceedings of the 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI'09)*. IEEE, 77–84.

[32] Ross Mead and Maja J. Matari. [n.d.]. Toward robot adaptation of human speech and gesture parameters in a unified framework of proxemics and multimodal communication. ([n.d.]).

[33] S. Al Moubayed, M. Baklouti, M. Chetouani, T. Dutoit, and A. Mahdhaoui. [n.d.]. Multimodal feedback from robots and agents in a storytelling experiment. ([n.d.]), 43–55. file:///Users/johal/Library/ApplicationSupport/MendeleyDesktop/Downloaded/Moubayedetal.-Unknown-MultimodalFeedbackfromRobotsandAgentsinaStorytellingExperiment.pdf.

[34] Salvatore M. Anzalone, Serena Ivaldi, Olivier Sigaud, and Mohamed Chetouani. 2013. Multimodal people engagement with iCub. *Advances in Intelligent Systems and Computing* 196 AISC, Figure 1 (2013), 59–64. DOI: http://dx.doi.org/10.1007/978-3-642-34274-5_16

[35] John T. Bruer. 1997. Education and the brain: A bridge too far. *Educational Researcher* 26, 8 (1997), 4–16.

[36] F. Muckler. 1984. The future of human factors. *Human Factors Society Bulletin* 27, 2 (1984), 1.

[37] Jakob Nielsen and Rolf Molich. 1990. Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 249–256.

[38] Mary Beth Rosson and John Millar Carroll. 2002. *Usability Engineering: Scenario-Based Development of Human-Computer Interaction*. Morgan Kaufmann.

[39] John M. Carroll. 2000. *Making Use: Scenario-Based Design of Human-Computer Interactions*. MIT Press.

[40] Clayton Lewis, Peter G. Polson, Cathleen Wharton, and John Rieman. 1990. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 235–242.

[41] Gerhard Fischer. 2017. Exploring richer ecologies between designers and users. In *Conversations Around Semiotic Engineering*. Springer, 21–29.
[42] Fatima Isiaka. 2017. *Modelling Stress Levels Based on Physiological Responses to Web Contents*. Ph.D. Dissertation. Sheffield Hallam University.
[43] Ruth Chatelain-Jardon, Jesus S. Carmona, and Ned Kock. 2016. An extension to simulated web-based threats and their impact on knowledge communication effectiveness. *International Journal of Technology and Human Interaction (IJTHI)* 12, 3 (2016), 64–77.
[44] Lisa Dorn. 2017. *Driver Behaviour and Training*. Vol. 2. Routledge.
[45] Ghasan Bhatti, Roland Brémond, Jean-Pierre Jessel, Nguyen-Thong Dang, Fabrice Vienne, and Guillaume Millet. 2015. Design and evaluation of a user-centered interface to model scenarios on driving simulators. *Transportation Research Part C: Emerging Technologies* 50 (2015), 3–12.
[46] Molla Ramizur Rahman. 2018. Understanding the human-computer interaction behavior in electrical and power systems. In *Industrial Safety Management*. Springer, 143–151.
[47] Christian Reuter and Marc-André Kaufhold. 2018. Usable safety engineering sicherheitskritischer interaktiver systeme. In *Sicherheitskritische Mensch-Computer-Interaktion*. Springer, 17–40.
[48] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 25–32.
[49] Emily Wang, Constantine Lignos, Ashish Vatsal, and Brian Scassellati. 2006. Effects of head movement on perceptions of humanoid robot behavior. In *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 180–185.
[50] Paul Ekman and Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1, 1 (1969), 49–98.
[51] Thibault Lucien Christian Asselborn, Wafa Johal, and Pierre Dillenbourg. 2017. Keep on moving! Exploring anthropomorphic effects of motion during idle moments. In *Proceedings of the 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'17)*.
[52] Raquel Ros, Marco Nalin, Rachel Wood, Paul Baxter, Rosemarijn Looije, Yannis Demiris, Tony Belpaeme, Alessio Giusti, and Clara Pozzi. 2011. Child-robot interaction in the wild: Advice to the aspiring experimenter. In *Proceedings of the 13th International Conference on Multimodal Interfaces*. ACM, 335–342.
[53] David Cameron, Samuel Fernando, Emily Collins, Abigail Millings, Roger Moore, Amanda Sharkey, Vanessa Evers, and Tony Prescott. 2015. Presence of life-like robot expressions influences children's enjoyment of human-robot interactions in the field. In *Proceedings of the AISB Convention 2015*. The Society for the Study of Artificial Intelligence and Simulation of Behaviour.
[54] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1, 1 (2009), 71–81.
[55] Karen Wilson and James H. Korn. 2007. Attention during lectures: Beyond ten minutes. *Teaching of Psychology* 34, 2 (2007), 85–89.
[56] Holly A. Ruff and Katharine R. Lawson. 1990. Development of sustained, focused attention in young children during free play. *Developmental Psychology* 26, 1 (1990), 85.
[57] Caroline Larboulette and Sylvie Gibet. 2015. A review of computable expressive descriptors of human motion. In *Proceedings of the 2nd International Workshop on Movement and Computing*. ACM, 21–28.
[58] Paul Ekman and Wallace V. Friesen. 1977. Facial action coding system.
[59] Paul Ekman and Erika L. Rosenberg. 1997. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS)*. Oxford University Press.
[60] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: An open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision (WACV'16)*. IEEE, 1–10.
[61] A Freitas-Magalhães. 2012. Microexpression and macroexpression. *Encyclopedia of Human Behavior* 2 (2012), 173–183.
[62] Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20, 11 (1998), 1254–1259.
[63] Derrick Parkhurst, Klinton Law, and Ernst Niebur. 2002. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* 42, 1 (2002), 107–123.
[64] Shelley E. Taylor and Susan T. Fiske. 1978. Salience, attention, and attribution: Top of the head phenomena. *Advances in Experimental Social Psychology* 11 (1978), 249–288.
[65] Kshitij Sharma, Patrick Jermann, and Pierre Dillenbourg. 2014. "With-me-ness": A gaze-measure for students' attention in MOOCs. In *Proceedings of International Conference of the Learning Sciences 2014*. ISLS, 1017–1022.
[66] Séverin Lemaignan, Charlotte E. R. Edmunds, Emmanuel Senft, and Tony Belpaeme. 2018. The PInSoRo dataset: Supporting the data-driven study of child-child and child-robot social dynamics. *PloS One* 13, 10 (2018), e0205999.

[67] Emma Bould and Neil Morris. 2008. Role of motion signals in recognizing subtle facial expressions of emotion. *British Journal of Psychology* 99, 2 (2008), 167–189.
[68] Gwen Littlewort, Jacob Whitehill, Ting-Fan Wu, Nicholas Butko, Paul Ruvolo, Javier Movellan, and Marian Bartlett. 2011. The motion in emotion—A CERT based approach to the FERA emotion challenge. In *Face and Gesture 2011*. IEEE, 897–902.
[69] Keith Oatley, Dacher Keltner, and Jennifer M. Jenkins. 2006. *Understanding Emotions*. Blackwell Publishing.

Received February 2018; revised March 2019; accepted May 2019