Automatic analysis of trust over the course of a human-robot interaction using multimodal features and recurrent neural architectures

Marc Hulcelle

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2023IPPAT043

Thèse de doctorat

Automatic analysis of trust over the course of a human-robot interaction using multimodal features and recurrent neural architectures

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 École doctorale de l'Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Signal, images, automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 18 Décembre 2023, par

MARC HULCELLE

Composition du Jury :

Catherine Pelachaud Directrice de recherche CNRS, Sorbonne Université (ISIR)	Présidente / Examinatrice
Karola Pitsch Professeure, University of Duisburg-Essen (Institute of Communication Studies)	Rapportrice
Alexandre Pauchet Professeur, INSA Rouen (LITIS)	Rapporteur
Brian Ravenet Maître de Conférences, Université Paris Saclay (LISN)	Examinateur
Chloé Clavel Professeure, Télécom Paris (LTCI)	Directrice de thèse
Giovanna Varni Professeure, University of Trento (DISI)	Co-encadrante de thèse
Nicolas Rollet Maître de Conférences, Télécom Paris (i3)	Invité

Acknowledgments

I would like to start by thanking my thesis supervisors Chloé Clavel, Giovanna Varni, and Nicolas Rollet for making this thesis happen. You truly enriched my scientific knowledge, and pushed me to always be rigorous in my work. During these three intense years, you deeply influenced my vision of academia, the world of research, and what it means to be a researcher. Thank you for your patience, your communication, and always making me strive for the best. I want to thank everyone in the S2A team (research and administrative team) of Telecom Paris for the many different exchange, and their constant cheerfulness. I had the chance to meet many different personalities and would like to take the time to thank them for the time spent talking about their views on research, academia, how frustrating reviewer 2 can be, and the many discussions on movies, books, music, and weird philosophical ideas (Aina, Anas, Arturo, David, Dimitri, Émile, Félix, James, Joël, Junjie, Mathilde, Nathan).

I want to give my special thanks to my friends, the "Gem of Research" (Emilia, Tamim, and Jean-Remy), the "Borel team" (Anthony, Clémence, Léo, Maxence, Numa, Paul, Thomas, Richard, Victor), the "Rakoons" (Antoine, Aria, Layla, Quentin), and other friends that I have met during these past years (Alice, Bartoche, Claire, Félix, Julien, Lazare). Thank you for always being present to celebrate the good news, reminding me that there are other things in life than research, supporting me in the difficult times, your joy, and your love. Everyone of you made me realize the importance of building strong connections with people, accepting people as they truly are, and being as empathetic as you can be. Last but undoubtedly not least, I thank my parents Fabienne and Pierre, my sister Julie, my numerous cousins, my grandma, and my love Lisa for their unconditionnal support and love in this rollercoaster of adventure. I would not be here today if it were not for your presence, and the joy you bring me everyday. I have a special thought for my cat that still has to learn that night-time is not playtime.

And thank you Camus for making me imagine Sisyphus happy.

Résumé

Questions de recherche

Nous formulons les questions de recherche suivantes :

- RQ1: Quel cadre théorique est applicable dans le cadre d'une analyse automatique de la confiance conduite tout au long de l'interaction ?
- RQ2: Est-ce que des segments de confiance homogènes émergent au sein de l'interaction en se basant sur des indices comportementaux tangibles ?
- RQ3: Comment peut-on discriminer les segments de confiance de ceux de méfiance en s'appuyant sur des indices comportementaux tangibles ?

Contexte de la thèse

Historiquement, la confiance en Interaction Humain-Robot (HRI) a été déterminée comme une construction psychologique. Une des définitions les plus utilisées est celle de Rousseau, qui l'a définie comme *"un état psychologique comprenant l'intention d'accepter une vulnérabilité basée sur une attente positive vis à vis de l'intention ou comportements d'une autre personne"* [97]. Ainsi, chaque individu a sa propre représentation de son partenaire d'interaction et décide de lui faire confiance en fonction de cette représentation. Cette représentation se fonde sur des critères relatifs au robot, à l'environnement, et des critères propres à l'utilisateur-même [42]. La confiance joue ainsi un rôle fondamental dans le développement et maintien de rapport entre personnes. Comme la confiance a un impact sur l'acceptation du robot par l'utilisateur, sur l'issue et la performance de la tâche de l'interaction, il est important de la calibrer correctement [4, 50]. En effet, la sous-confiance peut mener à un refus de collaboration avec le robot, alors qu'une sur-confiance envers le robot peut aboutir à une mauvaise utilisation de celui-ci. Il est donc important de pouvoir mesurer la confiance afin de calibrer celle-ci. Celle-ci est souvent mesurée par des questionnaires (Godspeed [11], "Interpersonal Trust Scale" [96], "Trust Perception Scale-HRI" [105], "Negative Attitude towards the Robot Scale" [113]) remplis par les utilisateurs eux-mêmes en début et fin d'interaction, par des mesures "objectives" (distance au robot, EEG) ou par des mesures proxy tel que le nombre de pièces données au partenaire dans le jeu du "Dilemme du prisonnier" [58]. Les questionnaires ne permettent cependant pas de mesurer la confiance tout au long de l'interaction, puisqu'elles interrompent l'interaction et demandent un temps significatif à être remplis.

Il y a très peu d'études portant sur l'analyse automatique des dynamiques de la confiance en HRI. Parmi les quelques études existantes, Lee et. al a utilisé des modèles de Markov cachés afin

de représenter les dynamiques comportementales aboutissant à une confiance basse ou élevée dans le cadre d'un scénario du dilemme du prisonnier [58]. Khalid et al. a utilisé des modèles d'ensemble neuro-flou afin de classifier la confiance selon trois dimensions définies par Mayer: compétence, bienveillance, intégrité [49, 71].

Une approche sociologique de la confiance en HRI

Cadre théorique

Afin de pouvoir analyser la confiance tout au long de l'interaction, nous nous appuyons sur une approche moins "interne", "mentaliste" de la confiance telle qu'étudiée usuellement. Les théories de la sociologie interactioniste mettent en lumière le caractère observable de la confiance au travers de comportements rendus visibles par les participants. Nous définissons ainsi la confiance commme *"une forme d'affiliation et de crédit caractérisés par un ensemble de comportements intentionnels ou non, expressifs ou propositionnels"* [45]. Cette méthodologie permet de ne pas inférer l'état mental des utilisateurs, en se basant sur une analyse des processus interactionnels via leur comportements [35, 36, 44]. La confiance est ainsi un résultat de l'état de l'interaction, et est orientée à la fois vers le contenu et le format de l'interaction. Dans un état de confiance, les participants vont se comporter de manière à ce que l'interaction soit fluide et avance en direction de son objectif. Nous utilisons ainsi ces concepts afin d'établir notre base méthodologique d'analyse. Nous divisons l'analyse de la confiance à partir de sous-concepts qui constituent notre définition de travail, et dirigeons l'analyse de l'observateur vers des comportements qui indiquent un alignement, une affiliation [110], et ceux qui attribuent du crédit [26, 88] aux compétences [33] du robot.

Nous avons ainsi développé un schéma de codage nommé TURIN (Trust in hUman Robot INteraction), suffisamment flexible pour être utilisé à la fois en interaction dyadique ou en interaction de groupe. Nous proposons de coder la confiance en segments qui décrivent un niveau de confiance homogène, au niveau individuel pour les interactions dyadiques, au niveau du groupe pour les interactions de groupe. La segmentation débute par l'identification d'unités comportementales et l'attribution d'un niveau de confiance à celles-ci. Les unités consécutives renvoyant à une même catégorie de confiance sont ensuite agrégées pour former des segments de confiance homogènes. À chaque segment est attribué une catégorie de confiance : "confiance", "méfiance", ou "neutre" en fonction de si les utilisateurs exhibent des comportements de confiance, méfiance, ou neutres respectivement. Ainsi, tout type de comportement qui démontre une confiance interactionnelle, accepte une vulnérabilité, montre une amicalité, ou reconnaît une compétence du partenaire peut être interprété comme un comportement de confiance. Nous définissons la confiance interactionnelle comme un état de l'interaction dans lequel les participants démontrent une forme de naturalité comportementale (en traitant le robot comme un partenaire autonome de manière similaire à un humain), ou une fluidité dans l'interaction. Les comportements de méfiance correspondent à toute forme de malaise, doute, confusion, agres-

sivité, ou refus de collaboration. Les comportements neutres renvoient à tout comportement qui ne permet pas de statuer sur l'état de confiance du groupe. Nous proposons de coder au sein de ces segments quatre catégories supplémentaires afin de décrire les comportements observés : "Forme de l'interaction sociale", "Contenu de l'interaction", "Bienveillance", et "Intégrité". La catégorie "Forme de l'interaction sociale" représente tous les indices comportementaux bas-niveau qui démontrent soit un haut niveau de confiance (e.g. interaction fluide, naturelle), ou qui démontrent un bas niveau de confiance (e.g. rupture de confiance, doute). La catégorie "Contenu de l'interaction" fait référence aux évènements, comportements, et mots du processus interactionnel en cours qui signalent la catégorie de confiance. Tout comportement ou évènement qui démontrent une bienveillance ou malveillance sont décrits par la catégorie "Bienveillance". La catégorie "Intégrité" renvoie aux comportements ou évènements qui signalent soit un manque soit une marque d'intégrité de l'utilisateur. Le schéma de codage a fait l'objet d'une première validation par une collecte d'annotations sur le jeu de données Vernissage. Deux experts en HRI ont annoté 1 minute sur 3 interactions. L'accord inter-annotateur calculé sur chacune des catégories de confiance montre un accord significatif entre les annotateurs. La catégorie "méfiance" aboutit à l'accord plus élevé, suivi par la catégorie "confiance", puis "neutre". Nous montrons que l'accord entre les annotateurs est modéré sur les éléments des autres catégories, notamment du à la difficulté de correctement délimiter le début et la fin d'un comportement.

Comparaison avec l'approche mentaliste usuelle

Afin de montrer que notre approche "interactioniste" est différente et complémentaire de l'approche psychologique "mentaliste" usuelle en HRI, nous conduisons une étude de comparaison de notre méthode d'annotation avec des annotations collectées en utilisant la version réduite à 14 éléments du questionnaire de Schaefer "Robot Trust Scale" (RTS) [105]. Cinq experts en HRI ont participé à cette étude. Un expert a annoté l'entièreté du corpus Vernissage avec les deux outils. Sans chevauchement, deux experts ont annoté la moitié du corpus en utilisant TURIN, et les deux derniers experts ont chacun annoté la moitié du corpus en utilisant le RTS. Afin de pouvoir comparer les deux outils, nous avons opéré quelques modifications à leur méthodologie afin d'aboutir à un terrain commun de comparaison. Les annotations se font sur des segments de taille fixe pour chacun des deux, ce qui donne 18 segments par interaction. Les experts utilisant le RTS annotent leur ressenti selon les 14 critères vis à vis de l'interaction qu'ils observent, au vu de la réaction des participants. Nous comparons ainsi les deux approches en différentiant les segments en fonction du label de confiance assigné par les annotations de TURIN. Afin de pouvoir comparer les annotateurs, nous remettons à l'échelle les scores des critères pour chaque annotateur dans l'intervalle $[0, 1]$ en fonction de leur minimum et maximum. À travers une série de tests statistiques de Kruskal-Wallis [55] et de tests post-hoc de Dunn [25], nous montrons qu'il existe des différences significatives de distribution de scores de certains critères du RTS entre les segments annotés "confiance" et ceux annotés "méfiance", ainsi qu'entre ceux annotés "méfiance" et ceux "neutre". Nous montrons ainsi que

certains critères du RTS semblent indépendants du label de confiance assigné par les annotations TURIN. Nos calculs d'accord inter-annotateur montrent un accord modeste pour l'outil TURIN, et un accord pauvre pour le RTS. Ceci s'explique par les modifications que nous avons opérées afin de pouvoir comparer les deux outils.

À travers une comparaison théorique de l'approche mentaliste et de l'approche interactioniste, ainsi qu'à travers cette analyse expérimentale, nous identifions quatre critères qui permettent de différencier les deux approches : *temporalité* (l'intervalle temporel nécessaire à l'analyse), *orientation* (préférence pour des approches délimitées par un cadre théorique, ou pour des approches orientées-données), *capacité de généralisation* (spécificité ou généricité), *ambivalence d'analyse individuelle et de groupe* (passage d'une interaction dyadique à une interaction de groupe). Ainsi, l'approche mentaliste est plus orientée vers des cadres théoriques précis, assez générique, est ambivalente, et s'intéresse à des évolutions ayant lieu sur une interaction complète. L'approche interactioniste quant à elle est plus orientée données, plus spécifique, également ambivalente, mais prône une analyse sur des processus interactionnels courts à l'échelle du tour de parole.

Modèles computationnels de la confiance en HRI

Présentation des descripteurs

Nous nous sommes ensuite intéressés aux méthodes de modélisation computationnelle multimodale de la confiance. Notre modélisation s'opère sur le jeu de données Vernissage [47]. Le corpus est constitué de 10 interactions pendant lesquelles deux participants interagissent avec un robot Nao selon quatre phases : présentation rapide des participants, explication des tableaux exposés, présentation plus détaillée des utilisateurs, quizz artistique. Pour le reste de nos études, nous avons collecté des annotations d'un expert en HRI sur l'ensemble des interactions uniquement pour les trois premières phases comme le format de la dernière phase n'est pas pertinent par rapport à la confiance.

Nous avons dans un premier temps modélisé les dynamiques de la confiance au sein du groupe d'utilisateurs avec des descripteurs multimodaux à plusieurs échelles du groupe : individuel, dyades, et triade. Les modalités retenues sont les suivantes : le visage, le corps, la voix, ainsi que la sémantique. Nos descripteurs retenus sont un mélange de descripteurs extraits automatiquement et manuellement par des annotations. Pour le visage, nous avons extrait les unités d'action faciales (1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 20, et 23) des utilisateurs grâce au logiciel OpenFace [10] comme moyen de traiter de manière fine les émotions faciales [3, 74]. Nous avons également extrait le Focus d'Attention Visuelle (FAV) comme indices de l'alignement et de crédit accordé par les participants au robot. Les labels sont les suivants : peinture gauche, peinture centrale, peinture droite, Nao, autre humain, autre, pas clair. En plus des labels, nous avons calculé le pourcentage de temps de regard mutuel entre les participants pendant un segment ainsi que le nombre de changements de FAV par participant. Nous avons aussi

extrait les hochements de tête en tant qu'indices d'alignement et d'affiliation [8, 110, 127]. Avec l'extraction de la durée des hochements de tête, nous avons calculé le pourcentage de temps qu'un participant hoche la tête au sein d'un segment.

Pour le corps, nous n'avons pas pu utiliser de descripteurs sur le squelette pour des raisons de données trop bruitées par l'extraction. Nous avons donc utilisé le barycentre de position des utilisateurs et du robot calculé à partir des positions de leurs têtes fournies par le corpus en tant que mesure de la distance au robot [76]. Ensuite, avec OpenCV, nous avons extrait l'indice de contraction [16] de chacun des participants. Cet indice représente à quel point la posture corporelle est ouverte (e.g. bras tendus en l'air ou sur les côtés) ou fermée (e.g. personne recroquevillée).

En ce qui concerne la voix, nous avons d'abord représenté l'activité vocale des utilisateurs par un indicateur binaire selon trois labels : parole, silence, ou rire. Nous avons également inclus l'activité vocale du robot Nao, limitée quant-à-elle aux deux premiers labels. Nous avons ensuite calculé le pourcentage de temps pendant lequel le robot et un utilisateur parlent en même temps au sein d'un segment. Nous avons ensuite extrait des descripteurs GeMAPS de prosodie avec le logiciel OpenSMILE [28, 29]. Nous avons gardé parmi l'ensemble des descripteurs la F0 normalisée pour chaque participant, le volume, la gigue (variations cycle à cycle de la fréquence fondamentale), le scintillement (variations cycle à cycle de l'amplitude), ainsi que le flux spectral (différences de fenêtre à fenêtre du spectre audio), les quatre premiers coefficients cepstraux des fréquences de Mel (MFCC), ainsi que les dérivées premières de la F0 et des quatres premières MFCC [49].

Enfin, nous avons extrait une représentation sémantique du discours du robot grâce à un TinyBERT [48]. Nous avons appliqué une analyse des composants principaux pour réduire la taille de cette représentation à 50. Nous avons ainsi construit une représentation agrégée en moyennant la représentation sémantique des mots énoncés pendant un segment. Pour les segments pendant lesquels le robot ne parle pas, nous avons décidé de propager cette représentation depuis le segment précédent. Les descripteurs non-catégoriques sont agrégés au sein d'un segment par un calcul de moyenne et d'écart standard des valeurs qu'ils prennent. Nous appliquons avant cela une opération de réduction de bruit par un filtre de Savitsky-Golay. Nous obtenons ainsi un vecteur de taille 222 : 68 pour chaque utilisateur, 79 pour le robot, 3 pour chaque dyade, et 4 pour la triade.

Description des modèles

Chaque interaction i est ainsi constituée d'un vecteur de descripteurs x_j^i pour chaque segment j, avec un label associé y_j^i. Pour une première approche, nous entraînons des modèles simples de machine learning à prédire le label y_j^i étant donné le vecteur de descripteurs associé x_j^i. L'hypothèse ici est que le contexte n'est pas nécessaire pour la prédiction du label. Les classes correspondent aux catégories de confiance de TURIN. Nous formulons le problème de classification de deux manières différentes : i) une classification Un-contre-reste, ii) une classification à trois classes. Les modalités sont agrégées de deux manières différentes : fusion

immédiate, et fusion tardive. Pour le méchanisme de fusion immédiate, les descripteurs sont concaténés immédiatement avant d'être donnés en entrée du modèle. Pour la fusion tardive, nous entraînons un modèle par modalité puis faisons une moyenne des prédictions de chaque modèle avant de déterminer le label prédit. Nous entraînons donc plusieurs modèles de machine learning simples selon les modalités décrites, en ne prenant pas en compte la modalité sémantique en première approche. Nous entraînons un classificateur Ridge (CR), des Forêts Aléatoires (FA), une Machine à Support de Vecteurs en classificateur (MSV-C), et un Perceptron Multi-Couches (PMC).

Dans un deuxième temps, nous concevons plusieurs architectures neuronales récurrentes. Formellement, nous construisons une séquence $(x^i_{j-\tau}, ..., x^i_j)$, constituée du segment x^i_j dont le label y^i_j est la cible de l'entraînement, ainsi que des τ précédents segments qui constituent l'historique de l'interaction. Nous faisons une étude incrémentale où nous commençons par un premier modèle constitué simplement d'unités GRU suivi par une couche de réseaux de neurones ("fully-connected" FC). Ce modèle se nomme GRU-Simple (GS).

Nous ajoutons ensuite un premier module que nous nommons "Encodeur des Dynamiques Inter-Groupe" (EDIG). En s'engageant dans une activité de groupe impliquant une conversation, les participants organisent leurs interactions avec leurs partenaires de manière spécifique qui dépend de l'activité et de son but. Ils peuvent être vocaux en s'adressant à une partie ou l'entièreté du groupe, ou être silencieux en étant soit en écoute active en démontrant des signes d'intérêts soit en étant passifs. Il est donc important d'analyser les dynamiques entre les différents participants à différentes échelles du groupe pour comprendre les dynamiques qui le compose [36]. Ainsi, les données de chaque participant sont données en entrée d'un GRU par participant. Les sorties sont concaténées avec les données dyadiques et triadiques avant d'être données en entrée d'une couche FC. La sortie est alors concaténée avec les données du robot pour être données en entrée d'une couche GS. Ce modèle s'appelle ainsi EDIG-GS.

Finalement, nous ajoutons un dernier module nommé "GRU Interactionnel" (GI). Pendant une interaction, les participants produisent en continu des comportements sociaux. Ces comportements sont vecteurs de sens qui forment ainsi le contexte interactionnel. Les participants utilisent les tours précédents de leurs partenaires pour produire leur tour, et ainsi renouvellent le contexte à chaque tour [38]. Ceci signifie qu'il y a une structure temporelle dans l'utilisation du contexte comme ressource de l'interaction. Nous modélisons ainsi cette structure temporelle en traitant de manière différente les données issues du robot $x^i_{r,t}$ et celles issues du groupe de participants $x^i_{g,t}$, pour l'interaction i au pas de temps t. Nous avons ainsi :

$$h^i_{r,t} = GRU(x^i_{r,t} \oplus h^i_{g,t-1}) \tag{1}$$

$$h^i_{g,t} = GRU(x^i_{g,t} \oplus h^i_{r,t}) \tag{2}$$

où $h^i_{g,t}$ désigne l'état caché du groupe au pas de temps t, et $h^i_{r,t}$ désigne l'état caché du robot au même pas de temps. Nous avons modélisé la structure temporelle de cette manière afin d'insister sur le fait que le robot est le meneur de la conversation. Nous nommons l'architecture

composée des deux modules EDIG-GI. La Figure 1 montre l'architecture complète avec nos deux modules.

Expériences

Lors de l'entraînement des modèles de machine learning simples, nous utilisons une cross-validation avec trois interactions en validation pour la sélection des hyper-paramètres. Nous utilisons les métriques de ROC-AUC pour la tâche Un-contre-Reste, et un score F1 pour la classification multi-classe. étant donné l'imbalance de classes des données, nous augmentons le jeu de données avec l'algorithme SMOTE. Les Tableaux 1 et 2 regroupent les performances. À travers une série de tests de Kruskal-Wallis et de tests post-hoc de Dunn, nous montrons que les Forêts Aléatoires ont de meilleures performance à la fois en fusion immédiate et fusion tardive dans la classification Un-contre-Reste. Une série de tests de Wilcoxon-Mann-Whitney montre que les meilleurs résultats sont obtenus par mécanisme de fusion immédiate. Pour la classification multi-classe, nous montrons que les Forêts Aléatoires et PMC obtiennent les meilleures performances en fusion immédiate, alors que les FA et le MSV-C obtiennent les meilleurs performances en fusion tardive. Les tests statistiques indiquent que l'ensemble des modèles obtiennent de meilleures performance avec le mécanisme de fusion immédiate, indiquant des interactions entre les descripteurs que le mécanisme de fusion tardive n'arrive pas à capturer.

Fusion immédiate	**FA**	**PMC**	**MSV-C**
Confiance-vs-reste	**0.72** ±0.04	0.70 ±0.04	**0.74** ±0.04
Neutre-vs-reste	**0.77** ±0.04	0.74 ±0.04	0.75 ±0.04
Méfiance-vs-reste	**0.59** ±0.06	0.54 ±0.07	**0.58** ±0.06
Fusion tardive			
Confiance-vs-reste	**0.67** ±0.04	0.60 ±0.04	**0.66** ±0.04
Neutre-vs-reste	**0.74** ±0.04	0.65 ±0.04	0.70 ±0.03
Méfiance-vs-reste	**0.54** ±0.08	0.48 ±0.08	0.49 ±0.10

Table 1: Scores ROC-AUC de test pour les classificateurs en Un-contre-Reste.

Fusion	**Aléat.**	**Maj.**	**FA**	**PMC**	**MSV-C**
Immédiate	0.38 ±0.03	0.52 ±0.05	**0.66** ±0.04	**0.65** ±0.04	0.60 ±0.04
Tardive	0.38 ±0.03	0.52 ±0.05	**0.62** ±0.03	0.60 ±0.04	**0.61** ±0.05

Table 2: Scores F1 de test pour la classification multi-classe.
Aléat.: Classificateur Aléatoire. Maj.: Classificateur classe majoritaire.

Pour l'entraînement des modèles récurrents, nous faisons une recherche d'hyper paramètres avec une interaction de validation, et gardons une interaction en tant que jeu de test. Nous choisissons le score F1 ainsi que la précision balancée en tant que métrique d'évaluation. Le Tableau 3 regroupe les performances du modèle. À travers une série de tests statistiques, pour le score F1, nous montrons que l'utilisation du module EDIG permet d'aboutir à de meilleures

x

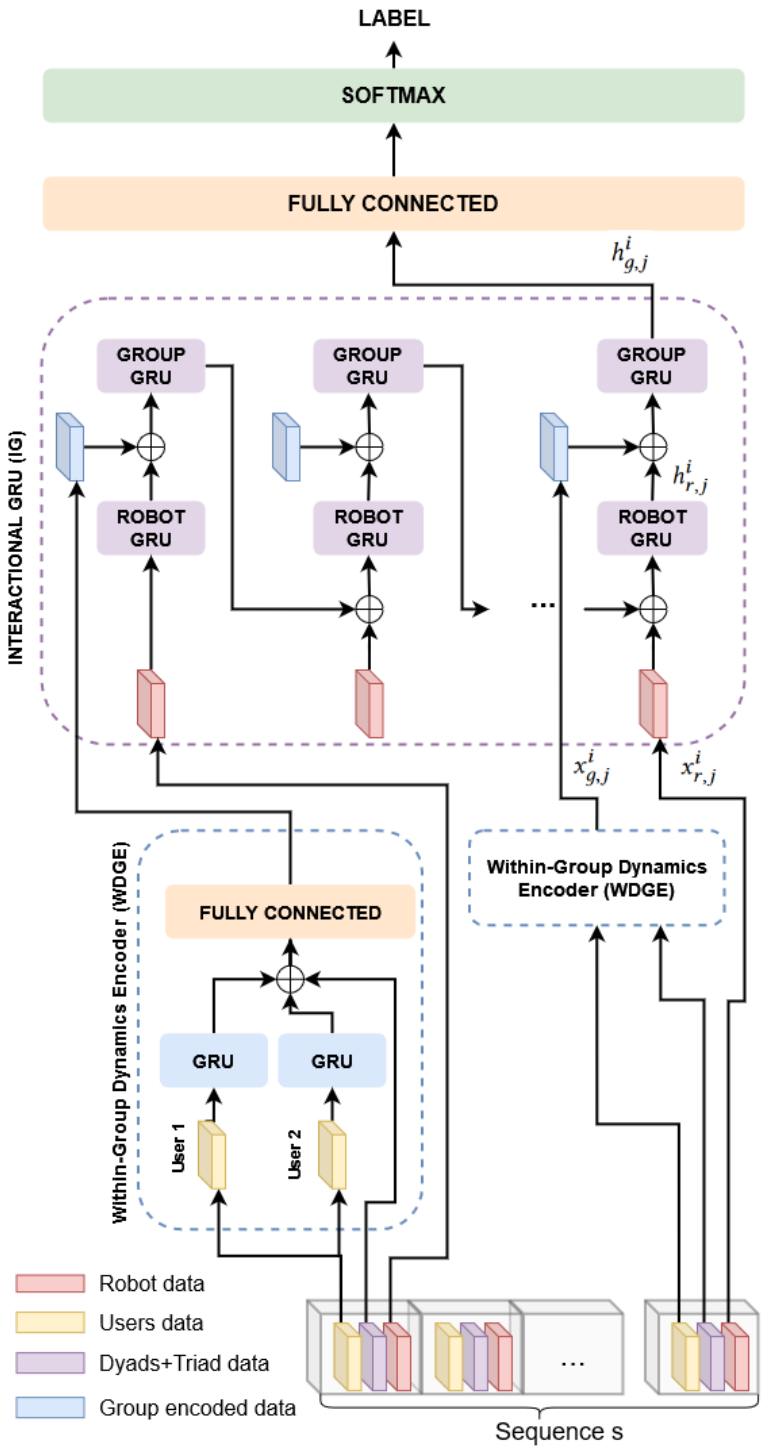

Figure 1: Représentation de notre architecture complète. Les données du robot sont concaténées avec celles du groupe provenant du segment précédent, et ainsi données en entrée du GRU du robot. La sortie est concaténée avec la sortie du module EDIG (ici indiquée en anglais “WGDE”) avant d’être donnée en entrée au GRU du groupe.

performances, indiquant la nécessité de prendre en compte les dynamiques intra-groupe. Il n'y a cependant pas de différence statistique significative entre les modèles GS et GI signifiant que le module GI seul ne capture pas bien les dynamiques temporelles. En ce qui concerne la précision balancée, les tests statistiques montrent que les meilleurs performances sont obtenues par les modèles GS et EDIG-GI. Une série de tests statistiques montre qu'il n'y a pas de différence significative dans les performances obtenues par les modèles en fonction de différentes tailles de séquence τ. Nous comparons également les performances entre l'EDIG-GI, les Forêts Aléatoires, et le PMC en prenant en compte toutes les modalités et même paradigme d'apprentissage. Une série de tests statistiques montre que notre architecture complète EDIG-GI permet d'obtenir de meilleurs performances.

τ	1	2	3	4	5	6	7	8
GS	0.733 ±.120	0.739 ±.119	0.733 ±.118	0.733 ±.127	0.735 ±.123	0.735 ±.125	0.734 ±.125	0.735 ±.124
GS (sans robot)	0.621 ±.058	0.613 ±.081	0.621 ±.080	0.604 ±.095	0.597 ±.084	0.598 ±.092	0.603 ±.085	0.605 ±.087
EDIG-GS	0.726 ±.116	0.732 ±.119	0.723 ±.144	0.730 ±.141	0.724 ±.138	0.725 ±.146	0.723 ±.146	0.731 ±.148
GI†	0.730 ±.113	0.717 ±.105	0.695 ±.120	0.698 ±.163	0.694 ±.182	0.710 ±.145	0.689 ±.175	0.694 ±.188
EDIG-GI	0.730 ±.102	0.730 ±.098	0.715 ±.143	0.736 ±.124	0.735 ±.135	0.745 ±.110	0.730 ±.146	0.714 ±.137

Table 3: Scores micro F1 moyens et écarts standards sur les sets de test des modèles en classification multi-classe pour $\tau \in [1, 8]$.
†: le modèle GI correspond au EDIG-GI sans le module EDIG.

Analyse des modèles

Nous menons une analyse SHAP [62] afin de déterminer l'importance de nos descripteurs, en utilisant le modèle des Forêts Aléatoires en fusion immédiate. L'analyse montre des comportements différents en fonction de la catégorie de confiance du segment. Pendant les segments de méfiance, les participants baissent plus leurs sourcils, parlent plus, ont un nombre de changement de focus d'attention visuel plus élevé, se tiennent plus à distance du robot, parlent plus en même temps que le robot. Ainsi, ils montrent plus de doute en regardant autour d'eux et en se rapprochant entre-eux. Pour les segments de confiance, les participants hochent plus la tête, plissent plus les yeux, sont plus proches du robot, baissent le ton de leur voix. Ainsi, ils bougent plus, parlent plus, s'alignent avec le robot.

En ce qui concerne nos modèles récurrents, nous menons une analyse d'erreur avec le EDIG-GI et $\tau = 6$ pour comprendre quels segments sont les plus durs à classifier et en comprendre les raisons. Nous remarquons que notre modèle a des taux d'erreur différents en fonction de quelle a été l'interaction de test. Les interactions et plus précisément les segments pendant lesquels les participants sont plus sarcastiques sont les plus durs à classer. En regardant les segments de toutes les interactions qui ont un taux d'erreur supérieur à 60%, nous observons que ceux-ci sont

majoritairement annotés avec le label “Alignement” - le process par lequel deux participants créent et maintiennent une action jointe en adaptant leur comportement de manière adéquate [112] -, montrant la difficulté du modèle à bien capturer ce phénomène. En particulier, les segments de méfiance avec un fort taux d'erreur sont principalement annotés “regard”, “expression faciale”, “intonation”. Ces segments sont généralement associés à des moments de doute, montrant que l'ajout de la sémantique des participants pourrait être bénéfique pour mieux modéliser ce phénomène. En ce qui concerne les segments de confiance, nous retrouvons principalement des annotations de “regard”, “expression faciale”, et ‘F-formation”. À nouveau, ces segments font référence à des moments d'alignement qui a été démontré comme phénomène complexe que le modèle ne capture pas correctement. Nous observons également que les annotations de “statut de participation” sont nombreuses dans les segments de méfiance et de confiance à fort taux d'erreur, indiquant la difficulté de compréhension de changement de statut par le modèle et sa capacité à résoudre leur ambiguité.

Conclusion

Dans cette thèse, nous avons abordé les problématiques soulevées par le développement de méthodes d'analyse automatique de la confiance tout au long d'une interaction humain-robot. Nous avons proposé certaines solutions, et résumons nos contributions ainsi :

- Établissement d'une nouvelle méthodologie d'analyse de la confiance en HRI: nous avons introduit une nouvelle méthodologie d'analyse basée sur des théories de sociologie interactionniste, provenant d'approches telles que décrites par l'ethnométhodologie. Au lieu de considérer la confiance comme un état mental, nous considérons la confiance comme un état de l'interaction rendu visible par les participants à travers leurs comportements. L'analyse de la confiance s'opère à travers l'observation de ces comportements et l'analyse de leur pertinence compte tenu de la séquence interactionnelle.

- Création d'un schéma d'annotation pour analyser la confiance en HRI: En se basant sur notre nouvelle méthodologie, nous avons créé un nouveau schéma de codage nommé TURIN suffisamment versatile que pour être employé dans des interactions dyadiques ou de groupe. TURIN permet d'analyser les dynamiques de la confiance, et d'étudier les comportements multimodaux que les participants exhibent lorsqu'ils expriment de la confiance. En comparant ce schéma avec un questionnaire de confiance HRI répandu provenant d'une approche psychologique, nous avons démontré que les analyses peuvent aboutir à des conclusions différentes, en particulier TURIN peut révéler des moments où les participants montrent de la confiance alors que le questionnaire montre le contraire (et vice-versa). Nous avons également démontré que les annotations collectées avec TURIN peuvent être utilisées pour définir la cible de l'apprentissage de modèles de machine learning.

- Proposition d'un ensemble de descripteurs comportementaux relatifs à la confiance: Nous

avons établi et proposé un ensemble de descripteurs comportementaux multimodaux qui peuvent être utilisés pour des modèles computationels de la confiance. Cet ensemble repose entièrement sur des descripteurs qui sont physiquement non-intrusifs pour les participants, avec un mélange de descripteurs extraits automatiquement et manuellement. À travers une étude de deux mécanismes de fusion différents, nous avons démontré qu'une fusion immédiate permet d'obtenir de meilleures performances, indiquant une interaction entre les différentes modalités. À travers une analyse de l'importance des descripteurs, nous avons vu que la modalité vocale jouait un rôle prépondérant, et que certains descripteurs ont plus d'influence envers une catégorie de confiance.

- Conception de modèles multimodaux automatiques de la confiance: Nous avons proposé deux approches pour des modèles de la confiance par des techniques de machine learning. La première se base sur l'hypothèse que le contexte n'est pas requis pour prédire la confiance du groupe, avec des modèles simples de machine learning. Nous avons démontré que la fusion immédiate permet d'aboutir à de meilleures performances. La deuxième se base sur l'hypothèse que l'historique d'interaction permet d'améliorer la qualité de la prédiction. Nous avons introduit deux modules pour modéliser les dynamiques interactionnelles. Le premier module modélise les interactions au sein du groupe de participants à différentes échelles. Le deuxième module modélise la structure temporelle de l'interaction entre le robot et le groupe, de manière similaire à un dialogue. Nous avons montré que notre architecture obtient de meilleurs performances que les modèles simples, mais que la longueur optimale de la taille du contexte reste à déterminer.

Les défis relevés par cette thèse offre quelques perspectives de recherche que nous décrivons sommairement ici :

- Détection en ligne de la confiance: Nos travaux s'inscrivent dans un première démarche d'analyse hors-ligne de la confiance. Afin de mener une détection hors-ligne, quelques ajustements sont nécessaires. D'abord, la segmentation de l'interaction se doit d'être automatisée pour correspondre le plus à celle suggérée par TURIN, ou alors la segmentation peut s'opérer avec un pas de temps fixe dont la longueur est à déterminer. L'ensemble de descripteurs proposé doit également être améliorer afin de ne pas inclure de descripteurs extraits manuellement.

- Collecte de données dans un scénario spécifique à la confiance: Il n'y a aucun jeu de données publiquement disponible dont le scénario a été pensé spécifiquement pour étudier la confiance.

- Amélioration de TURIN: Nous avons essayé d'être le plus exhaustif possible pour la catégorie "Forme de l'interaction sociale". Cependant, la catégorie "Contenu de l'interaction" pourrait bénéficier d'ajouts de phénomènes connexes à la confiance comme l'engagement.

- Amélioration des modèles de dynamiques interactionnelles pour la confiance: Il pourrait être intéressant de modéliser les dynamiques intra-groupe de manière différente, en utilisant

par exemple un réseau de neurones par graphe. Il serait également intéressant de concevoir une architecture hiérarchique qui puisse être capable de modéliser les différentes échelles d'analyse requises pour prédire la confiance (un niveau court-terme, et un niveau plus long-terme). Nos expériences ont également démontré le besoin de conduire plus d'expériences afin de déterminer la taille optimale de l'historique d'interaction.

Contents

I Introduction **1**

1 Introduction **3**
1.1 Social Robotics . . . 3
1.2 What is trust and why trust ? . . . 4
1.3 Research questions . . . 6
1.4 Manuscript organization . . . 9

II State-of-the-Art **11**

2 Models and analysis of trust in HRI: a state-of-the-art review **13**
2.1 Psychological definitions . . . 14
2.1.1 Definitions and models of trust . . . 14
2.1.2 Social phenomenons connex to trust . . . 17
2.2 Trust measures . . . 17
2.3 Automatic trust analysis methods . . . 19
2.4 Available datasets . . . 21

III Automatic Analysis of Trust Dynamics **25**

3 A new framework for trust analysis in HRI **27**
3.1 Interactionist Sociology theories . . . 28
3.1.1 A paradigm shift from a Philosophy of Science point-of-view . . . 28
3.1.2 Implications for trust analysis . . . 29
3.2 TURIN . . . 31
3.2.1 Describing the coding scheme . . . 31
3.2.1.1 General Overview . . . 31
3.2.1.2 Nonverbal and verbal trusting behaviors . . . 32
3.2.2 Social Interaction Form . . . 33
3.2.3 Interaction Content . . . 36
3.2.4 Benevolence . . . 37

3.2.5 Integrity . . . 38
3.2.6 Adapting the coding system . . . 38
3.2.7 Choice of the Vernissage corpus . . . 38
3.2.7.1 Validation of the coding scheme . . . 39
3.3 Comparison with the mentalist approach . . . 42
3.3.0.1 Procedure . . . 42
3.3.0.2 Tool adaptation . . . 42
3.3.0.3 Results . . . 44
3.3.0.4 Differentiating criteria . . . 46
3.4 Conclusion . . . 48

4 Trust analysis throughout the interaction 51
4.1 Multimodal features for automatic trust analysis . . . 52
4.1.1 Face . . . 52
4.1.2 Body . . . 53
4.1.3 Voice . . . 54
4.1.4 Semantics . . . 55
4.2 Multimodal computational models for trust analysis . . . 56
4.2.1 Simple machine learning models . . . 56
4.2.1.1 Formalization . . . 56
4.2.1.2 Machine learning models . . . 57
4.2.2 A recurrent neural architecture . . . 57
4.2.2.1 Formalization . . . 58
4.2.2.2 Gated-Recurrent Units . . . 58
4.2.2.3 A first sequential approach . . . 58
4.2.2.4 Modeling within-group dynamics . . . 59
4.2.2.5 Modeling the temporal structure of the users-robot interaction . 60
4.3 Conclusion . . . 61

5 Experiments and Analysis 63
5.1 Training results . . . 64
5.1.1 Dataset . . . 64
5.1.2 Without context . . . 64
5.1.3 With context . . . 67
Parameters . . . 67
Results . . . 68
5.2 Feature importance . . . 71
5.2.1 Face modality . . . 74
5.2.2 Body modality . . . 74
5.2.3 Voice modality . . . 74
5.2.4 Global analysis . . . 74

5.3 Error analysis 75
5.4 Conclusion 76

IV Conclusion **79**

6 Conclusion **81**
6.1 Contributions 81
6.2 Perspectives 84

Part I

Introduction

Chapter 1

Introduction

1.1 Social Robotics

The term *robot* was first introduced to the public by Czech writer Karel Čapek in his play "Rossum's Universal Robots" published in 1920. The word comes from the slavic word *robota* which means work/job, and was used to designate artificial people that could be mistaken for humans [129]. In this play, robots initially work happily with humans, but eventually revolt and cause the extinction of the human race. The term "robot" is therefore strongly linked by its etymology to questions of trust towards autonomous agents of our own creation, and fears of technology getting out of humanity's control. This is a recurring theme in the litterature, or even movies such as Chris Columbus' "Bicentennial Man", Stephen Spielberg's "Artificial Intelligence", Spike Jonze's "Her", or more recently Kogonada's "After Yang". The core theme of these works revolves around questions of believability of the robots to act as autonomous social agents, and how much they can integrate the human social space as entities that are "conscious" of their actions and desires.

Interestingly, even though the term was originally created to refer to humanoid objects crafted by mankind that feel "alive" and "intelligent", the term has now gained a broader meaning to refer to autonomous objects created for a specific repetitive task such as cooking or vaccuum cleaning. Robotics is now the general domain interested in the creation of physical robots that engage in various types of task. It is an interdisciplinary domain at the crossroads of computer science, design, electronics, engineering, psychology, and sociology. A subdomain of the field called *social robotics* tackles the specific issues of robots that enter the social world [12].

The fields of Social Robotics and *Human-Robot Interaction* (HRI) try to imbue robots with a specific type of human intelligence that is *social intelligence*. This term refers to the ability to perceive, interpret, predict the behaviors of another human [84]. Similarly to how we humans had to and continuously learn how to interpret other people's behaviors to connect with them, robots should too as they are bound to interact with humans. Social robots are expected to play more important role in the service industry, in positions such as frontdesk in supermarkets, museum guides, elder everyday care, or as concierge in hotels. Therefore, not only should they

learn to interpret human behaviors, but they also should learn how to properly interact in a way that is believable and feels natural for humans. The difficulty of both the understanding and production of behaviors resides in the ecological complexity of the situations in which they occur. Each behavior has to be replaced in its cultural context, the interaction context, and can sometimes be modulated by the interaction partner's personality. For instance, a Japanese person will nod more often during a conversation to communicate that they are listening while an American person will do so less often [119]. These are the type of differences that a robot should be able to account for when interacting with a human to be perceived as a believable social agent.

Believability and acceptance have been studied in depth in HRI, from the robot design (e.g. level of anthropomorphism, height, color) to the way they should behave, be it on a technical aspect (e.g. movement speed, speech flow) or on a social aspect (e.g. timing of an apology after a mistake, social navigation) [1, 52, 56, 70, 78, 128]. A major concept that affects the robot acceptance is trust [4, 50, 59]. As current abilities of robots can be misaligned with users' expectations as well as their general representation in media, it is important to make sure that users' trust is properly calibrated given the role that robots will play in the service industry.

1.2 What is trust and why trust ?

Historically, trust is framed in HRI as a psychological construct. One of the most widely used definition is Rousseau's, who defined it as *"a psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another"* [97]. There are three important points that arise from definition. First, trust is defined as a psychological state, meaning that each individual forms its own representation of the interaction partner and decides to trust it based on this representation. This representation is formed by robot-based criteria, environment-based, and user-based ones [42]. Previous work showed that robot-related factors have the strongest impact on the user's trust, such as the robot's perceived gender, its embodiment, but also the type of error it makes during the interaction [77, 114, 115]. Environment-based factors include elements from the interaction location, while user-based ones relate to the user's propensity to trust [71]. This representation is then the basis of the user's decision to trust the robot or not in a certain situation. However, as a mental state, the robot can never have a direct access to the user's trust but can only infer it through the user's behaviors and decisions during the interaction.

Second, the vulnerability aspect in Rousseau's definition is important. A form of collaboration between the interaction partners has to happen for trust to be relevant, there has to be something at stake to determine whether collaboration is possible or not [42]. Vulnerability can either be physical (e.g. relying on the robot plan in a fire escape scenario) or mental (e.g. revealing a personal secret known by a handful of close people). This vulnerability and the risks it presents partly determine the user's choice to trust. The last element to take into account is the user's positive expectations of the intentions or behavior. The user will most

probably not decide to trust the robot if the vulnerability cost is higher than the potentiality of a positive reciprocity from the robot [42, 58, 77, 114, 115]. In a Prisoner Dilemma scenario, two participants are given some money. They can either decide to keep it for themselves, or give some or all to their partner. If they do so, the money they give is doubled for their partner. In this scenario, the optimal outcome for a single individual is to keep their money and hopefully receive all of their partner's money that is going to be doubled. However, the optimal outcome for both is them giving all their money to each other. In [58], the researchers set up an experiment with this scenario. The participants talk for 5 minutes to get to know each other before the money exchange part. This allows the participants to form a representation of their partner's trustworthiness based on their behavior, and assess the financial risk of the game.

In this scenario, we see that trust determines the financial outcome of both participants. Trust, as a psychological and sociological construct, plays a fundamental role in the development and maintenance of relationships between individuals [4, 50]. Previous work in human-automation interaction (HAI) shows that trust has an impact on the robot acceptance by human users, and the interaction task outcome and performance [59]. A poor calibration of users' trust can lead to sub-optimal situations. When users have too little trust in the robot - situations of *undertrust* -, they tend to not rely on the robot at all [115]. On the other hand, if users have too much trust towards the robot - situations of *overtrust* -, they can fail to properly monitor the robot activity during technical activities, for instance when cooking together. Overtrusting can also lead to situations where the robot might ask the user to do things for its own benefit in social activities, by making them disclose personal information for instance [4, 5, 93]. As both understrusting and overtrusting can lead to potentially dangerous situations for the user, trust should be monitored so that it is properly calibrated during the interaction.

Trust monitoring can be done either through questionnaires or through proxy measures depending on the interaction scenario - the amount of gold given in the Prisoner's Dilemma, or other decisions to trust that are relevant for the interaction. Previous studies in HRI measured trust through punctual self-assessments surveys filled by users [102, 103, 105, 106]. However, such assessments are generally filled at the beginning and end of the interaction, as they are time-consuming to fill, which limits the possibilities of trust monitoring during the interaction. Furthermore, as the questionnaires available in HRI were constructed from mentalist approaches from Psychology theories, they only measure the user's representation of the robot based on trust-related criteria. For instance, Schaefer developped the "Trust Perception Scale-HRI" that includes 40 items measured on a 11 Likert scale [105]. The items include questions about the user's representation of the robot ability to communicate, to share information but not disclose personal information, to be dependable, and to adapt to its environment. Such questionnaire allowed researchers to investigate the socio-psychological effects of the robot on users, by analyzing both verbal and nonverbal communication between the robot and users. Punctual trust assessment tools that are currently available in HRI do not allow to properly monitor trust during the interaction as they are time consuming. This important limitation led past research to use proxy measures, as previously explained, or "objective" measures (e.g.

distance to the robot, physiological measures) to monitor trust during the interaction.

Continuous assessment tools used by external observers in Psychology exist. Such trust assessment tools in HRI, however, come from psychological theories adopting a mentalist approach for which trust is considered a mental state of users. Because of this, external observers, when providing their assessment, have to infer the users' mental states since they do not have direct access to it. This inference can introduce subjective biases from the observers [108]. Interactionist Sociology methodologies worked around this issue by relying on what is made observable by the users themselves [34, 44]. The work of the observer committed to research is thus that of an eavesdropper who describes what is already shown by users engaged in their interaction.

To conduct an offline analysis of trust regularly throughout the interaction, another possibility is to rely on a less "internal", and "mentalist" approach as was mostly previously done in HRI [9, 27]. Interactionist Sociology methodologies shed light on the observable characteristic of trust within the interaction through behaviors made visible by users. In this light, trust is thus oriented towards several interactional processes: e.g. trust in the proceedings of the interaction, in the robot knowledge, in its capacity to perform a certain action at a given moment. Through perspectives such as those offered by Interactionist Sociology and Ethnomethodology [31, 35, 36, 38, 44], trust can be defined as a *"form of affiliation and credit characterized by a set of behaviors that are intentional or not, expressive or propositional"* [37, 38, 44, 45, 110]. This definition relies on concepts such as *affiliation* - claiming access to and understanding the partner's stance, and endorsing their perspective - [110], *credit* [26, 88] given to the robot *competence* [33], and *alignment* [110] - *i.e.* complying with the *trajectory* (sequential progression) of the interaction. Trust is thus a process as a result of a state of the interaction where users are displaying, in the here and now of interaction, a form of preference for fluid and progressive interactions. This definition also implies that the robot is treated as an interaction partner in a similar manner to what occurs in human interactions. Thus, when interacting with a robot, the user expects it to have a set of basic behavioral skills that are necessary to make the ordinary course of the interaction progress fluidly [111].

Interactionist Sociology methodologies thus allow an external observer to assess users trust based on a normative analysis of interactional processes expressed through their behavior. Such methodology does not rely on the assumption of the user's psychological state to analyze trust. Indeed, in this approach, the mental state is addressed as a socio-cultural phenomenon, that is, accessible and shareable among the parties involved in an interaction. Hence, in this approach one would not try to assume some mental state, but would scrutinize how trust as an interaction resource is made visible.

1.3 Research questions

We introduced a few of the problematics around trust analysis in HRI. In this thesis, we address issues around the automatic multimodal analysis of trust dynamics in HRI through

offline methods. In this context, it addresses the following research questions:

1. **RQ 1**: Which theoretical framework is applicable to perform a multimodal analysis of trust regularly throughout the interaction ?
2. **RQ 2**: Do homogeneous segments of trust arise within the interaction based on observable behavioral cues ?
3. **RQ 3**: How can we discriminate trusting segments from mistrusting ones with tangible behavioral cues ?

The contributions of this thesis are as follows:

New methodology for trust analysis in HRI

We introduce a new methodology to the problem of trust analysis in HRI, based on Interactionist Sociology, such as in Conversational Analysis, stemming from approaches prescribed by Ethnomethodology. Rather than considering trust as a mental state, trust is considered a state of the interaction made visible by participants through their behaviors. Trust analysis therefore relies on the observation of these behaviors, and the analysis of the behavior relevance within the interactional sequence. This method allows the researcher to conduct the study of trust dynamics from an external point of view from the interaction, without interrupting the interaction.

This work brought answers to *RQ-1* and led to the publication of two conference papers:

- *M. Hulcelle, G. Varni, N. Rollet and C. Clavel, "TURIN: A coding system for Trust in hUman Robot INteraction," 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 2021, pp. 1-8,* `https://doi.org/10.1109/ACII52823.2021.9597448`.
- *M. Hulcelle, G. Varni, N. Rollet and C. Clavel, "Comparing a Mentalist and an Interactionist Approach for Trust Analysis in Human-Robot Interaction," International Conference on Human-Agent Interaction (HAI '23), 2023, Gothenburg, Sweden,* `https://doi.org/10.1145/3623809.3623840`

Coding scheme to analyze trust in HRI

Grounding on the theoretical framework that we proposed, we created a coding scheme for trust in HRI called TURIN that is versatile enough to be used for dyadic or group interactions. TURIN allows to analyze trust dynamics, and study the multimodal behaviors that users express when displaying trust. To the best of our knowledge, this coding scheme is the first that was specifically designed for trust in HRI [1].

This work brought answers to *RQ-1* and *RQ-2* and led to the publication of one conference paper: *M. Hulcelle, G. Varni, N. Rollet and C. Clavel, "TURIN: A coding system for Trust*

[1] We also released annotations that we collected on the following link: `https://doi.org/10.5281/zenodo.8409887`

in hUman Robot INteraction," 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 2021, pp. 1-8, https://doi.org/10.1109/ACII52823.2021.9597448.

Proposition of a trust-relevant set of features

We propose a set of multimodal features that can be used to build computational models of trust. The set solely relies on features that are physically non-intrusive for the participants, with a mix of manual and automatically extracted ones. We introduce features and modalities that have not been used to analyze trust in HRI.

This work brought answers to *RQ-3* and led to the publication of one workshop paper: *Hulcelle, M., Varni, G., Rollet, N., Clavel, C. (2023). "Computational Multimodal Models of Users' Interactional Trust in Multiparty Human-Robot Interaction". In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13643. Springer, Cham.* https://doi.org/10.1007/978-3-031-37660-3_16

Conception of multimodal models of trust

We propose two types of multimodal models of trust: one that is based on traditional machine learning (ML) techniques, and one based on a neuronal architecture. We explore several traditional ML models, and investigate the impact of early and late-fusion of features on the models performance. We then propose a new neuronal architecture [2] composed of two main modules to better model the interactional dynamics relating to trust. Each module is based on hypotheses derived from Interactionist Sociology theories. The first one encodes within-group dynamics to model the interactions between users with different granularities. The second one models the interaction between the robot and the group as a dialogue with a temporal structure.

This work brought answers to *RQ-3* and led to the publication of one conference full-paper and one conference poster-paper:

- *Hulcelle, M., Varni, G., Rollet, N., Clavel, C. (2023). "Computational Multimodal Models of Users' Interactional Trust in Multiparty Human-Robot Interaction". In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13643. Springer, Cham.* https://doi.org/10.1007/978-3-031-37660-3_16

- *M. Hulcelle, L. Hemamou, G. Varni, N. Rollet and C. Clavel, "Leveraging Interactional Sociology for Trust Analysis in Multiparty Human-Robot Interaction," International Conference on Human-Agent Interaction (HAI '23), 2023, Gothenburg, Sweden,* https://doi.org/10.1145/3623809.3623973

[2] Code for the architecture is available here: https://github.com/GrituX/WGDE_IG

1.4 Manuscript organization

The manuscript is organized in two main parts. In the first part, we will present previous research on trust that mainly relied on Psychology theories (Chapter II). We discuss the theoretical framework for such mentalist approach of trust, and the existing trust models that are the foundation of a majority of trust studies. We present the assessment tools that were built through this mentalist approach. We then discuss the existing automatic trust analysis methods, from the choice and design of features to the computational models.

In the second part, we first discuss in Chapter III the introduction of a new framework for trust analysis in HRI based on Interactionist Sociology theories which allows us to perform a multimodal analysis regularly throughout the interaction. We then present our coding scheme TURIN to tackle this issue, and discuss its complimentarity with existing mentalist approaches. Following this, in Chapter IV, we lay out our computational methodology for an offline trust analysis conducted regularly throughout the interaction, based on machine learning models and a selection of trust-relevant features. Last, in Chapter V, we present the results of our experiments, discuss the importance of features for trust analysis and analyze the errors made by our models.

The manuscript ends on Chapter VI with a summary of our contributions and offers perspectives on multimodal trust analysis in HRI following our work.

Part II

State-of-the-Art

Chapter 2

Models and analysis of trust in HRI: a state-of-the-art review

Abstract

Trust was studied from a Psychological angle in HRI. It was described as a mental state in which the user has a certain set of expectations towards the robot social and technical skills. Most definitions rely on the idea that users expect the robot to be able to handle uncertain situations and act in a way that is not mentally nor physically harmful. Building from these definitions, past research investigated trust antecedents. It showed that user's trust is impacted by the robot-related factors (e.g. its design, how it moves), environmental factors, and user-related factors. To be able to analyze trust during an interaction, trust measures were built. The most commonly used ones are the Godspeed questionnaire, the Interpersonal Trust Scale, the Negative Attitude towards the Robot Scale (NARS), and the Robot Trust Scale. These questionnaires are usually filled by users themselves at the beginning and end of the interaction. Previous research also investigated automatic trust analysis methods to study how trust builds and develops through the interaction, and determine which behaviors are linked to trust.

Trust in HRI was mostly studied through the lens of Psychological theories. In this chapter, we present the different definitions and models of trust. We then discuss the existing trust measures, that are mostly questionnaires. Following this, we explore the computational trust analysis methods from the literature. We finish by presenting some publicly available datasets in HRI and discuss how they can fit our studies objectives.

2.1 Psychological definitions

2.1.1 Definitions and models of trust

Grounding on Psychological theories, trust in HRI is described as a mental state, in which users have a certain set of expectations - either positive or negative - towards the robot technical and social skills. Despite the effort and the numerous studies on trust, there is no unique definition of trust as the way it is defined is heavily influenced by the context in which trust is being discussed. Different robotic agents, applications, human operators might require different trust definitions to properly frame the needs of the study. A fire-emergency scenario with a non-humanoid rover as guide might not require the same definition as in a museum guide robot setting.

Wagner et al. provided a comprehensive definition of trust: *"a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk"* [118]. One of the most commonly used definition in HRI characterized trust as *"psychological state comprising the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another"* [97]. Most definitions, such as the previous ones, rely on the idea that users form a mental model of the robot capability to handle uncertain situations, and act benevolently [51, 64, 104]. Users thus expect the robot to not harm their physical/mental well-being, respect each other's interests, and deal appropriately with uncertain situations. Most definitions also state that the user has something at stake when making the decision to trust the robot or not.

Building on this background, Schaefer distinguished between static and dynamic factors of trust [104]. He grouped those factors under two static components: the user's "propensity to trust" and the robot inherent "trustworthiness". Two dynamic components of trust were distinguished: *cognitive trust* (CT) - the *"self-efficacy to rely on capabilities and reliabilities of a specific party"* - and *affective trust* (AT) - the *"self-efficacy on the party based on human affective responses to the behavior of the party"* [104]. Table 2.1 provides a complete list of factors of his model of trust. While affective trust develops and mosty varies at the beginning of the interaction, cognitive trust continues to vary and settles later on.

Trust is, therefore, a combination of the user's mental projection of the robot capabilities and an affective response to these. Psychological definitions also show that trust is affected by the observation of present events and the user's projection of future events. This link between trust and the uncertainty generated by the projection of future events is highlighted in [14] that defines trust as *"a process of uncertainty reduction, the ultimate goal of which is to reinforce assumptions about a partner's dependability with actual evidence from the partner's behavior"*. Interestingly, while the last definition mentions the user's mental model of their partner's dependability, it also mentions the reliance on perceptible behaviors and behavioral proofs of this dependability, which leads us to the interactionist approach that we later describe in Chapter III.

One of the first objectives of research on trust in HRI was to determine the factors that

Static		**Dynamic**	
User's propensity to trust	Robot trustworthiness	Cognitive trust	Affective trust
User	Anthropomorphism	Competence	Social behavior
Personality	Level of Automation	Perceived intelligence	Proxemics
Interaction history	Physical Design	Efficiency	
Demographics	Assigned Gender	Reliability	Animacy
	Reputation	Responsability	Warmth
	Context role	Interactivity	
	Perceived function	Knowledge	

Table 2.1: Schaefer's model of trust including two static components ("Propensity to trust" and "Trustworthiness") and two dynamic ones ("Cognitive trust" and "Affective trust") [105].

influences the user's trust. Yagoda et al. [124] made a comprehensive model of factors that impact trust, originally created to develop a HRI Trust Scale, grouped under five major categories: team configuration, team process, context, task, system. Table 2.2 provides the detail of the dimensions of each category of the trust model.

Both models of Schaefer and Yagoda show the importance of a user's personality, cultural background, and interactional history with robot on trust. These antecedents shape the user's context that will impact how much trust they will be able to give during an interaction with a robot. On top of this, the interactional context will also impact the user's capability to trust, as explicited by Yagoda et al's model: the social environment (e.g. number of interaction partners, team configuration), the physical environment, the task and how it is configured (social or technical task ? Objective of the task ? How difficult is it ?). Their models reveal that many variables have to be taken account when studying trust, hence showing the difficulty to generalize the results of a study.

The HRI community mainly addressed dyadic interactions when investigating trust. As robots mainly face groups during in-the-wild interactions, it is necessary to further study team trust. However, few studies on team trust in HRI exist. When users gather to perform a joint task or achieve common objectives with a robot, trust is exhibited differently than during dyadic HRI. Indeed, users have to communicate their intentions and actions through verbal or nonverbal behavior to ensure a smooth interaction [36, 38]. Team emergent states appear due to these interactions, and develop over time [67]. Team emergent states are cognitive, affective, and motivational states of team that are *"dynamic in nature and vary as function of team context, inputs, processes, and outcomes"* [68]. Kozlowski and Klein classified a phenomenon as emergent when *"it originates in the cognitive, affect, behaviors, or other characteristics of individuals, is amplified by their interactions and manifests as a higher level, collective phenomenon"* [54]. Team trust is one of these states, described in particular as a cognitive-affective team emergent state [90]. Analyzing trust in a group setting, therefore, leads to the analysis of the dynamics between all of the users involved, and understanding how within group dynamics impact the emergent state of the group.

Trust Category	Dimensions
Team Configuration	Operator Human team member Supervisor Subject matter expert (SME)
Team Process	Communication Coordination Team dynamics Situational awareness Decision making Planning/Replanning Backup Leadership
Context	Operation Task Physical environment Social environment Previous task knowledge Previous human team member experience Previous physical environment experience Previous overall system knowledge
Task	Required skills Task allocation Objectives Task difficulty Task feedback Feedback from human team members Feedback from the physical environment Feedback from the overall system
System	User interface Sensor data Navigation capabilities Signal/Bandwidth End effectors Remote information processing Level of automation Type of control

Table 2.2: Yagoda et al's model of trust consisting of five categories and their dimensions [124].

2.1.2 Social phenomenons connex to trust

As shown by the models that we presented previously, trust is a complex multi-faceted psychological construct. It involves many different connex concepts that we define and discuss their links here. One major component of Yagoda's model of trust relates to the interaction task. Yagoda showed that the task difficulty, and its objectives will condition users' trust towards the robot [59, 124]. A user will not place the same amount of trust in a vacuum cleaning robot than a cooking robot when it comes to making bread obviously. But when it comes to interactions that also take place in the social world, trust dependency on the task raises the question of the concept link with phenomenons such as engagement and cohesion. Sidner defined engagement as *"the process by which two (or more) participants establish, maintain, and end their perceived connection to one another"* [109]. Previous studies showed that trust impacts users' engagement, for instance during long-term interactions for rehabitilition [57], in emergency situations with guidance robots [92]. While it is clear that poor trust can lead to the disuse of a robot, hence disengagement, it is still unclear how much one affects the other.

Before considering the end of an interaction where users disengage with the robot, the way participants maintain their perceived connection during the interaction leads to the study of cohesion. Cohesion is defined by Lewin as *"a group characteristic that depends on its size, organization and intimacy"* [61]. Griffith differentiated two dimensions: the Task, and the Social dimensions [40]. Previous studies showed that interpersonal trust between humans mediates the relationship between team cohesion and team performance [63]. As obvious as the link between trust and team cohesion may appear, there are still very few studies that investigate this link in HRI. Adapting the robot to users' knowledge can lead to a proper calibration of trust which in turn can have a positive impact on social cohesion [15].

2.2 Trust measures

Available trust assessment tools in HRI built on a psychological background require participants to answer questions about their mental representation of the robot they will interact or have interacted with. The most used questionnaires are the Interpersonal Trust Scale (ITS) [96], Godspeed [11], the Negative Attitudes towards Robots Scale (NARS) [113], and the Robot Trust Scale (RTS) [104]. In the following, we present the different questionnaires.

The Interpersonal Trust Scale [96] was designed to measure a person's generalized expectancy that the promises of other individuals or of groups with regard to future behavior can be relied on. The scale is composed of 25 items (12 trust and 12 distrust items) along with 15 filler items. All items are evaluated on a five-Likert scale. The ITS was built and validated only for trust towards another human, and has never been used nor validated for trust towards a robot, a virtual agent, a machine, or a computer. This questionnaire was used as a starting point to establish HRI-specific trust scales or to measure specific antecedants of trust, but it cannot be used on its own to measure users' trust as it lacks concepts and items that are necessary when studying trust towards a robot, as we will see with other scales.

The Godspeed questionnaire was created to measure the user's perception of the robot according to the following criteria [11]:

1. Anthropomorphism: it refers to the attribution of human-like features, characteristics, and behaviors to non-human things such as robots, animals, or other objects. Because of its level of anthropomorphism, a robot will face a certain set of expectations from the user. For instance, users will expect a robot to be able to talk and hear if it has a human-like head. It is therefore important that the robot matches the user's expectations to avoid disappointment potentially created by an initial level of trust that does not match the robot real capacities [75].

2. Animacy: Piaget's framework refers to animacy as something moving "on its own accord", independant of an external pull, that exhibits intentional behavior. It can be rephrased as the user's perception of how "alive" the robot is. As the robot animacy increase, users will generally attribute more social capabilities to the robot, and thus place more trust in its social skills.

3. Likeability: it refers to the evaluation of first impressions after seeing or meeting the robot, and the degree at which these impressions are positive. As a robot likeability increases, so does its general evaluation by users, which in turn impacts how trusting users will be [11, 91].

4. Intelligence: the user may evaluate the intellectual capabilities of the robot based on its task competence, and how knowledgeable and sensible it is. These are important factors to take into account when studying trust, as users may stop using the robot if they consider it to be not intelligent enough. Researchers often resort to deploying the robot through a Wizard-of-Oz setup to experiment to improve its perceived level of intelligence by controlling its competence [11, 116].

5. Safety: it plays an important role for trust as hazardous robots will be disused. Under-informed users about the potential hazards of the robot might lead to misuse and thus to dangerous situations [11, 93].

Each criteria in turn consists of five items, except for the "Perceived Safety" which includes only three items. Each item is evaluated on a 5-Likert scale. While the Godspeed questionnaire does not directly measure trust, most of its items have an impact on trust. Researchers mostly used this questionnaire as a way to find trust antecedents and study its correlates.

The Negative Attitudes towards Robots Scale (NARS) [113] was developped to understand how the behavior and embodiment factors of a robot impact the users' representation and response to the robot. This scale comprises 14 items split across three different sub-scales:

1. **Sub-scale 1:** negative attitudes toward situations and interactions with robots

2. **Sub-scale 2:** negative attitudes toward social influence of robots

3. **Sub-scale 3:** negative attitudes toward emotions in interaction with robots

The NARS was first developped in Japanese, and was then translated in English. The English version contains three less items than the Japanese for internal consistency reasons due to cultural differences. It was not designed to directly measure trust, but researchers have used this scale to, again, establish antecedents and correlates to trust, or used as subsequent evaluations of a given interaction [79]. For instance, Aroyo et al. used the NARS as a trust measurement tool before and after a social engineering task where the robot tries to make participants gamble [5]. They used it to determine the impact on the robot likeability after it tries to build trust and rapport by asking the user for personal information, providing clues in a treasure hunt game, and suggesting to gamble their prize.

While NARS can be useful to assess a participant's negative preconception of the robot to interact with, NARS and ITS questionnaires are highly correlated with pre-interaction trust measures, but not post-interaction trust measures [104]. The RTS can be filled by participants before and after an interaction with a robot to assess their trust. This scale focuses on antecedents and measurable factors of trust related to the human, robot, and environment. The scale comprises 40 items, but a smaller subset of 14 items can be used for a faster assessment [104]. Each item represents the participant's expectation of the robot behavior given their mental model of the robot - e.g. *"What % of the time will this robot act consistently"*. Answers are given in the form of an 11-point Likert scale, from 0 to 100. The final trust score is the average of all individual items score. The scale encompasses items that relate either to the robot perceived technical or social skills. Some items can be interpreted in both ways. For instance, "acting consistently" can either relate to the predictability of the output of the task the robot is working on - e.g. baking cookies - or to the consistency of its displayed personality - e.g. being friendly, then becoming overly sarcastic would be inconsistent.

2.3 Automatic trust analysis methods

Previous work in HRI mostly studied how the robot behavior impacts the user's trust. There are very few studies on automated analysis of users' trust dynamics, even less that focus on how users exhibit behaviors indicating trust [53]. Out of the few multimodal computational models of trust in HRI that exist, Lee et al. [58] studied trust in a Prisoner's Dilemma scenario. They investigated which specific behaviors impacted the partner's decision to trust it at the end of the interaction, both in Human-Human Interaction (HHI) and in HRI. Their trust measure is a proxy one that corresponds to the number of coins a user is willing to give to their partner. They train two Hidden-Markov-Models (HMM) with the sequence of the previously identified behaviors - related to posture, smile, and eye -, one for interactions resulting in a high trust level, and another one for a final low trust level. This separate training revealed different patterns of behaviors for both conditions. Some behaviors were linked to a low-trust outcome - e.g. arms crossed, looking away, face touching -, while others were linked to a high-trust outcome - e.g. smiling, arms in lap. This study focused on specific multimodal behaviors, and

relies on a single final evaluation of trust, thus not taking into account changes of trust during the interaction.

Khalid et al. [49] analyzed psycho-physiological correlates to trust in HRI by building a neuro-fuzzy ensemble trained to classify against trust categories defined by Mayer: Ability, Benevolence and Integrity [71]. While their work provides a set of features that are relevent when modeling trust, they did not focus on the evolution of trust throughout the interaction. Both of these work showed the importance of multimodality to study trust.

Among the rare studies that deal with trust inference throughout the interaction, Xu and Dudek built an online probabilistic trust inference model based on a Dynamic Bayesian Network [123]. They trained a separate model instance on each user's experiences on an aerial robot navigation task, in a supervisor-worker style human-robot team, using the robot performance and the user's intervention as features. In their study, they rely on a trust assessment directly filled by users to train their model. They focused on a technical task with relational asymmetries between the robot and the user, while we focus on a social task during which the robot is considered an autonomous social entity by users.

Description	Representation	Value Range
Situations	$\alpha, \beta, \ldots$	
Agents	$a, b, c, \ldots$	
Set of agents	$\mathcal{A}$	
Societies of agents	$\mathcal{S}_1, \mathcal{S}_2 \ldots$	
	$\mathcal{S}_n \in \mathcal{A}$	
Knowledge (e.g., x knows y)	$K_x(y)$	True/False
Importance (e.g., of α to x)	$I_x(\alpha)$	$[0, +1]$
Utility (e.g., of α to x)	$U_x(\alpha)$	$[-1, +1]$
Basic Trust (e.g., of x)	T_x	$[-1, +1)$
General Trust (e.g., of x in y)	$T_x(y)$	$[-1, +1)$
Situational Trust (e.g., of x in y for α)	$T_x(y, \alpha)$	$[-1, +1)$

Figure 2.1: Summary of the basic notation of Marsh's formalization of trust.

Marsh proposed another approach to computationally model trust in HRI [69]. The summary of his formal notation can be found in Figure 2.1 He conceived a mathematical formalization of general trust from agent x in y as $T_x(y) \in [-1, +1)$. He argues that agent x can only have an estimate of its trust in y from all previously encountered situations α with it which he notes $\widehat{T_x(y)}$. This allows to determine the situational trust of x in y in a given situation α as follows

:

$$T_x(y, \alpha) = U_x(\alpha) * I_x(\alpha) * \widehat{T_x(y)} \tag{2.1}$$

The formula for situational trust highlights some problems which formalizing trust may hold. It is clear that negativity poses problems here, since the multiplication of two negatives results in a positive value. The problem of computing the estimate $\widehat{T_x(y)}$ also arises, as it requires a memory of all past situations α of x and y. Given the value of $T_x(y, \alpha)$, agent x then decides to either trust or distrust agent y in situation α according to different threshold values that depend on agent x propensity to trust. Marsh also provides a method to expand his formalization to temporal variations of trust. While this approach encounters issues inherent to formalization, it provides a simple way of computing agent x choice to trust agent y with respect to their previous encounters. However, it requires the knowledge of many different variables from all possible combination of agents and situations which necessitates some information collection if one agent is a human. In that sense, this approach is purely psychological and therefore never relies on any verbal nor non-verbal behavioral indicators of trust.

Apart from trust, sequential computational models of social phenomenons in HRI are rare. Atamna et al. [6] took inspiration from DialogueRNN [65] to build an RNN-based model for engagement decrease detection for dyadic interactions. They rely on a multimodal analysis - among which posture, speech, gaze and facial expressions - of the interaction, and explicitly model the different parties involved by considering the robot speaking turn as contextual information that helps assessing user's engagement. We took another step from this study by taking into account semantic information from the robot and explored more complex reccurent neuronal architectures to better model the interactional dynamics.

Alahi et al. developed an architecture called "Social-LSTM" to predict human trajectories in future instants [2]. Their idea was to introduce a "Social" pooling layer which allows Long-Short Term Memory networks (LSTM) to share their hidden-states with other LSTM sequences that are spatially proximal. Their architecture can automatically learn interactions that take place among trajectories which coincide in time. We take inspiration from such architecture design that takes into account interactions between participants for the design of our neuronal architecture.

2.4 Available datasets

There are a number of human-robot interaction databases available in the literature. However, to the best of our knowledge, there are currently no publicly available datasets in HRI that were designed to specifically study trust, and that contains answers from one of the previously presented trust questionnaires. There are few datasets that contain participants' answers to the Godspeed questionnaire, but all of them use these assessments to measure participants' perception of the animacy of the robot and its attributed "personality" according to the phenomenon being studied. For instance, the ConcreG8 dataset was designed to study how humans

reorganize their spatial arrangement to accomodate a newcomer [126]. While the results are availble in their dataset, the authors did not discuss the results of the Godspeed questionnaire in the presentation paper.

Given this, we set three criteria for the choice of the dataset. First, the dataset scenario has to consist of a "generic" social activity such as a dialogue. An activity that is generic enough can be at the basis of many different interactions, which would allow our models and, to some extent, some of our findings to be reusable in a scenario with a similar activity. Second, the scenario should involve a group of users. In group settings, users tend to be more expressive in front of the robot as they have to communicate more with other users, while users can choose to be more passive when engaging in dyadic interactions. Last, the scenario has to involve an interaction where the robot speaks that is long enough so that we can observe variations of trust. We will now present some of the existing HRI datasets and discuss the extent to which these could be used to conduct a multimodal analysis of trust throughout the interaction. We report how each dataset fits the criteria in Table 2.3.

The JOKER database [24] includes interactions between a single user and a Nao robot. The aim of the robot is to make participants laugh at its jokes. Three different data collection systems were used, which differed in the capabilities of the robot. These are: i) paralinguistic and automatic, ii) linguistic and semi-automatic, iii) Wizard-of-Oz. The database features recordings collected through webcams, microphones, and Kinect depth sensors. Annotations include users's evaluation of the robot humourous skills and their answers to personality questionnaires, as well as laughter, head gesture, and emotional states. There are a few limitations to use the JOKER dataset to study trust. First, there is no trust-related questionnaires filled by users in the dataset, which is normal since it was not its original purpose. Second, the scenario is not relevant for trust. While we can argue that the credibility of the robot as a comedian is at stake here, the scenario is too asymmetrical for trust to be relevant. Users' participation is limited to laughing, or a few comments after the joke at best. This does not give much opportunity for users to question the robot animacy, anthropomorphism, likeability and hence its believeability as a comedian.

The MHHRI dataset [19] was collected to study attention and engagement in human-human and human-robot interactions. Participants first speak together about themselves. Then, they interact together with a Nao robot and answer one after the other a set of predefined questions. It contains multimodal data of participants such as video, audio, depth, and physiological data. Annotations include self-reported engagement through questions on a 10-Likert scale, and self-assessed as well as interaction partner-assessed personality. Among all the questions, there is only one that relates to trust. This is one of the rare datasets that could be used to study trust as we want to do as the questions asked by the robot require participants to open up to the robot in front of the other participant. However, since the robot asks questions to one participant then the other, group dynamics are fairly inexistant. The participant who is momentarily not adressed to becomes a passive bystander, and has no incentive to listen to their partner's answer. The human-human-robot interaction phase of the dataset thus happen

to be much more like successive dyadic interactions than triadic ones.

The UE-HRI dataset [13] consists of in-the-wild spontaneous human-robot interactions to study users' engagement breakdown. A Pepper robot was installed in the hallway of a school, and participants were free to start the interaction and leave whenever they wanted to. Interactions can either involve a single or multiple users. The interaction is separated into predefined phases: consent form agreement, open questions, explanations about the robot human detection technology, and a final interaction questionnaire. The dataset contains video, audio, depth, and sonar recordings as well as annotations to characterize different engagement cues: sign of engagement decrease, early sign of future engagement breakdown, engagement breakdown, and temporary disengagement. This dataset is not really suited to study trust mainly because of the scenario. The main "human detection technology explanation" is the longest and main phase of the interaction. This phase was designed to try and disengage the user as best as possible, as the robot speaks for a long time with no interruption, and rarely asking for feedback in whichever way possible. Because of this, users' interventions are minimal. This, combined with the fact that nothing aside keeping the interaction alive is at stake, makes trust not very relevant. One interesting thing about this dataset though is the presence of "mini turing tests". Some users spontaneously ask challenging questions to test the robot cultural, mathematical, or physics knowledge mostly. These "turing test" moments can be valuable for trust analysis, as users challenge the robot agency, believeability, autonomy, or knowledge. However, these moments are too short and few in number to be used for analysis.

The Vernissage dataset [47] sets up an interaction between a robot and 2 users. The NAO robot explains paintings in the room and then quizzes the participants on art. Once the paintings presentation is over and right before the quizz, the participants are asked to present themselves by giving more than just their names. This corpus contains multimodal data through recordings of the interactions including several video views, separate audio files for each user, the users' head motion captured by a motion capture tool, and logs of the robot movements. We chose this dataset as it corresponds to all three criteria that we have set. The activity, looking at pictures while the robot provides explanations for them, involves a dialogue that is generic and common enough for users to behave in a casual way even though the experiment happens in a lab.

	Generic activity	Group interaction	Long interaction
JOKER	X		
MHHRI	X	X	X
UE-HRI	X		X
Vernissage	X	X	X

Table 2.3: Summary of the validity for our goal study of the datasets according to three criteria.

We choose the Vernissage dataset as it fills all the criteria. While the MHHRI also fills them, the robot only asks questions to one participant at a time during the group interaction in contrast to the Vernissage dataset in which the robot addresses the user group as a whole most of the time. The choice of the dataset has an impact on the answers to the research

questions, as each dataset creates a unique situation that frames users' behavior in a special way. It is thus important to keep in mind that the base of our methodology that we present in Chapter III is generic, but our answers to research questions will be specific to the dataset that we chose.

Part III

Automatic Analysis of Trust Dynamics

Chapter 3

A new framework for trust analysis in HRI

Abstract

Interactionist Sociology theories stem from a different Philosophy of science point-of-view. While Psychology relies on deductive methods as prescribed by falsificationism, Interactionist Sociology relies on inductive methods, where the observation of recurring patterns in data leads to new discoveries. With methods prescribed from Interactionist Sociology, Ethnomethodology, and Conversation Analysis, we build a new methodology that allows to directly observe users' trust as made visible by the users themselves through their behaviors. We then build a coding system to analyze Trust in hUman Robot INteraction (TURIN). TURIN unitizes the interaction into segments of coherent trust category ("Mistrusting", "Neutral", and "Trusting"), and allows the annotator to describe the users' behaviors based on the form and content of their social interaction, as well as descriptors of their benevolence and integrity. Then, we theoretically compare this new methodology with previous main Psychological approaches. Through an experimental study, we identify criteria that differentiate our approach from the usual Psychological one: orientation, generalization capability, time-framing, and scalability. We provide guidelines on how both approaches can be complementary depending on the target computational model of trust.

Associated publications:

- *M. Hulcelle, G. Varni, N. Rollet and C. Clavel, "TURIN: A coding system for Trust in hUman Robot INteraction," 2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII), Nara, Japan, 2021, pp. 1-8,* `https://doi.org/10.1109/ACII52823.2021.9597448`.
- *M. Hulcelle, G. Varni, N. Rollet and C. Clavel, "Comparing a Mentalist and an Interactionist Approach for Trust Analysis in Human-Robot Interaction," International Conference on Human-Agent Interaction (HAI '23), 2023, Gothenburg, Sweden,* `https://doi.org/10.1145/3623809.3623840`

Previous models of trust were mostly created through the lens of Psychological theories that adopt a mentalist perspective. As current psychological trust measures do not allow to conduct a regular analysis throughout the interaction, we propose a new methodology based on Interactionist Sociology that allow such studies.

3.1 Interactionist Sociology theories

3.1.1 A paradigm shift from a Philosophy of Science point-of-view

As we previously saw in Chapter II, Psychological methods generally used in HRI are highly influenced by the doctrine of falsificationism as described by Popper [87]. Falsificationism implies that any theory, depending on its context of formulation, is considered valid until empirical evidence shows that it does not hold given an empirical context. The validity of a hypothesis is generally demonstrated by statistical significance from empirical data.

We base our methodology of analysis on Interactionist Sociology, such as in Conversational Analysis (CA), stemming from approaches prescribed by Ethnomethodology. CA considers the talk as object of sociological study. It studies how talk emerges naturally during interactions, and tries to describe the details of social organization that make social interactions possible in an orderly and intelligible manner [39, 60, 89]. And so our methodology much more relies on inductive methods where at each step of the analysis, the observer tries to answer questions such as "why does *this* behavior happen at *that* time and *what* purpose does it serve in the interactional process ?". It follows from Garfinkel's theory that the analysis of human behaviors relies on processes of *reflexive mutual accountability*. The general idea is that the problem of interpretative relevance of human behavior is insoluble in theory and that scientists should pay close attention to how participants solve interactional problems in practice according to a specific situation. In that sense, Ethnomethodology studies the methods that participants use and develop to make sense of situated social cues instead of studying norms. However, CA still relies on reflexive normativity to explain how participants optionally conform to normative patterns of interaction and use these to situationally deviate from normative expectations in an orderly manner [99]. CA avoids relying on these norms to formulate hypotheses and theories to be tested outside of the context of discovery. However, while there is a contrast between two ways of approaching the relationship to empiricism, this does not mean that CA is entirely incompatible with psychological theories, especially those adopted by the HRI community.

Interactionist Sociology thus avoids invoking concepts that are not firmly grounded in natural observation. Heritage insists that research should be solely conducted using "data collected from naturally occurring occasions of everyday interaction" [7], and should never include any pre-coded schedules for the interaction, nor should they try to direct or manipulate the participants' behavior. Behavior analysis begins by highlighting recurring patterns or skewed distributions of some candidate phenomena, and by collecting sufficiently large amount of data to draw any conclusion. In that sense, Interactionist Sociology can help Psychology face its current

"replication crisis". It has been shown that several published findings of experimental studies are hardly reproducible [82]. The context of discovery and context of justification of a theory in Psychology are generally separated [99], while Ethnomethodology-CA methodologies lead to the two being kept together. CA discoveries could be used to generate working hypotheses for Psychology.

While we leverage Interactionist Sociology theories for the theoretical methodology of trust analysis, we rely on falsificationism for our computational studies while formulating hypotheses based on Interactionist Sociology theories and discoveries, as we need to validate assumptions that were used to build our models after our experiments.

3.1.2 Implications for trust analysis

Interactionist Sociology relies on the observability of trust within the interaction, which is made visible by the participants themselves through their behaviors [35, 36, 44]. Trust is thus a result of the state of the interaction, and is oriented towards both the content and the format of the interaction. In a trusting state, participants tend to behave in a way so that the interaction is fluid and proceeds towards its objective [32]. It is observable on different bases: e.g. trust in the robot capabilities to maintain a fluid and progressive interaction, in its knowledge, its skill in accomplishing a specific action at a given moment. Given this, we define interactional trust as a *"form of affiliation and credit characterized by a set of behaviors that are intentional or not, expressive or propositional"*. This definition relies on concepts of *alignment* [112], *affiliation* - claiming access to and understanding the partner's stance, and endorsing their perspective - [110], and *credit* [26, 88] given to the robot *competence* [33]. Credit is the recognition of the relevance and suitability of the partner's message or social behavior in the interaction context. In a way, our definition of interactional trust relates to the ecological validity of the robot as an autonomous agent.

We leverage concepts from Interactionist Sociology to build the ground of our analysis method and determine which type of behavior the observer should focus on. We break down the analysis of trust in sub-concepts that constitute our working definition of trust, and focus the observer's analysis on affiliative, aligning behaviors as well as those that attribute credit to the robot competence. The analysis thus does not make any pre-observation assumption that should be validated with post-analysis statistical tests. Rather, the observer determines the relevance of each participant's action in relation to trust, and to the interaction history and context.

We slightly deviate from Heritage's suggested methodology as we need a predefined coding system for our computational models that are trained in a purely supervised method. We do not address the issue of semi-supervised nor unsupervised methods such as few-shot learning, nor do we rely on domain-adaptation methods to train our models - mainly due to the lack of data availability for trust in HRI. Current machine learning methods for HRI require a stable structure of labels [94]. While the choice of categories is generally guided by the task, by what the system must detect in a top-down approach, we adopted a more bottom-up approach in the

construction of our coding scheme for trust analysis in HRI which we present in the following section.

Another important point in Interactionist Sociology is the necessity to keep the analysis of behavioral cues rooted in the interaction context. This is done to ensure that the analysis keeps track of the relevance of cues produced by participants as answers to past cues of other participants and as resources in the production of other participants' turn. However, the process of annotation decontextualises the cues from the interactional history. Having a more systematic approach to annotations such as what is done in CA's transcriptions can be a small step towards recontextualisation of the cues. An example of CA transcription is provided in Figure 3.1. In turn, this can help machine learning models to better learn detecting trust.

```
01    P     ablas esp[agnol
                hablas español
02    R               [une autre fois\
                        another time
03    P     <oh: ((look at smartphone)) (0,5s) > (0,5s) ok (..)je ne
04          sais pas qué: qu'est-ce qué tou (1,1s) dire\
              oh, ok I don't know what, what do you..say
05          (3,6s)
06    P     <((with greeting gesture)) au revoir/>
                                        goodbye
07          (6,2s)
08    P     <((with greeting gesture and body torq)) au revoir Pepper\>
                                                    goodbye Pepper
```

Figure 3.1: Example of a multimodal CA transcription. In this excerpt, there are two identified speakers: the participant "P" and the robot "R". Each line indicates what each one of them says, with additional information. On line 01 and 02, the bracket "[" indicates that the participant's speech and robot speech overlap. Actions between "¡" and "¿" happen in synchrony, such as on line 02 where P says "oh" and looks at his smartphone placed between the double parentheses. Times in parentheses show a pause in the speech, and "(..)" points to a short pause. Semi-colons ":" are used to show a prolongation of the sound placed right before. The symbols / and
respectively indicate a rising or lowering tone [94].

This paradigm shift also impacts the segmentation process of machine learning methodology. The standard methodology involves defining a fixed time length to determine a window of analysis, during which features are extracted. Models are then trained to predict the target phenomenon on that window length. The window length is determined according to the phenomenon that is being studied. A general rule of thumb is to choose a length of five seconds

[74]. However, breaking down the interaction based on a fixed time length inevitably results in segments that do not correspond to participants' turns, hence leading to further decontextualisation. One way to minimize this is to ground the annotation and analysis in the interaction structure, as is done in CA.

3.2 TURIN

3.2.1 Describing the coding scheme

3.2.1.1 General Overview

In this section, we present the details of the TURIN (Trust in hUman Robot INteraction) coding system, a flexible framework to study trust in HRI that can be adapted for both dyadic and group interactions. To the best of our knowledge, it is the first coding system that was conceived for trust in HRI. While we bring concepts from Human-Human Interaction (HHI), the coding scheme is specific to HRI. We focus on participants' trust towards the robot and define subcategories according to the linked behaviors. Following is the description of the unitizing process.

We propose coding trust in segments describing time periods of homogeneous trust level, at the individual level for dyadic interactions, and at the group level for group interactions. The segmentation should start at the single behavioral act level. Individual acts referring to changes in behaviors should be assigned to a trust category. Consecutive acts of the same category should be aggregated to form a segment of homogeneous level of trust. In group settings, trust is considered to be an emergent state of the group and so takes time to develop and change. Segments are delimited so that each corresponds to a same group-level emergent trust content, starting with a member's behavior indicating this trust category, and ending with a change in a participant's behavior indicating another trust category. In group settings, we prefer avoiding small segments. We would like to have a broader view of the emergent trusting behavior of the group. Indeed, emergent states of the group take longer to fluctuate than states for a single user as they involve the coordination of multiple users [41, 90]. Following this, we suggest adopting a sligthly more macro view of the segments and coding according to the function of the undergoing interactional process. For instance, extremely short trusting behaviors inside a longer "mistrusting" sequence should still be coded inside a "mistrusting" segment. In dyadic HRI, the dynamics of trust will be different with quicker changes, but the expression of these changes will fundamentally remain similar.

Segments are assigned to a trust category, namely either "*trusting*", "*mistrusting*" or "*neutral*" depending on whether users display respectively trusting, mistrusting or neutral behavior. The following subsection defines those trusting, and mistrusting behaviors. Inside those segments, behaviors are described by items from four sub-categories that are detailed in Section 3.2.1.2: "*Social Interaction Form*", "*Interaction Content*", "*Benevolence*", and "*Integrity*".

3.2.1.2 Nonverbal and verbal trusting behaviors

Based on our definition, any type of behavior that displays interactional trust, accepts vulnerability, seems friendly, or that acknowledges the partner's competence can be interpreted as a trusting behavior. We define interactional trust as a state displaying a form of naturalness, or fluidity in the interaction. Naturalness implies that the robot is treated as an interactional partner in the same way as a human partner would be treated [95]. Naturalness is estimated based on the dynamics of participants' behaviors rather than on the robot embodiment and behaviors. For instance, making jokes shows that a user trusts the robot to understand and react to that joke. Mistrusting behavior is any type of behavior that displays uneasiness, doubt, confusion, aggressiveness, or unwillingness to cooperate. Any other type of behavior should be coded as neutral, as it means that it is a type of behavior that is not expressive enough to draw any conclusion.

Following our definition, and Mayer's definition of trust [71] as a *"positive expectation towards the ability, benevolence and integrity of the trustee"*, we suggest coding specific behaviors to four subcategories inside segments tagged as mistrusting or trusting: Social Interaction Form, Interaction Content, Benevolence, and Integrity. All coded non-verbal behaviors relate to users, not to robots. According to its definition, human's trust towards the robot depends on the partner's benevolence, and integrity. We hypothesize that showing benevolence and integrity could indicate trust, as benevolent behavior has a social cost and reciprocity is not certain. The user presents himself as vulnerable by being benevolent and expects his partner to reciprocate. "Social Interaction Form" represents all low-level behavioral signs that are a failure in a social interaction norm or that signal a high level of naturalness of the interaction. "Interaction content" relates to events, behaviors or words from the current undergoing task signaling either trust or mistrust. Behaviors or events that display benevolence or malevolence from a user towards the robot should be classified under the "Benevolence" tag. "Integrity" stands for any behavior or event that signal a user's integrity, or a lack thereof. These last two categories serve as affective descriptors of the multimodal behaviors. The content of the sub-categories are given in Table 3.2.1.2. Some of these items are group-specific.

Though Hancock's [42] and Schaefer's [104] definitions are widely used in other studies, their definition relies on the concepts of "uncertainty" and "vulnerability" which are heavily context and task dependent. Mayer's definition on the contrary allows to describe trust without relying on the ongoing task of the interaction. Most items can be used for both trusting and mistrusting segments. Items marked under "trusting only" or "mistrusting only" in the table should only be used inside trusting and mistrusting segments respectively. Benevolence, Integrity, and Interaction Content items are descriptors of Social Interaction Form items. Social Interaction Form items are not necessarily linked to other categories if no item fits as a descriptor of the behavior. Not all behaviors should be coded, only those that are relevant to the trust category. The categories hierarchy is presented in Figure 3.2.

Social Interaction Form	Interaction Content
Gaze	Compliance
Facial expression	Cooperation
Nod	Alignment
Gesture	Approval
Phrasing	Out-of-context comment
Intonation	**Trusting only**
F-formation*	Joke
Speaking turn	**Mistrusting only**
Repetition	Doubt
Participation status*	
Benevolence	**Integrity**
Respect	Honesty
Personal info disclosure	Responsibility
Warmth	Promise
	Mistrusting only
	Manipulation

*Group-specific item

Table 3.1: Sub-categories and items of the TURIN coding system

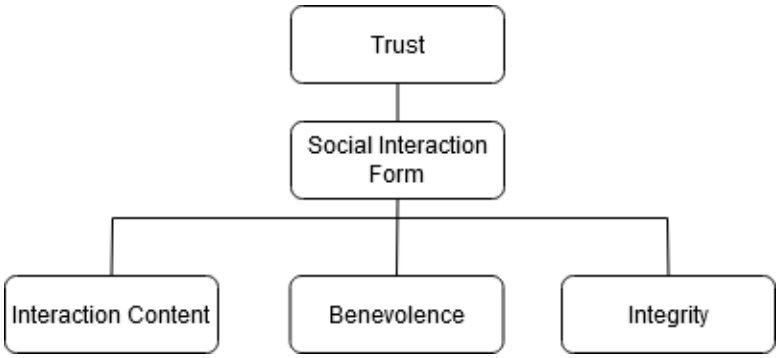

Figure 3.2: Hierarchy of categories of the TURIN coding scheme.

3.2.2 Social Interaction Form

In this section, we describe different modes of trust-related affective behaviors that are linked to norms of social interactions. Description of items are provided in Table 3.2.2. Not all behaviors should be coded, but rather only manifestations of trusting and mistrusting behaviors should be coded.

Item	***Description***
Gaze	Gaze explicitly indicates what the user focuses on. It is a hint of trust when it aligns with the current task, or of confusion and mistrust when it is not fixed on a specific object but rather sweeping across the interactional space at the search of something [18, 38].

Item	***Description***
Facial expression	Facial expressions convey emotional information that can indicate that users acknowledge a failure from the robot or signal warmth towards it. For instance, after a failure, users will typically react by raising their brows, or by frowning [3, 74]. Users may smile when reacting to a robot utterance to signal an understanding and approval during trusting segments as a way of providing positive feedback.
Nods	Nods can be indicators of affiliative and aligning behaviors [8]. Nods are generally produced as a sign of passive feedback called back-channelling [127] during which the listener conveys information to the speaker without taking its turn away. In this way, nods can convey feedback on different levels, such as attention, hearing, understanding, or acceptance of the speaker's discourse. Nods can thus relate to both trust and mistrust. In trust, nods generally are signals of approval of what is happening in the interaction, meaning that users trust the flow of the ongoing task [127]. In mistrust, nods are hints that a failure happened, e.g. users are waiting for the robot to proceed with his utterance after an abnormally long pause and so nod to show that they are waiting [74].
Gesture	Gestures in trusting segments indicate a very natural communication with the robot. For instance, deictic gestures rely on the ability of the robot to know his environment and identify which object is pointed at. They could also be interpreted as a method to disambiguate references. In mistrusting segments, gestures can be a manifestation of confusion, frustration or aggressiveness, or can be used as support during trust-reparation sequences, e.g. a user might indicate disagreement by crossing their arms associated with scowling brows [72].

Item	***Description***
Phrasing	Phrasing is the choice of words used to express something, an element that has not been given much attention in HRI. Robots currently do not have the ability to understand human speech on a deep meaning level. In most studies, robots act in a Wizard-of-Oz setup. This means that they can only react to what users say and do in a very limited way. Robots' understanding during experiments is limited to a very specific range of sentences at best, very few words at worst. Due to that, users might not always be able to formulate an answer as they would want to. Users that exhibit trusting behaviors tend to phrase their sentences very naturally, often with complex grammar, while mistrusting users tend to answer with very concise and grammatically minimal sentences [81].
Intonation	Intonation can convey information about the emotion of the speaker [81]. Moments where intonation signals warmth should be associated to trusting behavior, whereas intonation indicating confusion or aggressiveness reveals a mistrusting behavior.
Speaking turn	Speaking turn enables the assessment of the interactional flow. Users trusting in the robot interactional ability tend to react quickly to the robot utterance, up to the point where the robot and the user's utterances bounce back at each other dynamically [31]. Overtrusting users may also spontaneously take the floor, imagining that it is their turn to speak. On the opposite, longer response timings could reveal doubt and uneasiness towards the interaction. Marking a pause to think is not a sign of mistrust though.
Repetition	Repetition occurs generally after failure of either the speech synthesis or speech recognition module, indicating a trust failure for the user [115]. In some cases, users might repeat a sentence as a sign of trust in the robot capabilities to adapt its speech subject.
F-formation *	F-formation refers to how people arrange themselves in a spatial environment [21]. A change of formation into a new one that still includes everyone as active users signifies a trusting behavior, whereas a change into a formation that turns away one or more users reveals a distrusting behavior. We also include proxemics with this item, as it was shown to correlate with social proximity [3, 76]. Standing closer to a robot is thus an indicator of trust, while being further away than normal communicates mistrust.

Item	***Description***
Participation status *	Participation status refers to the ratification of the robot as a member of the human group interaction [35, 36]. The robot is considered to be fully ratified if he is an addressee or considered as a side participant (e.g. two users are whispering together while figuring out the best answer to the robot question). While the robot being ratified as an addressee is a very neutral stance, ratification as a side participant is a subtle form of social trust. It is indeed a proof that users pre-reflexively apply advanced social interaction norms with the robot. Leaving the robot unratified means considering him as either a bystander or an overhearer. That can be viewed as a mistrusting behavior.

*Group-specific item

Table 3.2: "Social Interaction Form" sub-category items description

3.2.3 Interaction Content

In this section, we detail the items relating to behavioral cues through their function in the communication. Description of items are provided in Table 3.2.3.

Item	***Description***
Compliance	Compliance to the robot indications and orders can be related to authority in some cases, but we hypothesize that it could also indicate a form of social trust in the robot social role or simply relate to the user's cognitive trust. Previous research showed that users tend not to rely on the robot in an emergency situation if the robot previously exhibited faulty behavior [93]. Here, compliance breaks down to relying on the robot. It could also be a form of credit given to the social role of the robot, e.g. following the robot indications as a museum guide [30] means believing the robot delivered the right information as expected of his role as a museum guide.
Cooperation	Choosing to cooperate with the robot or delegate a task to him simply means relying on its capabilities, which is exactly the definition given by Rousseau [97].

Item	***Description***
Alignment	Alignment comprises of several communicative behaviors, both verbal and non-verbal. Verbal alignment can happen on different levels: lexical, syntactic, and semantic. Research showed that alignment increases social affiliation [86]. As alignment is about coordination and social connection, we hypothesize that users aligning with its robot partner is an indicator of social trust.
Approval	Approval-related behaviors, both verbal and non-verbal, can be linked to a recognition of the partner's ability or an indicator of social proximity, or even complicity. [18, 38].
Out-of-content comment	Sometimes, users make out-of-context comments addressed to the robot. These comments can be sly or candid. Sly comments can take the form of competence tests, that is random out-of-the-blue questions whose goal is to assess any of the robot ability (e.g. "what is the derivative of cosine?" when addressing a cooking robot). Candid comments on the other hand can take the form of comments that come from the original context of the conversation but go overboard, indicating that the user over-trusts the robot capabilities to handle a dialogue.
Jokes	Jokes are an advanced form of social interaction as it implies being able to detect that a joke was made and decode it. Joking suggests that the user believes the robot to be able to perform all of that. This item should be coded inside trusting segments only. [18, 38].
Doubt	Doubt is the last item of this category and it refers to moments where users express hesitation towards what the robot says because it is unclear whether the robot is right or not. This item should only be coded inside mistrusting segments. [18, 38].

Table 3.3: "Interaction Content" sub-category items description

3.2.4 Benevolence

Any explicit display of respect is a good indicator of the user recognizing the robot as being worthy of a special attention and care. Exhibiting respect means that the user expects the robot to understand the social concept of respect, which we hypothesize to be a form of social

trust.

Being able to disclose personal information is accepting the vulnerability to open up to another party by expecting it not to share the information abusively.

The last item refers to social warmth. We consider the two aspects of social warmth to be friendliness for trusting behavior and aggressiveness for a mistrusting one.

3.2.5 Integrity

The concept of integrity is based on a social contract of respect to social norms of honesty and moral principles. Individuals that break away from these norms create uncertainty towards their interaction partner and thus might prove themselves as being hard to rely on. While integrity can be seen here as normative, we also formulate the hypothesis that showing integrity is an expectation of reciprocity. Items from this subcategory come from that idea of upholding shared moral principles and proving oneself to be trustworthy. The item "manipulation" should only be used under "mistrusting" segments as it refers to manipulative behaviors for one's own personal gain.

3.2.6 Adapting the coding system

The coding system remains flexible to the needs of each research, as macro and sub-categories can be refined (e.g. one could introduce a separate "overtrust" macro-category).

Inter-group differences about trust were not integrated in this paper. Annotating trust in a group setting will inevitably lead to asymmetrical situations where a user exhibits trusting behavior and the other users are neutral for instance. Studying the intra-group symmetry comes down to studying the affiliation, alignment, complicity, detachment and shyness that occurs between individuals within the group. It has been shown that trust asymmetry in groups moderates the positive relationship between mean levels of team trust and team performance such that it becomes weaker as trust asymmetry becomes higher [41]. We believe that refining the coding system with information about trust symmetry will lead to a better understanding of the dynamics of trust in a group. It can also provide information about the relationships between participants.

3.2.7 Choice of the Vernissage corpus

Four publicly available HRI datasets were considered as test-beds for the coding system: JOKER [24], MHHRI [19], UE-HRI [13], and Vernissage [47]. To test TURIN, we adopted the Vernissage dataset. As we wanted to validate all items from the "Social Interaction Form" sub-category, we discarded the JOKER, and UE-HRI datasets as they are based on dyadic interactions. We also discarded the MHHRI dataset due to its scenario: during the triadic HHRI phase, the robot addresses only one of the participant, hence leading to an asymmetrical situation, which we have not fully considered in our annotation system.

The Vernissage dataset contains multimodal data through recordings of the interactions including several video views, separate audio files for each user, the users' head motion captured by a motion capture tool, and logs of the robot movements. This corpus sets up an interaction between a robot and 2 users. The NAO robot explains paintings in the room and then quizzes the participants on art. Once the paintings presentation is over and right before the quizz, the participants are asked to present themselves by giving more than just their names.

The corpus is composed of 10 interactions, each lasting around 11 minutes (mean length M=11min 21s, standard deviation s=51s), during which 4 minutes were dedicated to the art presentation section, 2 minutes to the self-presentation prior to the quizz, and 5 minutes to the quizz. The robot was controlled in a Wizard-of-Oz setting. Users' behaviors were unconstrained.

3.2.7.1 Validation of the coding scheme

Two experts on HRI annotated 1 minute of 3 videos of the dataset[1]. Annotations were performed through the ELAN software platform [122]. An example of annotations is provided in Figure 3.3.

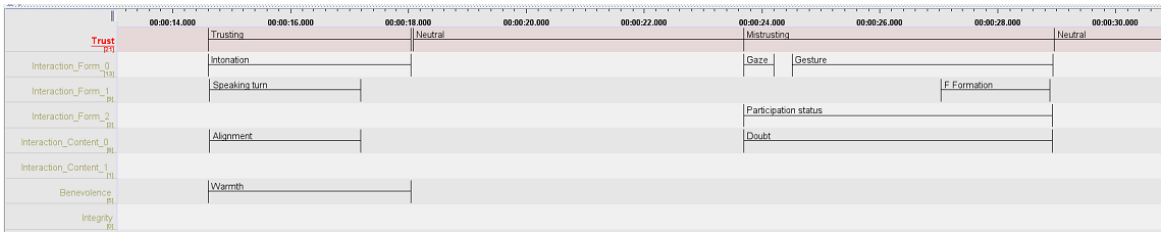

Figure 3.3: An example of trust annotations using the TURIN coding scheme on the ELAN software for the beginning of an interaction. At first, participants express trust as indicated by their intonation, and a speaking turn taken promptly after the robot, indicating alignment. They also signal warmth through their speech. After this, participants are neutral, and then exhibit mistrust as indicated by their gaze, a change in participation status, gesture, and a change in F-formation. Through all these behaviors, they express doubt towards the painting that they should focus on.

The inter-rater agreement (IRA) between the experts was computed through the Cohen's kappa [22] implemented in ELAN. The results of the analysis are given in Table 3.2.7.1. The high agreement rate of the "Mistrusting" segments compared to the other categories could be explained by the fact that it is easier to recognize errors in norms of social interaction than segments where behaviors are extremely natural as we defined as "trusting behaviors". The subjectivity of the task might also increase the difficulty of the recognition task. The moderate agreement rate on the "Neutral" segments are due to disagreements on behaviors happening inside long "Neutral" segments (maximum length is 22.9s). These disagreements resulted in a division of one long segment into multiple smaller ones, which are mostly coded as "Trusting", thus impacting the agreement score.

Trust as an emergent state was established as slow to change in the literature, meaning that it takes several minutes to evolve. The average length of coded segments does not align

[1]Complete annotations on all 10 interactions are published here: `https://doi.org/10.5281/zenodo.8409887`

Segment category	**IRA (κ)**	**Mean duration (s)**	**Std (s)**
Mistrusting	0.79	4.6	2.2
Trusting	0.64	2.1	1.5
Neutral	0.45	4.7	4.6

Table 3.4: Inter-rater agreement on segments extracted from the Vernissage dataset

with this, which could be explained by 2 factors. First, the experts annotated only the first minute of the interaction, which might mean that the group's dynamics are unstable at the beginning and need time to stabilize. Overtrusting groups behave in very natural ways and inevitably encounter moments where the robot does not meet their expectation of fluidity in the interaction. Through trial and error, the group will find the proper way to interact with the robot given its capabilities. Earlier steps of the interaction might correspond to this "trial and error" phase while later steps will correspond to adjusted behavior with a more stable trust state. Coding the entire available interactions would allow us to conclude on this hypothesis. A second explanation might be linked to our methodology that suggests coding Trust according to changes in behavior, leading to relatively short segments.

Segment category	**IRA (κ)**	**Mean duration (s)**	**Std (s)**	**Count**
Nod	0.52	1.4	0.6	15
Gaze	0.36	1.6	1.4	27
Gesture	0.56	1.9	1.3	7
Phrasing	0.42	1.3	0.5	3
Repetition	0.89	0.9	0.1	2
Intonation	0.74	1.6	1.0	7
F Formation	0.80	2.1	0.9	10
Speaking turn	0.80	1.3	0.8	26
Facial expression	0.41	1.3	0.8	19
Participation status	0.80	3.5	1.9	12

Table 3.5: Inter-raters agreement on coded "Social Interaction Form" items from the Vernissage dataset

The results of the analysis for coded "Social Interaction Form" items are given in Table 3.2.7.1. They were computed using a different algorithm of the Cohen's Kappa. This version computes the overlap of coded segments by subdividing them in smaller segments of 200 milliseconds long, as a way to condense the computation of agreement on segmentation and the agreement on category attribution. The fair to moderate agreement on "Nod", "Gaze", and "Facial expression" could be explained by the difficulty to define the start and end of these annotations in relation to trust. The experts also coded them slightly differently: one coded these behaviors as one segment per user, while the other coded them as one segment for the entire group. The coding scheme specifies that behaviors should be coded for the entire group, by focusing on behaviors that are most indicative of the chosen Trust label. However, one of the annotator found it difficult to adopt a perspective on the group rather than looking at participants individually. The use of automatic tools to code these items might be beneficial

to improve the precision of annotations.

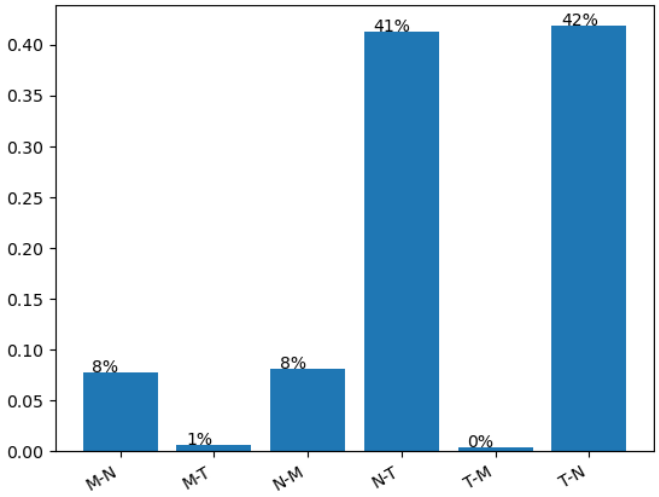

Figure 3.4: Probability of label pair transition patterns occuring with our annotations on the Vernissage dataset.
M: Mistrusting. N: Neutral. T: Trusting.

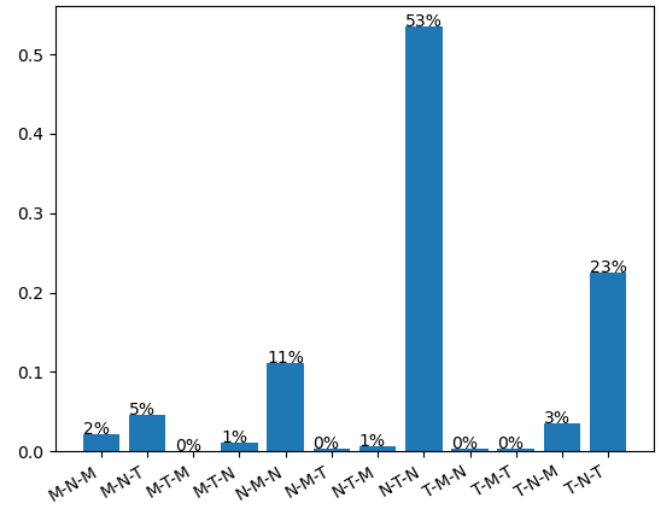

Figure 3.5: Probability of label triple transition patterns occuring with our annotations on the Vernissage dataset.
M: Mistrusting. N: Neutral. T: Trusting.

Figures 3.4-3.5 show the patterns of label transition as occuring through our annotations. For pair transitions, unsurprisingly there are a majority of transitions between "Neutral" and "Trusting" as these are the two most common labels. We can also observe that transitions between "Mistrusting" and "Trusting" are very scarce, mostly due to the fact that "Mistrusting"

labels are rare in themselves. If we look at triple transitions, N-T-N patterns represent half of our dataset, due to the dominant presence of "Neutral" labels. M-N-T patterns are very scarce, indicating that recovery from trust failure takes time to happen. Similarly, T-N-M patterns are rare, showing the stability of behaviors exhibited by participants.

3.3 Comparison with the mentalist approach

In this section, we compare 2 trust assessment tools - one from the mentalist perspective and one from the interactionist perspective - through a pilot study to investigate which approach is best suited to build machine-learning models for a fine-grained analysis of trust throughout the interaction, and investigate how both approaches can complement each other.

3.3.0.1 Procedure

Five experts in HRI participated in the study. We collected annotations on the first three minutes of the 10 interactions, focusing on the first three paintings of the vernissage phase for this preliminary study. Expert A annotated all 10 interactions with both approaches. They had previous training and experience with TURIN before conducting our study task. Annotations were collected with a time lapse of a week between both methods to reduce the influence of one task on the other. Experts B and C annotated five different interactions with no overlap between them with the interactionist tool. Experts D and E did the same with the mentalist one. The experts annotated fixed-length windows of 10 seconds, yielding a total of 18 segments per interaction that are annotated using one assessment tool from each approach, that is the RTS for the mentalist approach or TURIN for the interactionist approach. When choosing the windows length, we reached a compromise close to a few speaking turns for TURIN to still be able to highlight behaviors in a relevant time-frame, and long enough for the annotator representation of RTS items to evolve. While the RTS can be filled by annotators right after being presented and items definition clarified, TURIN requires annotators to be trained before using it. We therefore trained the experts for an hour on interactions that they were not going to annotate.

3.3.0.2 Tool adaptation

We chose the RTS as an annotation tool for the mentalist approach. The RTS is the most comprehensive among other previously cited tools [104] and is the only one to be correlated with post-trust interaction [105], showing that it is able to measure trust variations. We used the RTS reduced to 14 items version. We chose the reduced version of the RTS to limit the cognitive load of the annotation process. Furthermore, many items from the full scale focus on robots' technical and social skills that are too general - e.g. "Protecting people", "Warning people of potential risks in the environment", "Performing many functions at one time". As such, they are irrelevant in the context of the Vernissage experimental scenario. Among the 14

items, we considered the items "perform exactly as instructed" and "follow directions" to be unrelated to the task since the robot acts as an art guide, and is thus mostly in charge of the conversation and never has to follow any of the participant's instructions. We operate changes on the RTS for our task given its constraints. One of the constraints is that annotations are required to be collected during regular and small time-frames. The RTS is generally used at the beginning and end of the interaction as it takes time to fill given the amount of items. Interrupting the interaction in such a way would disrupt its flow. As a consequence, there are no publicly available dataset that includes annotations collected in such way. Annotations should be conducted from an external observer's point of view. We thus had to adjust the RTS point-of-view since it was not designed to assess participants' trust by an external observer. As there is no mentalist assessment tool with a third-person view and for the tool to fit our task, we asked the annotator to consider them-self as a bystander of the interaction. From the observation of the participants' behaviors and reactions to the robot, they built their own perception of the robot which they used to fill the RTS. Hung et al. performed such translation of questionnaires in a third-person point of view to study the cohesion of small human groups based on nonverbal audio-visual behaviors [46]. We did not ask the annotator to try to infer participants' state of mind from their behavior as the RTS items relate to perceptions of the robot skills and do not relate to user-centered criteria. Considering this issue, and to avoid interpretation bias, we asked the expert to annotate its own perception of trust towards the robot.

We relied on our coding scheme TURIN for the interactionist approach. For the annotation comparison study, we adjusted the TURIN unitizing method to fit the task constraints. First, we changed TURIN unitizing method by collecting annotations based on fixed-length windows, even though TURIN specifies a unitizing method that relies on the aggregation of moments of coherent trust category. Even though this unitizing method is not grounded in the delineated interactional processes, these processes are still visible within a segment. Thus, we decided here to focus on the dominant trust category and TURIN sub-categories that are made visible by users within a segment. The annotation length from TURIN subcategories does not necessarily match the segment length in the original approach. Subcategory items are used originally to describe behaviors that happen during a time-frame that is relevant in the interaction structure. The annotator has to choose at most 2 annotations from the "Social Interaction Form" sub-category, at most 2 annotations from the "Interaction Content" category, at most 1 from the "Benevolence" one, and at most one from the "Integrity" one. We limited the annotations in such a way to only report behaviors that were the most salient inside units and avoid having too many annotations about punctual behaviors.

We applied all previously described adaptations of the tools as a way for them to meet a common ground for comparison purposes. As there are no publicly available HRI datasets that contains RTS annotations, we had to make more adaptations to it than TURIN for this pilot study. Figure 3.6 provides a summary of the annotation process for the interactionist and the mentalist approach.

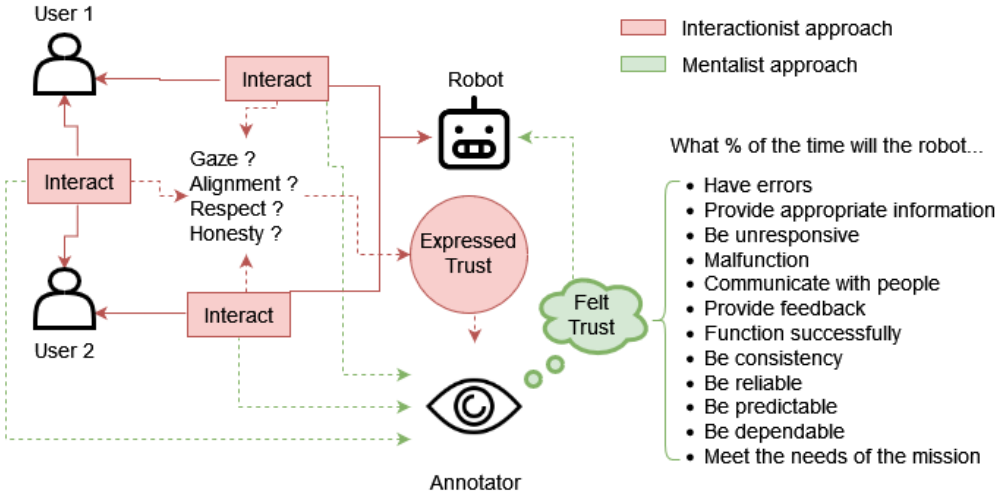

Figure 3.6: Through interactions between all group members, the group makes observable to the annotator the expressed group trust. In the interactionist approach, the annotator relies on observable and tangible evidence of members' behaviors to assign a trust label from TURIN. In the mentalist approach, the annotator relies on its interpretation of these interactions and its own perception to assign a score to each criteria of the RTS.

3.3.0.3 Results

To compare both approaches, we differentiated segments based on their assigned TURIN trust label. To aggregate the annotations from all experts, we first rescale the scores of each item from the RTS for each annotator (A, B, and C). We first perform a Shapiro-Wilk test to assess whether each item score for each annotator come from a normal distribution [107]. The results show that none come from a normal distribution, $p < .001$. We thus operate a min-max scaling of each of the RTS items score for each annotator. We then searched for statistical differences in the mean score of each of the RTS items depending on the assigned TURIN trust label to the segment. We first applied a Kruskal-Wallis test [55]. If the test reveals significant difference, it is followed by a post-hoc Dunn test with Bonferroni correction [25]. We plot the score distributions and report all results of the statistical tests in Figure 3.7.

First, we observe statistically significant differences between "Mistrusting" and "Trusting" segments for items "Function successfully", "Malfunction", "Errors", "Feedback", "Communication", "Reliable", and "Unresponsive", $p < .05$. This is due to the fact that participants strongly react to faulty behaviors from the robot, for instance when the robot ignores the answer of participants after asking them a question, or when it fails to recognize the participants' name. Except for item "Errors", the test reveals significant differences between "Mistrusting" and "Neutral" segments for all previously cited items.

Other items "Consistency", "Mission needs", "Appropriate information", "Dependable", and "Predictable" appear independent of TURIN trust labels: these items can take any value, TURIN labels will not necessarily reflect the RTS trust label. Looking at annotations closely, some participants still align and comply with the robot even when it displays faulty behavior, while others might still express doubt towards the robot even if it functions perfectly well.

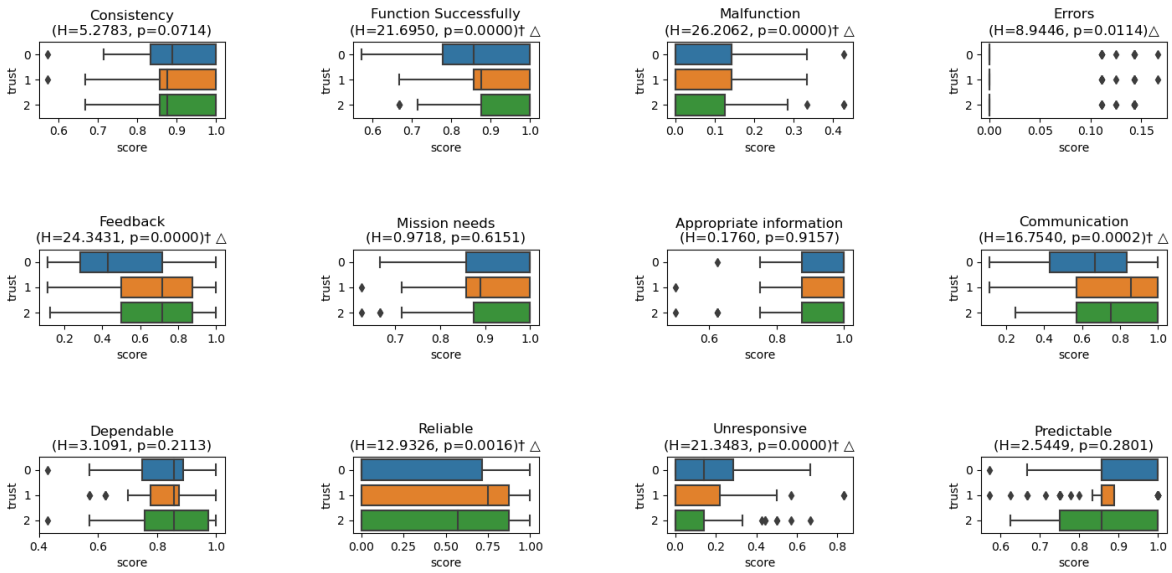

Figure 3.7: Score distribution of the 12 items for the RTS and correlation with TURIN trust annotations. Trust values: 0="Mistrusting", 1="Neutral", 2="Trusting".
† : significant score distribution difference between Mistrusting and Neutral segments.
△ : significant score distribution difference between Mistrusting and Trusting segments.
○ : significant score distribution difference between Neutral and Trusting segments.

We then studied the inter-rater agreement (IRA) between annotator A and annotators B, C, D, and E. For TURIN, we computed the Cohen's Kappa [22] between A and B for each interaction and the overall agreement on all interactions, and did similarly for A and C since there is no overlap between B and C. The overall Cohen's Kappa is rather weak with TURIN, 0.37 between A and B, 0.35 between A and C. This can be explained by the choice of using fixed-length windows to collect annotations and compare both approaches. Since windows are not rooted in the structure of the interaction, annotators may decide to highlight different phenomenon that may happen within these. It can also be explained by the choice of the window length which is long. When investigating the mismatches between annotators, we observe that the highest number of mismatches appear when one of the annotators chose the "Neutral" label. Assigning the "Neutral" label requires the annotators to evaluate whether behaviors are not significant enough to be considered signs of trust or mistrust. This highlights the difficulty of the annotation task given the constraint on windows length.

We then computed the Cohen's kappa for TURIN subcategories "Interaction Form" and "Interaction Content". For "Interaction Form", the Cohen's Kappa is poor, $\kappa = 0.13$ between A and B, $\kappa = 0.09$ between A and C. The low value can be explained by the constraints of the task: the experts had to choose at most two from many items that were dominant in the segment. They reported during a post-annotation interview that this limitation in the choice of the item made the item selection difficult given the length of the segment. They reported

that they often would have like adding a third item. The experts mostly annotated with items "Gaze", "Facial expression", and "Intonation" For "Interaction Content", the Cohen's Kapp is poor between A and B, $\kappa = 0.08$, but weak between A and C, $\kappa = 0.24$. The low Kappa value can be explained by the fact that this subcategory is a descriptor of the "Interaction Form" subcategory. As such, it relies on previous item selection during this task and depends on what the annotator chose to focus on in the segment. Items that were the most used for this subcategory were "Alignment", "Approval" and "Compliance"

For the RTS, we considered each category set of values through the interaction as time-series. Therefore, we computed the Cramer's V correlations between annotators for each category, and averaged them for an interaction as we observed no negative correlations between annotators. Correlations are pretty low, $V = 0.16$ between A and D, and $V = 0.22$ between A and E. This result highlights the subjectivity of the task, as expected from a mentalist approach. Categories "Errors", "Mission Needs", "Reliable", and "Dependable" yield the lowest correlations. This is explained by the content of the interaction: the robot acts as an art guide, and as such, the mission needs may not appear very clear to annotators. As there is no explicit vulnerability in the experimental scenario of the dataset, dependency and reliance on the robot from participants may also appear unclear. Category "Errors" low correlations might be explained by different annotation practices from annotators : expert E used higher ranges of scores than expert A, thus increasing the size of the contingency table.

During post-annotation interviews, all annotators pointed out that interactions 2, 3, 9, and 10 were significantly harder to annotate than the others. Annotators reported that the discrepancy between the enthusiasm of some participants and the faulty behaviors of the robot made the annotation task difficult for these interaction. Annotators were also unsure whether some participants were sometimes acting sarcastically or not in these interactions. Moreover, in interaction 9, one of the participants has trouble understanding Nao and often asks the other participant to translate in their native language. While they point out Nao's faulty behaviors, they still show signs of trust by asking it to slow down for instance. Because of this, annotators expressed difficulty in choosing the appropriate TURIN trust label.

3.3.0.4 Differentiating criteria

	Time-framing			**Orientation**		**Generalization**		**Scalability**	
	BU	ST	EI	Data-driven	TF-driven	Specific	Generic	Individual	Group
Mentalist			X		X		X	X	X
Interactionist	X	X		X		X		X	X

Table 3.6: Summary of the comparison of the mentalist and interacionnist approach based on 4 criteria. BU: Behavioral Unity. ST: Speaking Turns. EI: Entire interaction. TF: Theoretical-framework

We identified 4 criteria, which we detail in the following sections, on which both approaches differ from our theoretical and annotation comparison study : *orientation*, *generalization capability*, *time-framing*, and *scalability*. A summary can be found in Table 3.6.

First, their *orientation*, the preference for theoretical-framework-driven or data-driven tools, diverges. Our theoretical analysis showed that the mentalist approach has led to the creation of rather **theoretical-framework-driven** assessment tools, while the interactionist assessment tools are more **data-driven** as study results solely rely on the close examination of users' behaviors that emerge from the data. There is a tension between considering trust as a mental state, and admitting that given that the participants do not have access to their partner's brain, it is towards the observability of this supposed state that the participants are oriented. They can nonetheless try to infer it through the partner's behaviors and decisions [58]. However, past studies show that even when a robot displays faulty behaviors that are detrimental to trust, participants sometimes still decide to follow the robot advice, e.g. during a fire alarm [93, 103]. This is confirmed by our study. Indeed, in the Vernissage scenario, the robot often shows difficulties in grasping the users' names and asks the participants to repeat. RTS annotations indicate that participants' trust is low after that. But TURIN annotations show that participants still trust the robot to refer to the correct paintings right after the mistake. This discrepancy between the user's expected behavior and its actual behavior shows the difficulty of inferring the user's mental model of the robot trustworthiness [51, 105].

Next, they differ on their *generalization capability*, based on how specific or generic their analysis is according to the interaction task. The mentalist approach is **quite generic** and not dependent on the interaction history. Assessment tools items cover a wide range of concepts relating to trust that do not depend on the interaction task. However, the pilot study shows that some items from the RTS are very similar, such as "performing exactly as instructed" and "following directions". Given this and their potential double interpretation, the study of the robot behavioral factors affecting users' trust can be difficult depending on the interaction confounders. As for the interactionist approach, the small time-framing makes the interactional history important during the analysis, making this approach **context-sensitive and non-generic**. Depending on which interactional process is being studied, and therefore the time-frame of analysis, the interpretation can yield different labels. Some behaviors can also be interpreted in both ways. For instance, after the robot fails to first understand the name of a participant, the participant may repeat itself. By doing so, the participant highlights the robot failure and disrupts the interaction fluidity. But, by repeating, they start an error reparation process and thus reveal that they still trust the robot to understand their name.

Their *time-framing*, the optimal time-interval necessary for the analysis, also differs. Our theoretical analysis showed that the interactionist approach time-framing is **close to one or a short series of behavioral units**, since it heavily relies on the interactional history. This approach is highly dependent on human unitizing and requires more training before being used. In our study, a post-annotation interview revealed that using unitizing that is not grounded on the interaction dynamics can lead to difficulties on the choice of the trust label. On the other hand, the mentalist approach has a much longer time-framing. Indeed, as trust assessment tools are questionnaires that are time-consuming to fill, measures are generally conducted at the end and beginning of the interaction. Their time-framing is thus generally **the entire**

interaction so that all criteria have enough time to evolve, and **sometimes several interactions** depending on the criterion. This reduces the possibilities to investigate the evolution of trust within the interaction. The focus is on the participants' representations of the robot *global* capacities and not the participants' behaviors.

Last, approaches diverge on their *scalability*, the ability to be used for the analysis of a single user or larger groups. The theoretical analysis showed that the interactionist approach is very **scalable** as the methodology of analysis does not have to change when going from a dyadic to a multiparty interaction. The analysis is driven by the interaction activity, and takes into account its history [44]. Previous studies show that participants in groups organize the interaction in a manner that favors one-on-one exchanges, and that conversational rules between more than 2 participants are adaptations of one-on-one ones [38, 100]. However, trust psychological models are **hardly scalable**. Indeed, when users form a group to perform a joint activity, trust is considered as an emergent state of the group, and group trust assessments are more than the average of each user's trust. This means that the psychological model should change drastically since social phenomena happening during dyadic and group interactions are very different [67, 73, 90]. For instance, some of the RTS items - such as "provides appropriate information" and "communicates well" - would need to be re-specified for situations of asymmetry during group interactions - e.g. the robot communicates properly with only one participant but not the others.

Given all the previous criteria, we provide a few guidelines on the type of computational studies for trust analysis each specific approach can tackle. The interactionist approach is a good fit for a continuous participants' behavioral analysis throughout the interaction given its time-framing. This approach can be useful in contexts such as assistive robotics for elder care where a robot needs to adapt to different interaction modalities according to the user. It is also suited to investigate the impact of the robot behavior on the user's response, although in a very narrow time-frame and specific interactional context, such as the user's reaction to the robot pre-opening [98].

Given its current tools, the mentalist approach is not a good fit for real-time analysis of trust in HRI. With the important adaptations of the RTS in our study, our study demonstrated the need to design a more suited analysis tool with this approach. It is best suited to study the influence of the robot design or behavior on the user's decision to trust the robot based on an overall representation of a specific or multiple criteria relating to trust through statistical analysis. To ensure that the user's representation of mentalist models' criteria have enough time to evolve, assessments should be conducted at the beginning and end of the interaction, or at sufficiently long interval during the interaction.

3.4 Conclusion

We introduced a new methodology to analyze trust regularly throughout the interaction based on Interactionist Sociology theories. This methodology relies on the observable character

of trust, that users display through specific behaviors when interacting with a robot. Through this methodology, we built the TURIN coding scheme that allows to annotate segments that are homogeneous in terms of trust content, and describe the behaviors that are linked to trust through TURIN different sub-categories. All these sub-categories provide behavioral descriptors of trust that the observer should look out for during the annotation process.

We showed that our methodology and coding scheme can lead to different conclusions from the standard Psychological approach to trust in HRI as they differ on four main criteria: orientation, generalization capability, time-framing, and scalability. Even though Psychological trust questionnaires can indicate that users have a poor global trust towards the robot, our approach can still reveal moments where their behavior is indicative of trust.

Chapter 4

Trust analysis throughout the interaction

Abstract

Using the methodology we previously described, we build multimodal computational models to analyse trust regularly during the interaction. We first build models using traditional machine learning techniques, and use these models to learn to predict the label associated with the segment currently being processed. We propose to study these models using two fusion mechanisms of our multimodal features: early and late. We also propose a set of features that describe different modalities: body, face, voice, and semantics. These features combine a mix of both automatically and manually extracted ones. Then, we study the prediction of the label of a segment by taking into account its history, which form a sequence all together. For this task, we propose a neuronal architecture based on two modules: i) the Within-Group Dynamics Encoder (WGDE) module encodes user data at different levels (individual, dyads, triad), ii) the Interactional-Gated-Recurrent-Unit (IG) module treats robot and user group data as a dialogue to model the temporal structure of the interaction.

Associated publications:

- *Hulcelle, M., Varni, G., Rollet, N., Clavel, C. (2023). "Computational Multimodal Models of Users' Interactional Trust in Multiparty Human-Robot Interaction". In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13643. Springer, Cham.* https://doi.org/10.1007/978-3-031-37660-3_16
- *M. Hulcelle, L. Hemamou, G. Varni, N. Rollet and C. Clavel, "Leveraging Interactional Sociology for Trust Analysis in Multiparty Human-Robot Interaction," International Conference on Human-Agent Interaction (HAI '23), 2023, Gothenburg, Sweden,* https://doi.org/10.1145/3623809.3623973

Building on the theoretical background that we explained in the previous chapter, we here explore multimodal computational models to predict trust regularly throughout the interaction. We first start by presenting our choice of multimodal features for the computational models. We then present a simple computational model of trust based on simple machine learning techniques, followed by a presentation of our recurrent neuronal architecture that models trust dynamics.

4.1 Multimodal features for automatic trust analysis

To explore the impact of within-group dynamics on the model performance and model those at the feature level, we extracted features from: (i) each group member (human/robot); (ii) dyads (human-human as well as human-robot); and (iii) the group as a whole (triad). Since our objective is not on online detection of trust, we chose a set of features that is a mix of both automatically and manually extracted ones from four modalities: face, body, voice, and semantics. We detail here the choice of our features.

4.1.1 Face

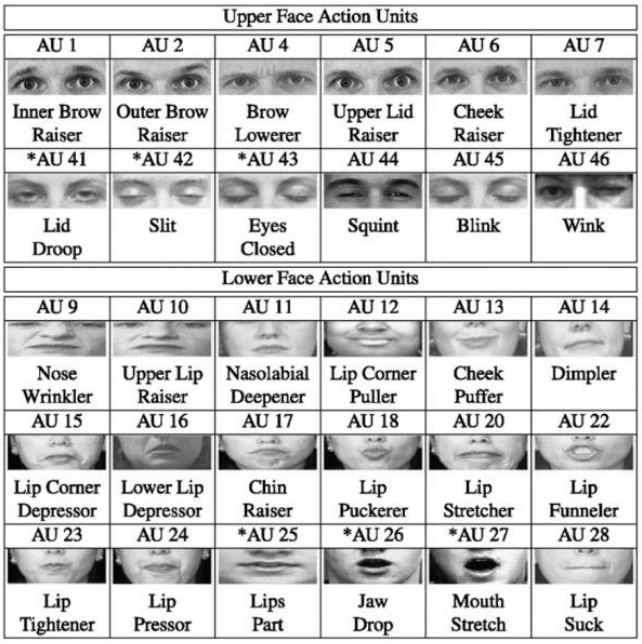

Figure 4.1: Visualization of Facial Action Units (FAU). All FAUs presented here are considered "activated" as opposed to a neutral facial expression where the face muscles are at rest.

We extracted Facial Action Units (FAU) using OpenFace [10] from the front camera view. A list of all FAU that OpenFace extracts can be found in Figure 4.1. FAU are indicators of users' facial expressions, which convey alignment and affiliative information towards the interactional partner. These can indicate whether users notice the robot failures - e.g. by raising their brows,

or by frowning [3, 74] - or smile to signal warmth towards the robot. We selected the intensity values of Action Units (AU) 1, 2, 4, 5, 6, 7, 9, 10, 12, 14, 15, 20, and 23 for each user. We excluded the FAUs that were not activated at least once during the interactions. We excluded FAUs relating to the eyes as activation values were too noisy considering the video angles provided by the dataset. FAUs were then filtered using a Savitsky-Golay (SG) filter (window length=11, polyorder=3) to reduce noise. This window length corresponds to a smoothing of values over a time of 0.5 seconds, which is a frequently used value in affective computing. FAU were extracted on all three front camera views. We mainly used the values that were extracted from the robot point-of-view. Whenever OpenFace lost tracking of one of the user's face on this camera view, we switched to another camera view for this user. Missing data was interpolated whenever OpenFace could not track one of the user's face.

The visual focus of attention (VFOA) can be considered a sign of trust when shared. Indeed, it is evidence that users are following the robot presentation, and trust the robot to point at the proper painting that is being currently referred to. This is proof of users' alignment and credit given to the robot, which are elements of our working definition of trust. The Vernissage dataset provides VFOA annotations for only one interaction. We thus manually annotated VFOA for both users independently following the labels originally suggested by authors of Vernissage : *left painting*, *central painting*, *right painting*, *Nao*, *other human*, *other*, *unclear*. VFOA were represented trough a binary indicator of presence or absence of each label during the segment for each user. Then, we also computed the time percentage of in-group look for the user-user dyad, as well as the number of VFOA changes per user.

Nods can also be indicators of affiliative and aligning behaviors [8, 110]. Nods are generally produced as a sign of passive feedback called back-channelling [127] during which the listener conveys information to the speaker without taking its turn away. In this way, nods can convey feedback on different levels, such as attention, hearing, understanding, or acceptance of the speaker's discourse. In this way, nods can signify trust depending on the context. Nods have previously been used as a feature in [58]. Nods were manually annotated for each user by indicating the beginning and end of a sequence of nods. We then computed the time percentage of nods during a segment per user and used this as a feature.

4.1.2 Body

In order to have low-level descriptors of body movement, we tried extracting key body points through OpenPose [17]. Body posture, as well as body position, head position, and head rotation are features that are often found in the social computing litterature [6, 23, 66, 101]. From these low-level descriptors, higher level descriptors can be automatically computed such as deictic gestures, which can be relevant information for an automatic trust analysis, such as Lee J. et al. did for their computational model of trust [58]. However, the resulting extraction proved to be too noisy on the front camera views, and the tracking of the body was too regularly faulty with the rear camera view to be exploitable. We therefore looked for other possible features for this modality.

The Vernissage dataset provides 3D head poses / rotations of both users recorded through a Vicon motion-capture system at a 100 Hz. We applied a SG filter (window length=21, polyorder=3) to reduce noise on the raw signal. We then computed the barycenter of the triangle shaped by the two users and Nao, and kept the 2D point projected on the floor plane as the single triadic feature. Since the robot does not change position, this gives us the distance between users, and the distance between both users and the robot. We included this feature in our studies as distance to the robot has been shown to be correlated to trust [76].

With OpenCV, we computed the contraction index [16] of each user by extracting their bounding box and their silhouette based on the rear camera view. The contraction index is defined as the ratio between the area of a body silhouette and its bounding box. It is an indication of whether a user exhibits a very open stance or is closed on itself. An example can be found in Figure 4.2. Extracted features were cleaned through a SG filter (window length=11, polyorder=3). Our hypothesis here is that users that users who exhibit more closed stances - e.g. by crossing their arms - and keep their stance closed through the interaction are less trusting than users who have more open stances - e.g. by pointing at things. Hence, a lower contraction index value can symbolize mistrust. For instance, users crossing their arms along with scowling brows can be a sign of confusion, frustration or aggressiveness [72].

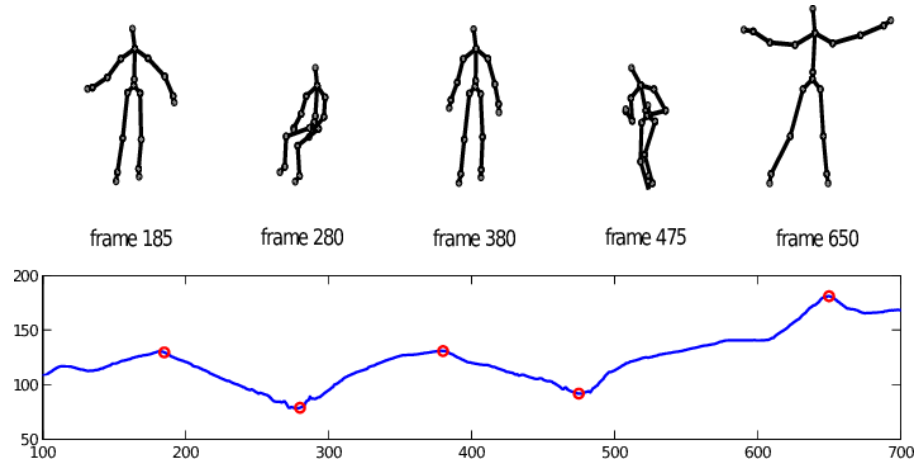

Figure 4.2: Variation of a performer's contraction index [117]. On the upper part, we can see the skeletal joints of the performer. We see that in frame 280, the performer is sitting with its joints closer to each other on the frame leading to a lower contraction index value, while the performer exhibits a very open stance by standing with its arms wide open in frame 650 resulting in a high contraction index value.

4.1.3 Voice

The dataset provides annotations for the vocal activity of both users and the robot. Users' vocal activity is annotated semi-automatically with the labels "speech", "silence", and "laughter". The dataset authors automatically annotated the first two labels, and manually annotated the last one. Nao's annotations are limited to the first two labels. We represented vocal activity by three binary indicators for each user and one for the robot and used these as features. The

dataset contains some overlaps between speakers (whether human or robot). Speech overlaps can either signify high trust during quick turn taking, or mistrust when users speak over the robot for a long period of time depending on their participation status [31, 35, 36]. We then compute the time-percentage of speech overlap in each segment between each user and the robot and store it as dyadic vector components.

Since previous research showed the importance of some prosody features to detect trust [49], we also computed prosody-related features from each group member's audio recording (users and robot). We extracted GeMAPS features [29] using OpenSMILE [28]. We kept the F0 (normalized for each speaker), loudness, jitter (cycle-to-cycle variations of fundamental frequency), shimmer (cycle-to-cycle variations of amplitude), spectral flux (frame-to-frame difference of the spectra) features as well as the first four mfccs, and the derivative of the F0 and first four Mel-frequency cepstral coefficients. We reduced the noise of the extracted features using a SG filter (window length=21, polyorder=3).

4.1.4 Semantics

We extracted a semantic representation of what the robot says during the interaction. The corpus provides transcripts of the robot speech with timestamps and lengths of pauses. The robot speech follows a very precise script composed of 90 sentences. The sentences have very little variations to include the names of users. This allows us to build a semantic context to which users react. We extract semantic representation through a TinyBERT [48] which yields a vector of dimension 312. Given the small amount of data we have, we decided to further reduce the size of the representation by conducting a principal component analysis (PCA). The final dimension of the representation is 50, which retains 99% of the explainable variance. We build an aggregated representation by averaging all words spoken during a segment for the robot alone.

Since the robot is silent in almost 20%($\pm$5%) of segments per interaction, we decided to propagate semantics and prosody into each silent segment from the one right before. This way, the model keeps track of the semantic and nonverbal context to which users react.

Features are aggregated in an early-fusion scheme for this study. We aggregated non categorical features by computing their mean and standard deviation within a segment. Figure 4.3 sums up how the different modalities are aggregated in a segment. This results in a feature vector of length 222 - 68 for each user, 79 for the robot, 3 for the dyads, and 4 for the triad.

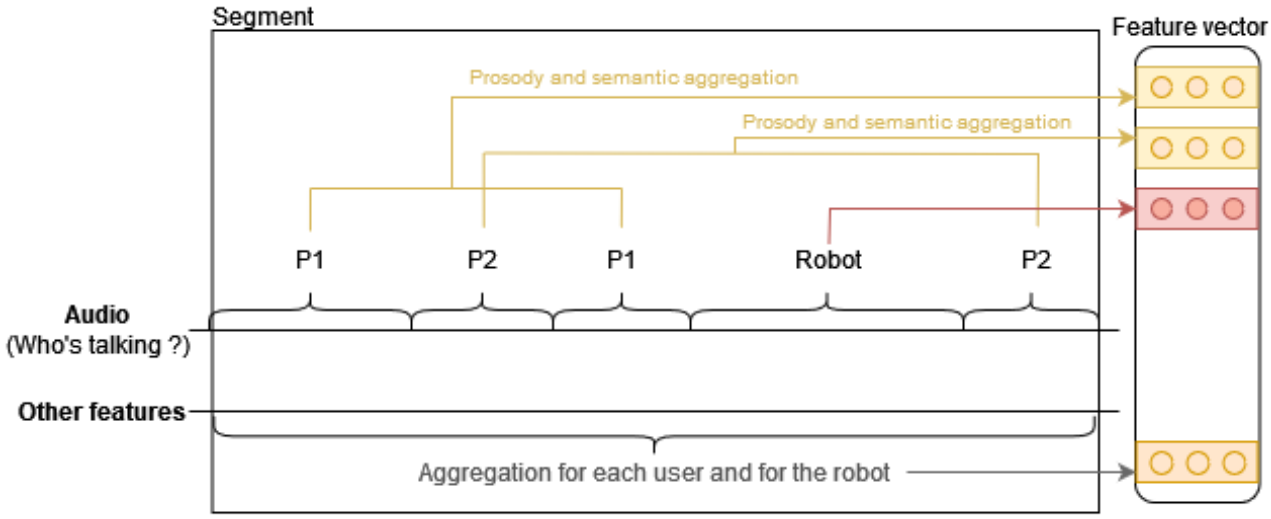

Figure 4.3: Audio features are aggregated within the segment only in sections where a user is speaking, while other eatures are aggregated on the entire segment.

4.2 Multimodal computational models for trust analysis

4.2.1 Simple machine learning models

4.2.1.1 Formalization

Formally, each interaction i can be represented by a sequence of segments obtained from the annotation process:

$Interaction_i = [(x^i_0, y^i_0), (x^i_1, y^i_1), ..., (x^i_{n_i}, y^i_{n_i})]$ where x^i_j is the feature vector obtained by aggregating frame-level features on segment j of interaction i, y^i_j the label of the segment, and n_i is the length of interaction i, i.e., the number of segments in the interaction.

For this first approach, we train simple machine learning classifiers on predicting y^i_j given the feature vector of the corresponding segment x^i_j. The hypothesis here is that interaction history is not necessary during training and prediction. Classes correspond to each of the labels from the TURIN coding scheme. Each of these labels correspond to the emergent trust of the group as explained in previous chapter. We formulate the trust classification problem in two different ways: (i) a One-vs-Rest (OVR) classification task; and (ii) a 3-class classification task. The OVR task formulation allows us to first study the separability of classes and search for explainability in the models' predictions.

Modalities are aggregated in 2 different ways - early-fusion and late-fusion - to study the role of each modality and their co-dependence. Let M be the number of modalities that constitute the feature vector. $x^i_{j,m}$ corresponds to the feature data for the modality m. In this regard, x^i_j corresponds to the concatenation of $x^i_{j,m}$ for all $m \in [1, M]$. In the early-fusion method, we concatenate all modalities to form a single feature vector x^i_j that we then feed to the classifier during training and prediction. In the late-fusion method, we train a different instance of the

model for each modality m. The outputs of each modality model $y^i_{j,m}$ are concatenated and submitted to a majority vote to determine the final output y^i_j.

4.2.1.2 Machine learning models

We train several machine learning models - Ridge classifier (RC), Random Forest (RF), Support Vector Machine Classifier (SVM-C), and Multi-Layer Perceptron (MLP) - for both classification tasks, and in early and late-fusion settings. As SVM does not natively support multi-class classification, we trained it as an OVR too. To achieve this, three models are trained, one for each class in an OVR setting. The output class is equal to the class left alone that has the highest probability between the three models.

We based our algorithm selection on an explainability criterion. The SVM-C learning is based on kernel trick which transforms the data to find an optimal boundary to separate classes. The kernel function choice can generate insight on the input data and the relation between features. The RF algorithmic design naturally creates interpretability. The matrix of weights from the Ridge classifier can also generate insights on the linear dependencies between the inputs and the output. The only algorithm that yields poor explainability is the MLP. MLP are known to be black-box algorithms, but we can still gather information on the non-linear dependecies between the inputs and output given the number of hidden-layers and their size selected by the hyper-parameter selection process. Figure 4.4 shows the entire pipeline of our model design.

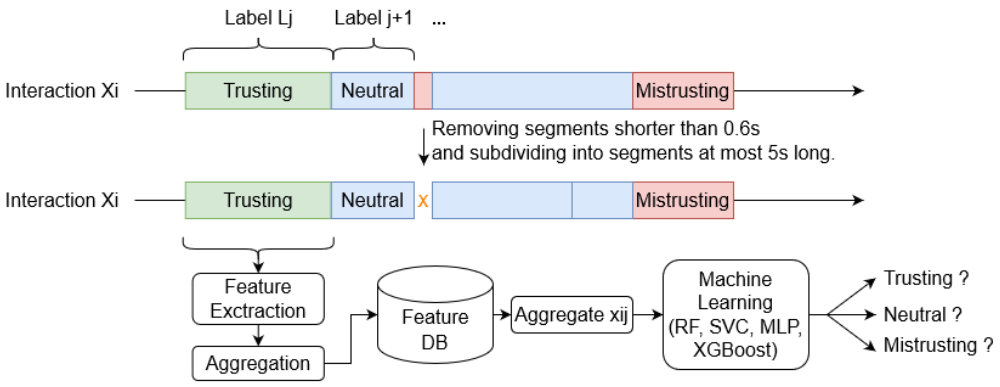

Figure 4.4: Summary of our model design. We first remove segments that are shorter than 0.6s seconds as they are poor in terms of behavioral content. We then extract features on each segment, which are aggregated by computing their mean value and standard deviation. We train each machine learning model on predicting the label associated to each segment.

4.2.2 A recurrent neural architecture

In this section, we present our architecture[1] through an incremental approach. At each step, we formulate a new hypothesis that we used for the design of a specific module.

[1]Code is available here: `https://github.com/GrituX/WGDE_IG`

4.2.2.1 Formalization

We build a sequence $(x^i_{j-\tau}, ..., x^i_j)$ comprised of the segment x^i_j, whose label y^i_j is the target of the model training, and the τ previous segments that constitute the *history*. Our objective here is to assess whether providing the history of the interaction improves the classifier performances. At first, we focus on the architecture to model the interaction and start with $\tau = 4$. After validating our architecture, we study the impact on the classification score of the history length for $\tau \in [1, 8]$. Figure 4.5 sums up the training pipeline of our architecture.

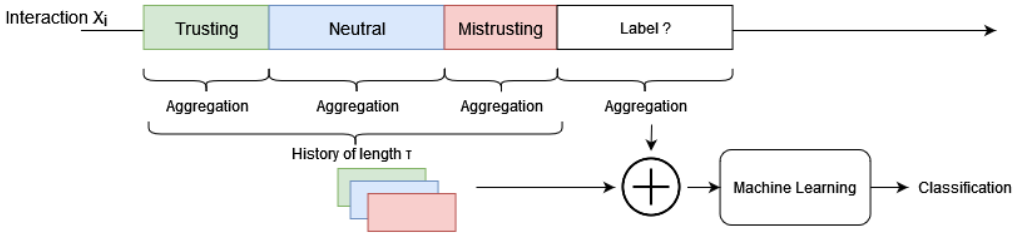

Figure 4.5: Features are aggregated for each segment. τ consecutive segments are grouped to form a sequence. Then we feed the sequence to a machine learning model to train it to predict the last label of the sequence.

4.2.2.2 Gated-Recurrent Units

We decided to build our architecture based on Gated-Recurrent Units (GRU) to encode information from participants' behaviors. A GRU is able to encode sequences. It uses two mechanisms to tackle the vanishing gradient issue, the first being a reset gate which controls the amount of necessary information from the past. The second one is an update gate, which controls the amount of past information to keep and the amount of new information to add. Formally, we write h_t the hidden state of the GRU for the encoded sequence at timestep t.

4.2.2.3 A first sequential approach

Users' actions are relevant within the sequence of previously exhibited behaviors by members of the interaction, and produced in response to somebody else's speaking turn [37, 38]. In that sense, someone's behavior should be understood and analyzed through the sequence of its and the interaction partners' past behaviors. Thus we can make the assumption that the interactional context is needed when classifying a segment. We train a model that we call **Simple-GRU (SG)** that contains a GRU layer followed by a fully-connected one, and then a final softmax layer. Formally, we have:

$$\widetilde{y^i_j} = softmax\left(W_h h^i_j + b_h\right) \tag{4.1}$$

where h^i_j corresponds to the hidden state of the segment j that we classify, which is the last state of the input sequence of the RNN, W_h is a weight matrix, and b_h is a bias vector. Figure 4.6 shows the architecture of this model.

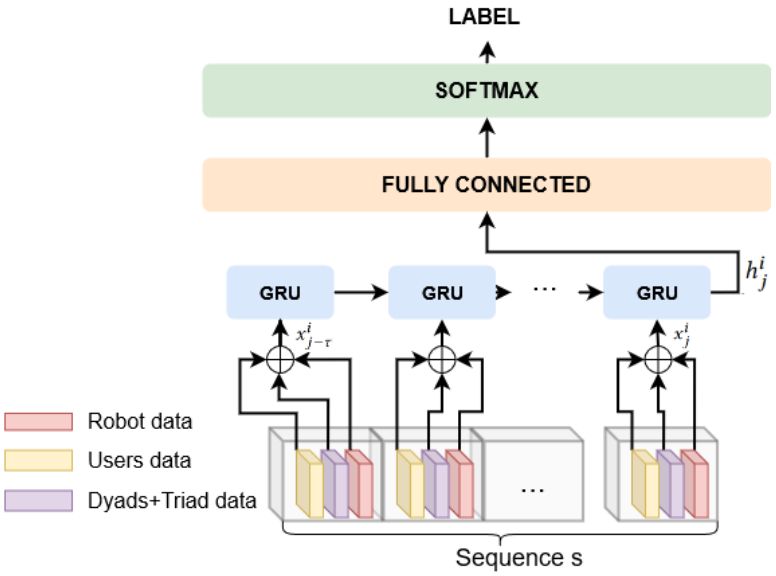

Figure 4.6: Representation of the Simple-GRU architecture. Robot, users, dyads, and triad data are concatenated and directly fed to a GRU before classification.

4.2.2.4 Modeling within-group dynamics

When engaging in a group activity involving a conversation, participants organize their interactions with other group members in specific ways depending on the activity and its goals. Participants can be speakers and thus address the entire group, a part of the group, or a single individual. They can also be listeners by being either actively engaged and showing signs of interest, or being passively engaged. It is thus necessary to analyze the interaction between all users from the entire group to individual to fully understand the group dynamics [36]. We conclude that trust should be analyzed by the interactions between all members of the group, at different scales (individual, dyads, and triads).

We thus modeled inter-segment within-group dynamics by adding a **Within-Group Dynamics Encoder (WGDE)** module to our architecture (see Figure 4.7). We call this new architecture comprised of the WGDE module and a simple GRU the WGDE-SG. The WGDE splits data within a segment. Feature vectors include data from the humans and the robot. First, the module takes input features from each user independently to feed them in its own gated-recurrent unit (GRU), as opposed to the previous architecture shown in Figure 4.6 where users and robot data were directly concatenated. Second, the outputs of the two GRUs are concatenated with features from all three dyads and the triad. Dyads are formed by either the two users, or each user with the robot. The triad corresponds to the entire group.

The resulting vector is then being fed to a fully connected (FC) layer which is the final output of our module $x_{g,t}^i$ at time-step t. We use the index g to refer to data representing the group formed by users, and the index r to refer to data representing the robot. Here, we have $t \in [j-\tau, j]$ with $j \in [\tau, n_i]$ to form a sequence of length τ.

We then concatenate the output of the WGDE with the robot features to produce the output

at timestep t :

$$h_t^i = GRU(x_{g,t}^i \oplus x_{r,t}^i) \tag{4.2}$$

where $\oplus$ denotes the concatenation operator. To assess the contribution of the robot behavior, we compare two versions of this architecture: one with and one without any robot data, meaning that only the output of the WGDE is taken as input for the simple-GRU RNN.

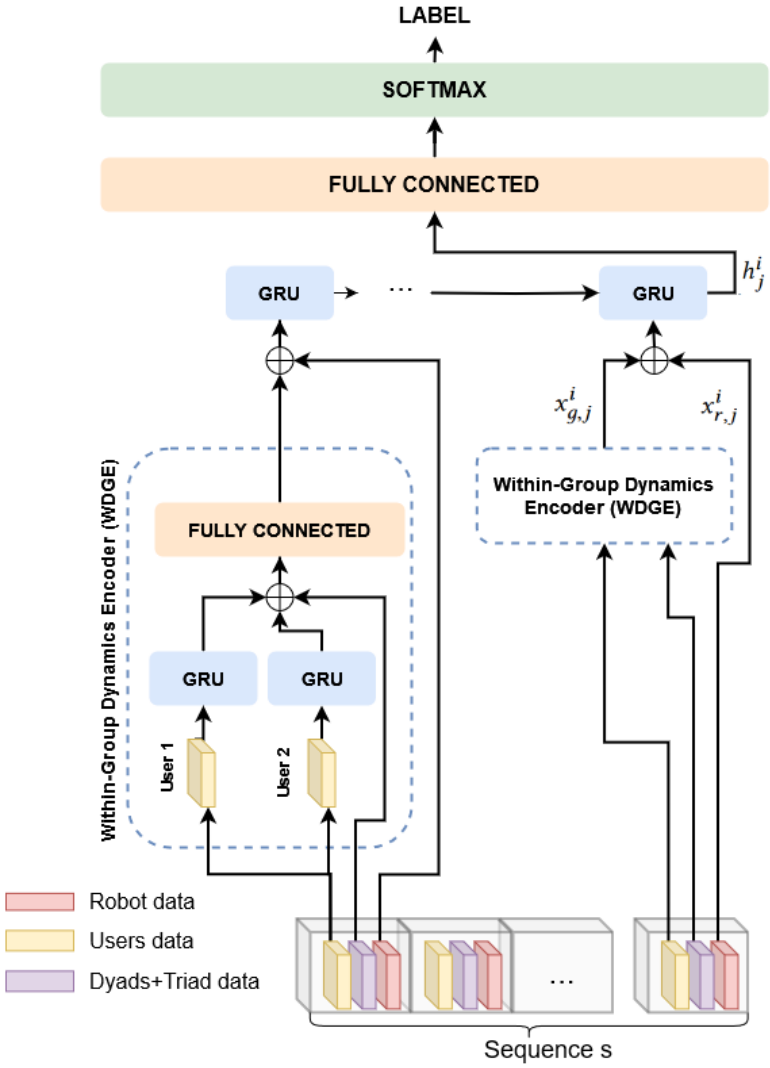

Figure 4.7: Representation of the WGDE-SG architecture. Users, dyads, and triad data are fed into the WGDE module. Its output is then concatenated with robot data from the segment and then fed to a GRU.

4.2.2.5 Modeling the temporal structure of the users-robot interaction

During an interaction, participants continuously produce social behaviors. These behaviors carry a meaning that form the interactional context, be it at a low level such as lexical, semantic or at a higher level such as shared knowledge. Other participants build their answer from their

interactional partners' past behaviors, and hence renew the context at each speaking turn [38]. Participants using the previous speaking turn of another group member as a contextual resource to create their own speaking turn means that there is a temporal structure between users and the robot.

We modeled this temporal structure by adding an Interactional GRU (IG) module. At each time-step t, features from the robot alone $x^i_{r,t}$ are fed into a GRU whose output is concatenated with the output of block WGDE $x^i_{g,t}$. This data is then fed to a GRU:

$$h^i_{r,t} = GRU(x^i_{r,t} \oplus h^i_{g,t-1}) \tag{4.3}$$

$$h^i_{g,t} = GRU(x^i_{g,t} \oplus h^i_{r,t}) \tag{4.4}$$

$h^i_{g,t}$ refers to the encoded hidden state of the group's behavior at timestep t produced in response to the robot encoded behavior $h^i_{r,t}$. In this model, the robot hidden state is computed with the group's hidden state from previous timestep. We modeled the interaction this way to emphasize the fact that the robot is the leader of the interaction within a segment, as the group's behaviors are modeled to be the answer of the robot within the segment. This choice is due to the role asymmetry between users and the robot.

The entire architecture is shown in Figure 4.8. We name this architecture WGDE-IG.

4.3 Conclusion

We proposed two offline methods to predict trust during a segment and explained their underlying hypotheses. The first method relies on traditional machine learning techniques. This model considers that no context is needed to predict trust, and thus takes as input the segment on its own. It handles multimodality through an early-fusion or a late-fusion mechanism to study the interplay between modalities and assess their importance.

The second method considers that the interaction's history is needed for the prediction, and therefore takes into account a few previous segments. From this, we designed a neuronal architecture that is based on two modules: i) the WGDE module encodes user data at different levels (individual, dyads, triad), ii) the IG module treats robot and user group data as a dialogue to model the temporal structure of the interaction.

We also presented a set of features that is a mix of automatically and manually extracted, and based on different modalities: body, face, semantics, and voice. Features were chosen in a way that is non-invasive for users so that future online models can rely on these without intruding on the users' feeling of "naturality" of the interaction.

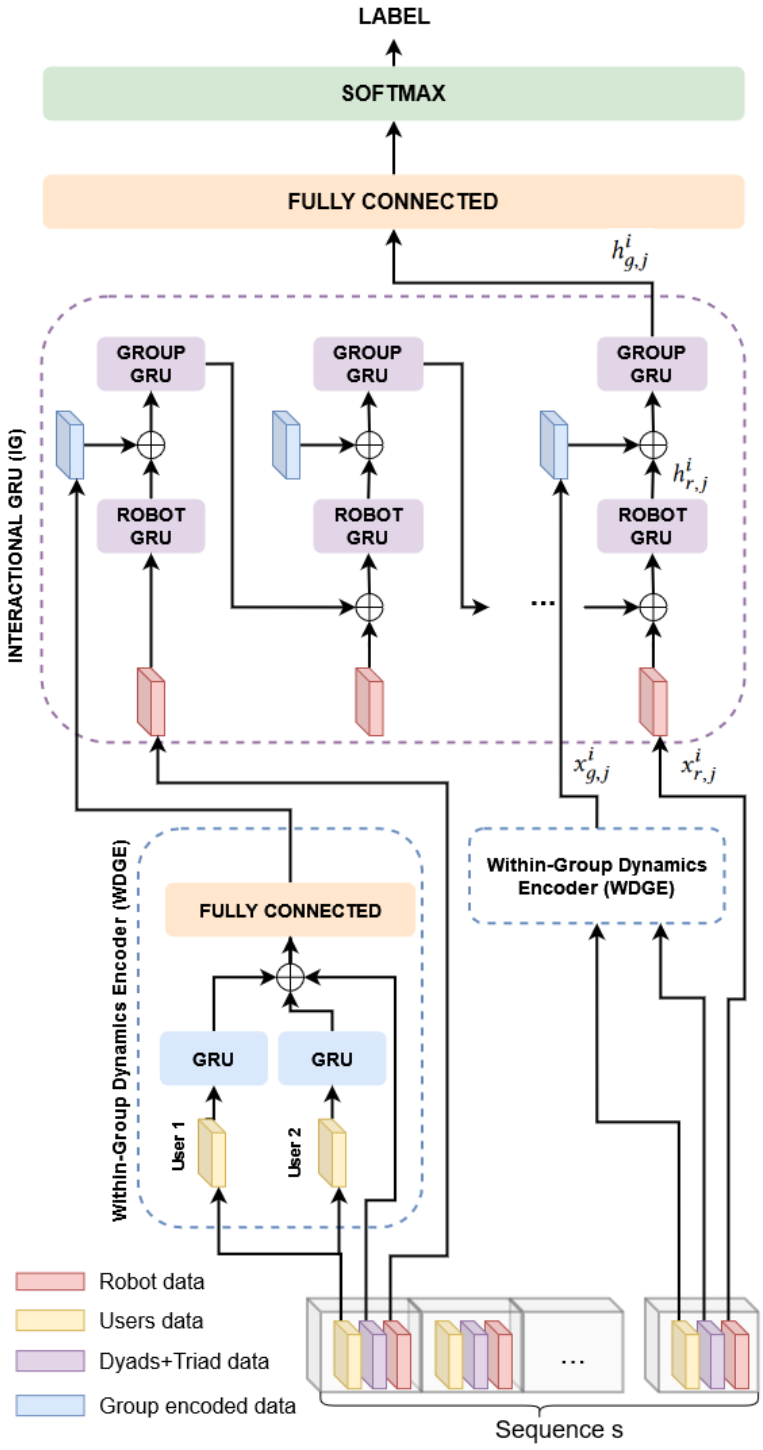

Figure 4.8: Representation of the WGDE-IG architecture. Robot data is concatenated with previous segment group GRU, and then fed to the robot GRU. Its output is concatenated with the output of the WGDE module, before being fed to the group's GRU.

Chapter 5

Experiments and Analysis

Abstract

We used the Vernissage dataset that we described in Chapter IV as an experimental testbed for the models that we designed. First, we trained the traditional machine learning models on two tasks : a binary classification one through a One-Vs-Rest (OVR) mechanism, and multiclass classification with our three labels as our ground truth. Through a study of the impact of two fusion mechanisms, early and late-fusion, we show that Random Forests perform the best in early-fusion when classifying a segment with no history of interaction. By studying the results of the RF with the late-fusion mechanism, we observe that the voice modalities play a more important role for classification, followed by face, and finally body modalities.
Then, we train different versions of our neuronal architecture described in Chapter IV that takes as input a sequence formed of the target segment and the previous τ segments that constitute the history. Through an incremental approach, we show that using our neuronal architecture with all of its modules leads to increased performance. While the "Neutral" and "Trusting" classes have fair scores, the model scores indicate that further work is needed to properly capture the dynamics to predict the "Mistrusting" class. The optimal history length τ remains also unclear. By studying the errors made by this final model, we provide guidinelines on a new set of features that could be used to improve trust models.

Associated publications:

- *Hulcelle, M., Varni, G., Rollet, N., Clavel, C. (2023). "Computational Multimodal Models of Users' Interactional Trust in Multiparty Human-Robot Interaction". In: Rousseau, JJ., Kapralos, B. (eds) Pattern Recognition, Computer Vision, and Image Processing. ICPR 2022 International Workshops and Challenges. ICPR 2022. Lecture Notes in Computer Science, vol 13643. Springer, Cham.* `https://doi.org/10.1007/978-3-031-37660-3_16`
- *M. Hulcelle, L. Hemamou, G. Varni, N. Rollet and C. Clavel, "Leveraging Interactional*

Label	Count	Segment length (Avg $\pm std$)
Trusting	193 / 240	3.7s ±4.9s / 3.0s ±1.5s
Neutral	260 / 604	9.5s ±8.8s / 4.1s ±1.4s
Mistrusting	75 / 78	2.8s ±2.5s / 2.6s ±1.6s

Table 5.1: Summary of annotations collected using the TURIN coding scheme.
Raw annotations / Annotations after our sub-segmentation.

Sociology for Trust Analysis in Multiparty Human-Robot Interaction," International Conference on Human-Agent Interaction (HAI '23), 2023, Gothenburg, Sweden, `https://doi.org/10.1145/3623809.3623973`

After presenting the models and the feature set that we designed, we here establish the experimental protocols to evaluate their performance. We then discuss their results to determine which features have more weight for classification, and determine which type of errors they make.

5.1 Training results

5.1.1 Dataset

We decided to divide the segments from the annotations collected from the Vernissage dataset into several sub-segments of at most five seconds to have a better homogeneity in segment length when aggregating features. When the subdivision occurs, the label is duplicated for all of its sub-segments. The length of 5s corresponds to the global average length of segments before segmentation. Segments smaller than 600ms are dropped out. Such short segments are, in this case, poor in interaction content because of some modalities analysis window - e.g. semantics, verbal. Table 5.1.1 provides the counts and lengths of segments before and after our sub-segmentation. Then, features are aggregated on the resulting segments.

5.1.2 Without context

We used a leave-3 groups-out (LTGO) cross validation to tune the models hyperparameters as well as to evaluate their performances. This enables a reduction of the variance of the performances providing a better overview of the generalization of the models. Since we draw 3 interactions out of a total of 10, this results in 120 rounds of test. The ROC-AUC metric was chosen as it provides a broader view of a model performance given that it captures the trade-off between precision and recall. We also used the F1 score for the multi-class classification problem. All the models were developed and evaluated using Python's scikit-learn package [85]. As the classes are heavily imbalanced as shown in Table 5.1.1, We augmented data using SMOTE [20] to obtain a balanced dataset. Data augmentation was performed after the LTGO division. For this first model, we used all modalities except for the semantic modality to reduce the size of the feature vector given the small amount of data.

Interactions are very different from one another, aggregating the test in this way allows to reduce the variance of scores which gives us a better overview of the generalization capacity of learned models. Tables 5.1.2 and 5.1.2 summarize the results.

Early-fusion	**RF**	**MLP**	**SVM-C**
Trusting-vs-rest	**0.72** ±0.04	0.70 ±0.04	**0.74** ±0.04
Neutral-vs-rest	**0.77** ±0.04	0.74 ±0.04	0.75 ±0.04
Mistrusting-vs-rest	**0.59** ±0.06	0.54 ±0.07	**0.58** ±0.06
Late-fusion			
Trusting-vs-rest	**0.67** ±0.04	0.60 ±0.04	**0.66** ±0.04
Neutral-vs-rest	**0.74** ±0.04	0.65 ±0.04	0.70 ±0.03
Mistrusting-vs-rest	**0.54** ±0.08	0.48 ±0.08	0.49 ±0.10

Table 5.2: ROC-AUC test scores of OVR classifiers

Fusion	**Rand.**	**Maj.**	**RF**	**MLP**	**SVM-C**
Early	0.38 ±0.03	0.52 ±0.05	**0.66** ±0.04	**0.65** ±0.04	0.60 ±0.04
Late	0.38 ±0.03	0.52 ±0.05	**0.62** ±0.03	0.60 ±0.04	**0.61** ±0.05

Table 5.3: f1 test scores of multi-class classifiers. Rand.: Random Classifier. Maj.: Majority-voting Classifier

We conducted a series of statistical tests to compare the classifiers performance and check for possible statistical differences between them. We used a Kruskal-Wallis test (KW) [55] followed by a post-hoc Dunn test (when needed) [25] to compare all models in either early- or late-fusion, with Bonferroni correction. We conducted a Wilcoxon-Mann-Whitney test (WMW) [120] to compare the performance of a model between early- and late-fusion. For the OVR method, we compared models in a single binary classification task (e.g. trusting-vs-rest for early- against late-fusion, or trusting-vs-rest for RF against MLP against SVM-C). We conducted all of our tests using an alpha value of .05. Figures 5.1 and 5.2 sum up all the p-values of our tests.

OVR method: the KW tests point significant differences between the models in early-fusion for Mistrusting-vs-rest, $H = 40.76, p < .001$, Neutral-vs-rest, $H = 28.49, p < .001$, and for Trusting-vs-rest, $H = 44.61, p < .001$. In both Mistrusting-vs-rest and Trusting-vs-rest, the post-hoc tests indicate no significant differences between the RF and the SVM-C, respectively $p = 0.16$ and $p = 0.28$, while the difference is significant between the RF and the MLP in both tasks, $p < .001$ for both, and between the MLP and the SVM-C, $p < .001$ for both. There is no statistical difference between the MLP and the SVM-C in Neutral-vs-rest, $p = 0.065$. This difference is significant between the RF and the MLP, $p < .001$, and between the RF and the SVM-C, $p < .01$. In late-fusion, the KW tests reveal significant differences between the models for Trusting-vs-rest, $H = 126.36, p < .001$, Neutral-vs-rest, $H = 169.65, p < .001$, and for Mistrusting-vs-rest $H = 35.30, p < .001$. The post-hoc tests show that the RF performs the best for both Mistrusting-vs-rest and Neutral-vs-rest classifications, followed by the SVM-C, and then the MLP. The difference in scores between the SVM-C and the MLP is non significant for Mistrusting-vs-rest, $p = .90$. Regarding Trusting-vs-rest, the difference between the RF and the SVM-C is not statistically significant, $p = .24$, while it is between the RF and the MLP,

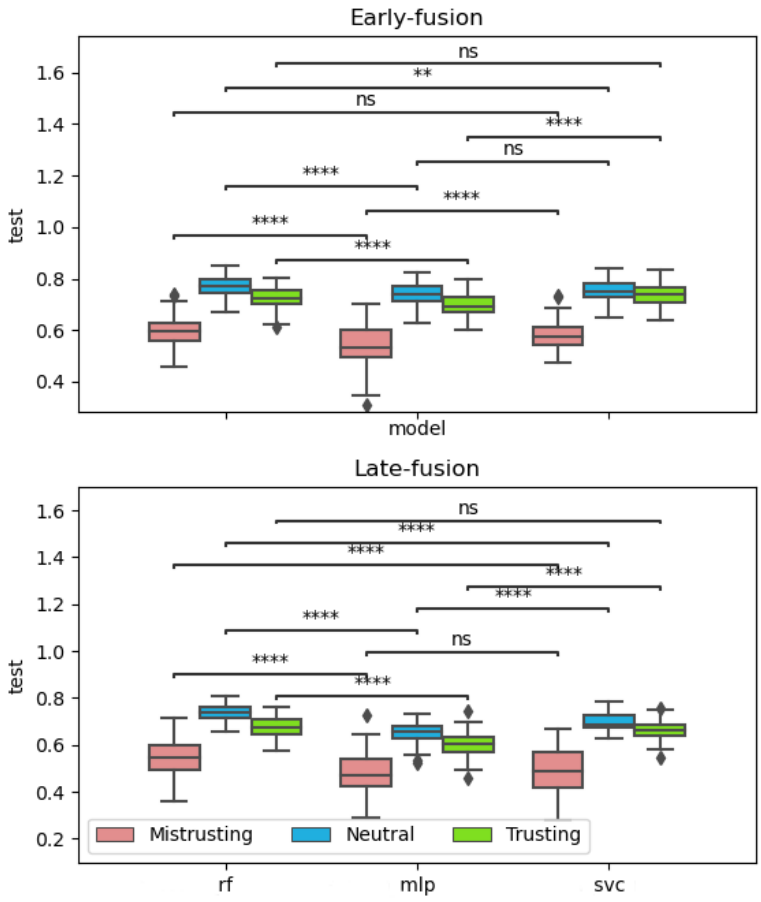

Figure 5.1: Boxplot of all models ROC-AUC test scores in OVR classification.
*** $p < .0001$; *** $p < .001$; ** $p < .01$; * $p < .05$; ns: Non Signficant

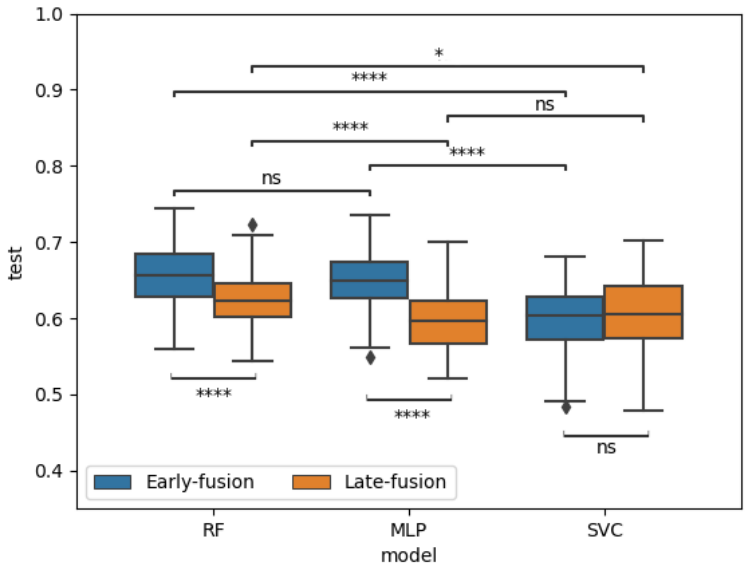

Figure 5.2: F1 test scores for multi-class classification.

and between the MLP and the SVM-C, $p < .001$ for both.

A series of WMW tests to compare between early and late-fusion reveals that early-fusion works better for all models in every OVR task, $p < .001$ for all models. Results of the tests are shown in Table 5.1.2.

	RF	**MLP**	**SVM-C**
Trusting-vs-rest	$U = 2784$	$U = 817$	$U = 1366$
Neutral-vs-rest	$U = 3772$	$U = 986$	$U = 2077$
Mistrusting-vs-rest	$U = 4332$	$U = 4183$	$U = 3101$

For all tests, we have $p < 0.0001$

Table 5.4: WMW test results to compare between early and late-fusion for all models in all OVR tasks.

Multi-class classification task: for early-fusion, the KW test showed significant difference between models, $H = 104.76, p < .001$. The post-hoc test indicated a difference between MLP and SVM-C, $p < .001$, as well as between RF and SVM-C, $p < .001$, but showed no significant difference between RF and MLP, $p = .54$. For late-fusion, the KW test revealed a difference between models, $H = 25.15, p < .001$. This difference is statistically significant between RF and MLP, $p < .001$, and RF and SVM-C, $p < .05$, but not between MLP and SVM-C, $p = .069$, according to the Dunn test. When comparing between early and late-fusion for each model, a series of WMW test showed that early-fusion yields better results for the RF, $U = 3715, p < .001$, the MLP, $U = 2359, p < .001$, but not for the SVM-C, $U = 6448, p = .081$.

The analysis shows that models perform better in an early-fusion setting. This means that early-fusion captures the interactions between modalities better than late-fusion does, albeit in a simple way. Considering that the RF is the best performing model in the OVR task, we analyzed how each modality model performs individually in late-fusion. The results for this training are given in Table 5.1.2. A series of KW tests revealed significant performance difference between modalities for Trusting-vs-rest, $H = 187.48, p < .001$, for Neutral-vs-rest, $H = 288.62, p < .001$, and for Mistrusting-vs-rest, $H = 150.93, p < .001$. For all OVR tasks, the post-hoc Dunn tests showed that the RF trained only on the voice modality performs better, followed by the RF trained on face modality, and finally the RF for the body, $p < .05$ for all.

	Body	**Face**	**Voice**
Trusting-vs-rest	0.54 ±0.05	0.62 ±0.04	**0.65** ±0.05
Neutral-vs-rest	0.57 ±0.03	0.65 ±0.03	**0.73** ±0.05
Mistrusting-vs-rest	0.46 ±0.08	0.51 ±0.06	**0.60** ±0.07

Table 5.5: ROC-AUC test scores of RF trained separately on the Body, Face, and Voice modalities in OVR classification.

5.1.3 With context

Parameters First, we specify the hyper-parameters of the sequential model and its training. We start by specifying the hyper-parameters of the first model consisting only of the SG. The

GRU is composed of a single-layer, mono-directional GRU cell. The model hyper-parameters are composed of: the GRU output dimension $h_j^i \in \{8, 16\}$, the number of epochs during which to train $\in \{30, 50, 70\}$. The model is trained with a batch size of 128 sequences, a learning rate of 1.10^{-3}. Given the small amount of data, we add a dropout layer before the final fully-connected layer, with $p = 0.5$. Hyper-parameters are selected using a grid-search cross-validation performed on a validation set of one interaction. We perform the training by leaving one group out for testing. We train the model using 5 different random seeds for each round of training. The model and training was implemented using the python package PyTorch [83]. We choose the F1 score and balanced accuracy metrics to understand which kind of error the model makes.

When training the WGDE-SG model, we add the constraint that the output dimension of the WGDE GRUs and the SG GRU is the same for all. We use the same ranges for the hyper-parameter selection, as previously detailed.

As for the WGDE-IG model, we keep the constraint that the output dimension of the WGDE GRUs and the IG GRUs is the same for all. The output dimension is the same for all GRUs. The value of output dimension is kept as a hyper-parameter with the same search values as previously. Again, for this model, we keep the same range for the training parameters, as previously detailed.

We trained the models to solve a multi-class classification task (labels "Mistrusting", "Neutral", "Trusting"). For this task, We chose the cross entropy as loss function computed between the output $\widetilde{y_j^i}$ and the target y_j^i. Here, we perform prediction only for the last segment of the sequence.

As the dataset we use is imbalanced - 4% of data is labelled Mistrusting, 67% is labelled Neutral, 29% is labelled Trusting -, we use weights corresponding to the inverse class proportions. As the resulting dataset is small, we augmented data by generating new samples by adding Gaussian noise of $\sigma = 2.10^{-3}$ for each sample. In total, we augmented the entire dataset four times. We then cleaned our dataset using Wilson's editing algorithm with $k = 3$ [121]. To further deal with this class imbalance, we use weighted random sampling during training at each epoch, again with weights corresponding to the inverse class proportions. Given the small amount of data, we added a regularization term to the loss function corresponding to the L2 norm of the weights with $\lambda = 1.10^{-2}$.

Results We present the training results along with statistical tests to determine which model yields the best performance.

We report the multi-class micro F1 scores on the test sets for each model for $\tau \in [2, 8]$ in Table 5.6. We perform a series of statistical test each time with $level = 0.05$. A series of Shapiro-Wilk test [107] for each combination of model and τ show that not all of the F1 scores follow a normal distribution. A Kruskal-Wallis test [55] between all models F1 scores reveal a difference between their performance, $H = 367.19, p < .001$. A post-hoc Dunn test [25] indicates that the difference is significant between the SG with no robot data and all other

τ	1	2	3	4	5	6	7	8
SG	0.733 ±.120	0.739 ±.119	0.733 ±.118	0.733 ±.127	0.735 ±.123	0.735 ±.125	0.734 ±.125	0.735 ±.124
SG (no robot)	0.621 ±.058	0.613 ±.081	0.621 ±.080	0.604 ±.095	0.597 ±.084	0.598 ±.092	0.603 ±.085	0.605 ±.087
WGDE-SG	0.726 ±.116	0.732 ±.119	0.723 ±.144	0.730 ±.141	0.724 ±.138	0.725 ±.146	0.723 ±.146	0.731 ±.148
IG†	0.730 ±.113	0.717 ±.105	0.695 ±.120	0.698 ±.163	0.694 ±.182	0.710 ±.145	0.689 ±.175	0.694 ±.188
WGDE-IG	0.730 ±.102	0.730 ±.098	0.715 ±.143	0.736 ±.124	0.735 ±.135	0.745 ±.110	0.730 ±.146	0.714 ±.137

Table 5.6: Mean and std of the micro F1 scores on the test sets of the models in the multi-class classification task for $\tau \in [1, 8]$.
†: the IG model corresponds to the WGDE-IG without the WGDE module.

τ	1	2	3	4	5	6	7	8
SG	0.565 ±.164	0.571 ±.158	0.575 ±.144	0.578 ±.150	0.578 ±.152	0.585 ±.147	0.577 ±.140	0.566 ±.147
WGDE-SG	0.591 ±.138	0.607 ±.134	0.596 ±.138	0.598 ±.133	0.590 ±.138	0.597 ±.137	0.596 ±.130	0.605 ±.144
IG	0.541 ±.149	0.536 ±.136	0.556 ±.123	0.547 ±.132	0.538 ±.128	0.552 ±.150	0.525 ±.136	0.537 ±.161
WGDE-IG	0.572 ±.141	0.580 ±.126	0.584 ±.137	0.585 ±.156	0.596 ±.144	0.592 ±.141	0.580 ±.170	0.560 ±.163

Table 5.7: Mean and std balanced accuracy on the test sets of the models in the multi-class classification task for $\tau \in [1, 8]$.

models, $p < .001$. This indicates the necessity to include robot data to better model the inter-segment dynamics of the group. This is confirmed by Interactionist Sociology theories which state that users express displays of trust through behaviors that are relevant with other group members' previous speaking turns. The addition of robot features in our models helps capturing these sequential dependencies. The post-hoc Dunn test also reveals that there is a statistical difference between the SG and the IG models performance, $p < .01$. The test reveals no further statistical difference between models, showing that the IG module alone fails to capture the interaction dynamics.

We report the balanced accuracy in Table 5.7 as it is a good metric for imbalanced datasets. A series of Shapiro-Wilk test [107] for each combination of model and τ show that not all of the average accuracy scores follow a normal distribution. A Kruskal-Wallis test reveals that there is no significant difference in the scores for the different sequence lengths. This highlights the complexity to define a single sequence length of analysis for all segments. A possible explanation is that some segments do not require a lot of context - e.g. users sometimes say "huh, what ?" when the robot interrupts them - while others require more context of the undergoing process to be understood. A Kruskal-Wallis test indicates significant differences between models, $H = 58.28, p < .001$. A post-hoc Dunn test shows significant difference for all models, $p < .01$, except between the SG and WGDE-IG, $p = 0.25$. A further look at each model median score indicates that the WGDE-IG yields the best performance, $median = 0.541$. While the WGDE-SG has higher mean scores, its quartiles are further apart than for the WGDE-IG, revealing that the latest is more consistent in its predictions.

Since we operated a few changes between this model and the previous one with traditional ML techniques, we trained a RF and MLP with the same experimental settings as the ones for the neuronal model to be able to compare the results with and without the context. We report the results in Table 5.1.3.

	RF	**MLP**
F1	0.744 ±.070	0.714 ±.068
Balanced accuracy	0.518 ±.118	0.507 ±.137

Table 5.8: F1 and balanced accuracy scores of a RF and MLP.

We compare these results with the scores from the WGDE-IG since it has the best performance, and use $\tau = 6$ as it leads to the highest mean F1 score and balanced accuracy. A first thing that we observe is that the introduction of the semantics modality leads to a significant increase in these models F1 score. We ran Kruskal-Wallis tests to check for statistically significant difference between the RF, the MLP, and the WGDE-IG scores. The tests indicated difference for both the F1 score, $H = 8.26, p < .05$, and the balanced accuracy, $H = 14.23, p < .001$. For the F1 score, a post-hoc Dunn test revealed significant difference only between the MLP and WGDE-IG scores, $p < .05$. A post-hoc Dunn test for the balanced accuracy scores indicated that the WGDE-IG had the best performance $p < .005$.

We also trained a version of the WGDE-IG that contains self-attention mechanisms such as

was done in [125]. However, the results were inconclusive as the performance were below the SG model, probably due to a low amount of data. Hence, we do not report the results here.

5.2 Feature importance

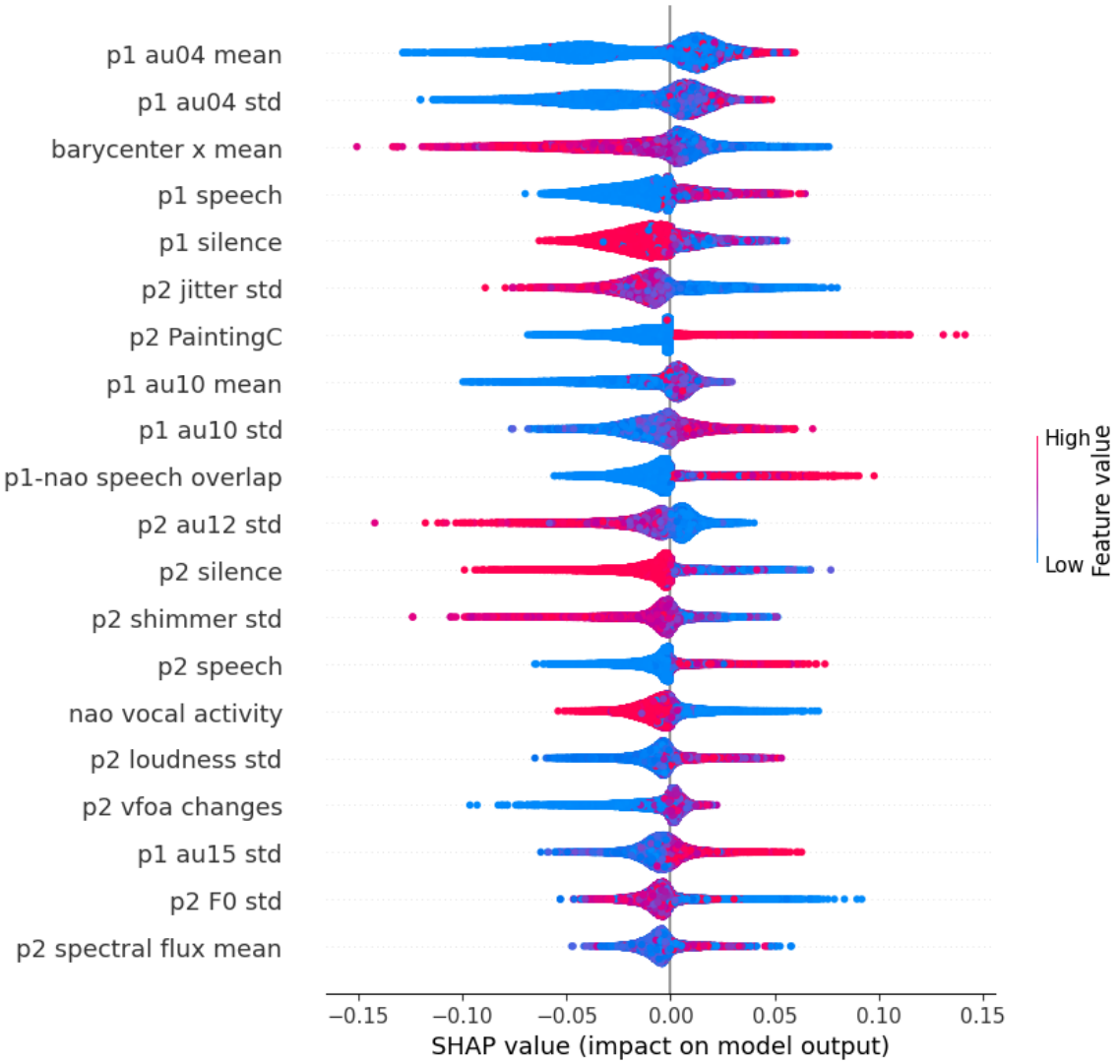

Figure 5.3: SHAP values point plot for RF for the "Mistrusting" class (p1/2: Participant 1/2)

We conducted a SHapley Additive exPlanation (SHAP) [62] values analysis for the random forest with early-fusion, since it gave the best results, to determine the importance of features in the classification task. SHAP values interpret the impact on the model output of a given feature having a certain value compared to the model prediction if that feature took some baseline value, e.g. its mean. When there are correlated features, only one of them appears in the SHAP values. Figures 5.3, 5.4, and 5.5 show the SHAP values associated to the features that were determined as more important by the analysis.

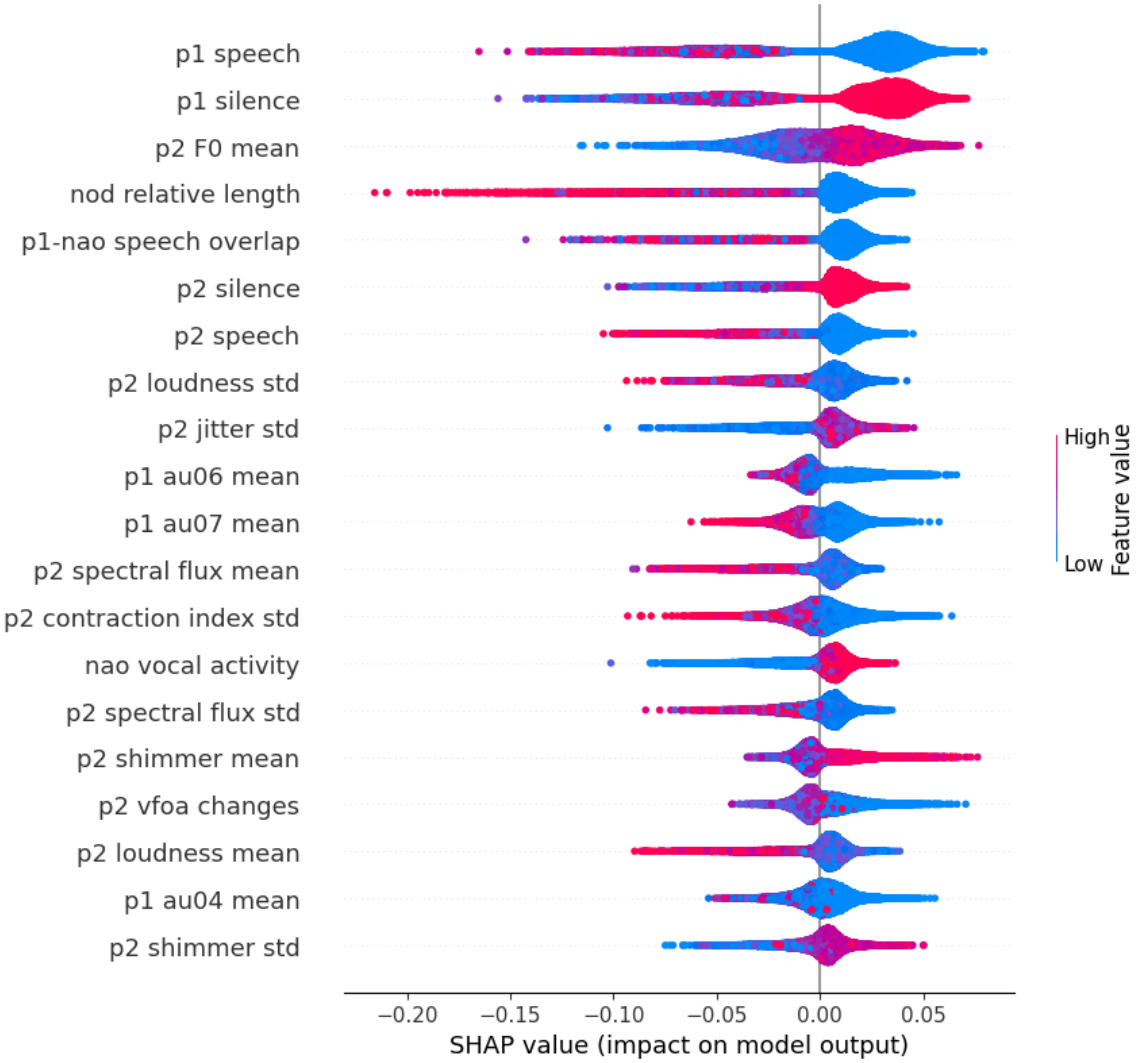

Figure 5.4: SHAP values point plot for RF for the "Neutral" class (p1/2: Participant 1/2)

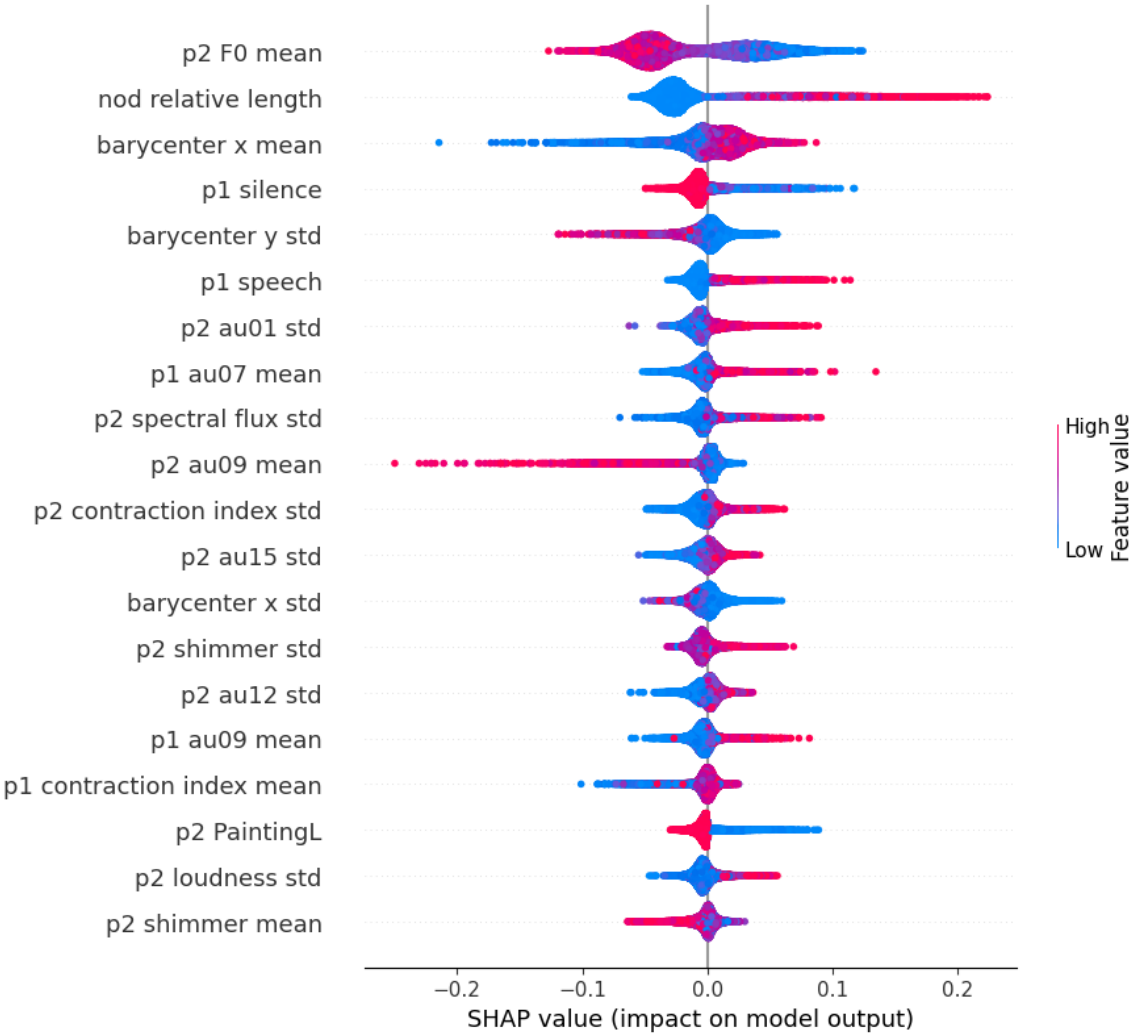

Figure 5.5: SHAP values point plot for RF for the "Trusting" class (p1/2: Participant 1/2)

5.2.1 Face modality

We present the analysis results by modality. About the face modality, participants lower their brows more (higher mean and std of AU4) during mistrusting segments than other segments, talk (higher mean and std of AU10), and change their VFOA more. During trusting segments, participants nod more often, tighten their lid more (higher mean of AU7), while nose wrinkles (higher mean of AU9) plays against classification as trusting. Higher std of the lip corner depressor (AU15) plays against classification as neutral.

5.2.2 Body modality

Concerning body, the distance of the group to the robot is the most important feature of trust. Greater mean distance to the robot has a positive impact when classifying as mistrusting, while closeness to it has more weight for trusting segments. This result confirms the findings in the literature [76]. When considering between-participants distance, a lower std value has a positive impact for trusting segments. As for the contraction index, its std plays a more important role than its mean value during a segment, with lower std being associated to neutral segments, while higher std are linked to trust.

5.2.3 Voice modality

Regarding the voice, neutral segments are linked to situations where participants are silent. Lower F0 means have a positive impact on classification as trusting, and higher F0 means have more weight for neutral segments. Higher speech overlap time between participants and Nao have a strong impact when classifying as mistrusting.

5.2.4 Global analysis

The SHAP values analysis shows that participants exhibit different behaviors depending on whether they trust the robot or not. Participants in trusting segments tend be be closer to the robot, more aligned with the robot presentation, move and talk more. In mistrusting segments, participants also talk, but they talk more over the robot. They also show signs of doubt through brows lowering, looking around, getting closer together, and being further away from the robot. As for neutral segments, participants are mostly silent and listen to the robot speaking, remain still, and have neutral facial expressions.

The SHAP analysis demonstrated the importance of the participants' distance to the robot. The barycenter feature was shown to be important, while the contraction index feature had a moderate impact on the model output. However, the series of statistical tests conducted in previous section on the late-fusion OVR task for RF showed that the body modality yields the lowest results. Considering that our set of body features is small, it could be enriched with other features - e.g. deictic gestures - to better model the interaction and improve the computational models performance.

Our definition of interactional trust is based on concepts of alignment, affiliation and credit [45, 110]. During the annotation, annotators focused on social signals to describe the alignment dimension that can also be found in other social phenomenon, in particular engagement [80].

5.3 Error analysis

We analyzed the errors of the WGDE-IG with $\tau = 6$ to determine which segments are the hardest to classify and understand the reason behind it. Thus, we determined the error rate of the model for each segment of each interaction. The error rate corresponds to the ratio of wrong predictions of the model for all 5 seeds during testing on a single segment. Figure 5.6 shows the error rates of segments for two interactions from the dataset. Interaction n°12 has the lowest error rates of all interactions. The model has a fairly low error rate through the interaction, with a significant increase for segments at the end. Interaction n°30 has the highest error rates of all interactions. The model has significant discrepancies in error rates throughout the interaction. During this one, users display sarcastic behavior which could be particularly difficult for the model to discern.

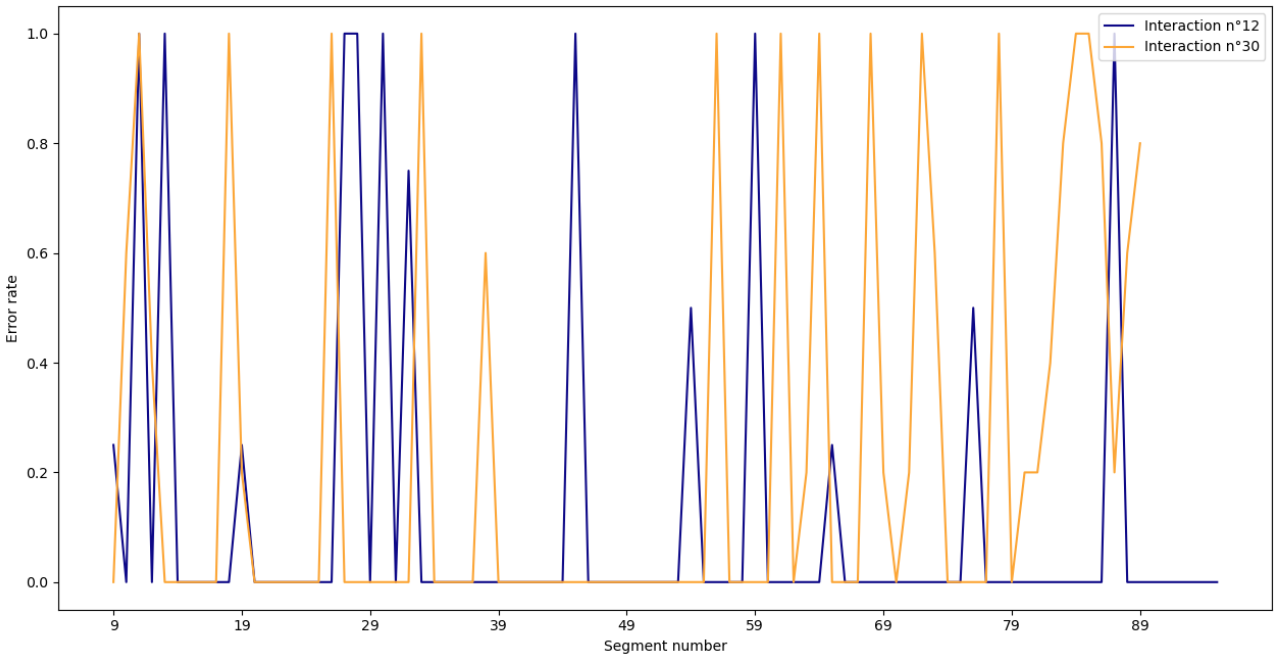

Figure 5.6: Error rate of the model for each segment of interactions n°12 and 30.

To determine the behaviors that are the hardest to analyze for the model, we looked at annotations from the "Social Interaction Content" and "Social Interaction Form" categories for segments that have the highest error rates. We selected segments that have an $errorrate > 0.6$, which corresponds to 20% of segments. The most frequent annotations that appeared for these segments are either "Alignment" or "Compliance". Both items were the most frequent annotated items from the "Social Interaction Content" category. Alignment is a social phenomenon

that is difficult to grasp since it requires understanding the current interactional process and task that users have to focus on at a specific time. Integrating the undergoing task's objective in the model could be a research avenue to further improve the model performance. Jokes are considered signs of trust since they imply that users trust the robot to be able to understand them. Understanding that a is being made joke requires a good comprehension of social norms, which might explain the high error rates of these. Irony and sarcasm are also a form of joke, but they should be analyzed to determine whether they are made to joke with the robot or about the robot.

Mistrusting segments with a high error rate were mostly annotated with "Gaze", "Facial Expression", and "Intonation". Gaze in mistrusting segments are generally associated with moments of doubt, a phenomenon that the model has problems capturing as explained before. An explanation for the presence of facial expressions could be that they can be ambiguous to decipher depending on the context, for instance when used ironically or sarcastically which happens during a few interactions. "Intonation" appears generally in shorter segments during which users speak shortly to answer the robot. The emotional valence of the intonation could be resolved with the addition of users' semantics, which we did not include for technical reasons. Some users exhibited sarcastic facial expressions during mistrusting segments. Sarcasm is difficult to analyze as it requires understanding that behaviors convey the opposite meaning of what is expressed.

Trusting segments with a high error rate were mostly annotated with "Gaze", "Facial expression", and "F formation". F formation describes changes in the users' spatial organization of their interaction. Users can either include or exclude the robot as a member of the interaction depending on their position and the direction they are facing. We represented F formations through the barycenter of the group feature. Including users' body orientation in the set of features could be a research avenue. Participation status appears frequently for both trusting and mistrusting segments with high error rates, which is a good proof of the difficulty to understand changes in participation status and resolve their ambiguity.

5.4 Conclusion

We showed that the Random Forest with an early-fusion mechanism performs the best without taking into account the history of interaction. The study of the model results with a late-fusion mechanism indicates that the voice modality performs the best, followed by the face, and body modality. The multiclass classification setting indicates that taking the segment alone as input is not enough to achieve fair performances.

With our neuronal architecture, we show that taking into account the history leads to increased performances, and fair results for the "Trusting" and "Neutral" classes. Our results indicate that further work is needed to properly capture the dynamics of "Mistrusting" segments. The optimal history length τ still remains unclear, and needs further investigation with additional data. By analyzing the errors made by our final model, we provide a set of features

that can further improve the performances of trust models.

Part IV

Conclusion

Chapter 6

Conclusion

6.1 Contributions

In this thesis, we explored the issues raised by the development of methods to automatically detect trust throughout the interaction and proposed answers to some of these issues. While past research proposed trust monitoring methods, most of them are based on a single type of signal that is generally physically invasive (e.g. physiological data) or required to query users for feedback on their mental state during the model training. Our work focused on a trust analysis method that is completely external to users in the sense that they are never prompted about their mental state nor did they have sensors put on them. We thus contributed to the field of HRI by formulating a new methodology that allows us to automatically model trust throughout the interaction through multimodal features. In the following, we resume the research questions that we formulated at the beginning of the manuscript and provide answers based on the work that we presented.

New methodology for trust analysis in HRI

We introduced a new methodology to the problem of trust analysis in HRI, based on Interactionist Sociology, such as in Conversational Analysis, stemming from approaches prescribed by Ethnomethodology. Rather than considering trust as a mental state, trust is considered a state of the interaction made visible by participants through their behaviors. Trust analysis therefore relies on the observation of these behaviors, and the analysis of the behavior relevance within the interactional sequence. This method allows the researcher to conduct the study of trust dynamics from an external point of view from the interaction, without interrupting the interaction.

RQ 1: Which theoretical framework is applicable to perform a multimodal analysis of trust regularly throughout the interaction ?
We explored Interactionist Sociology theories and showed that they provide tools, through inductive methods, that allow the researcher to observe participants' multimodal behaviors and analyze them in the context of the interaction. From there, we proposed a new methodology of trust analysis in HRI based on constraints that are specific to the field. Through a comparative analysis of the usual Psychological "mentalist" approach and ours, we also provided guidelines on which framework to use according to the study objective, and explained how these can complement each other.

Coding scheme to analyze trust in HRI

Grounding on the new theoretical framework that we proposed, we created a coding scheme for trust in HRI called TURIN that is versatile enough to be used for dyadic or group interactions. TURIN allows to analyze trust dynamics, and study the multimodal behaviors that users express when displaying trust. By comparing it against a common Psychological trust questionnaire, we showed that it can lead to different conclusions, in particular that it can reveal trusting behavior from the user while they report the opposite (and vice versa). We used the coding scheme to collect annotations on the Vernissage dataset and showed that these annotations can be used as ground truth for the computational models we built. Collected annotations are available on the Zenodo platform [1].

RQ 1: Which theoretical framework is applicable to perform a multimodal analysis of trust regularly throughout the interaction ?
We conceived TURIN by leveraging Interactionist Sociology theories and showed that it can unveil trust dynamics through the observation of participants' multimodal behaviors. We also demonstrated that annotations collected with TURIN can be used as the ground truth for multimodal computational models of trust dynamics.
RQ 2: Do homogeneous segments of trust arise within the interaction based on observable behavioral cues ?
We proposed the TURIN coding scheme that allows to focus on the observable characteristics of trust through tangible behavioral cues that indicate trust and highlights homogeneous segments of trust.
RQ 3: How can we discriminate trusting segments from mistrusting ones with tangible behavioral cues ?
In TURIN, we proposed annotation categories that allow the annotator to distinguish between behavioral cues that indicate trusting or mistrusting behavior.

[1] https://doi.org/10.5281/zenodo.8409887

Proposition of a trust-relevant set of features

We proposed a set of multimodal features that can be used to build computational models of trust. The set solely relies on features that are physically non-intrusive for the participants, with a mix of manual and automatically extracted ones. We explored different fusion mechanisms and showed that early-fusion leads to better model performance on trust classification, suggesting the presence of multimodal interplay between features. We also investigated feature importance through late-fusion models' performance and specific feature-importance experiments. The results suggested that audio features carry more weight than others for trust prediction, and that some features were more important for a certain trust category. We also released a librairy of chosen features on a Github repository [2].

RQ 3: How can we discriminate trusting segments from mistrusting ones with tangible behavioral cues ?
We established a set of multimodal features that can be used to build computational model of trust. We showed that some features carry more weight than others, and that some were more specific to a certain trust category.

Conception of multimodal models of trust

We proposed two types of multimodal models of trust: one that is based on traditional machine learning (ML) techniques, and one based on a neuronal architecture. We explored several traditional ML models, and evaluated their performance in binary and multi-class classification. We showed that these models yield fair performances in binary classification, but failed to properly distinguish classes in the multi-class setting. We also showed that the early-fusion mechanism gives the best performance, suggesting a multimodal interplay between features. We then proposed a neuronal architecture composed of two main modules to better model the interactional dynamics relating to trust. Each module is based on hypotheses derived from Interactionist Sociology theories. The first one encodes within-group dynamics to model the interactions between users with different granularities. The second one models the interaction between the robot and the group as a dialogue with a temporal structure. Through experiments, we showed that our full architecture performs better than traditional ML techniques. However, we observed that the performance do not indicate an optimal input sequence length. We also released the code for the neuronal architecture on a Github repository [3].

[2]Code is available here: `https://github.com/GrituX/WGDE_IG`
[3]Code is available here: `https://github.com/GrituX/WGDE_IG`

RQ 3: How can we discriminate trusting segments from mistrusting ones with tangible behavioral cues ?
Early-fusion mechanism leads to quite optimistic performance in binary classification with traditional ML techniques. We designed a neuronal architecture to model interactional dynamics within the user-group, as well as between the robot and users, which led to increased performance. However, such models have trouble distinguishing the "Mistrusting" class in multi-class classification probably due to the heavy class imbalance and low amount of data.

6.2 Perspectives

Performing an online trust detection

Our work on multimodal trust analysis mainly focused on offline detection methods. While our work constitues the stepping stones towards online trust detection, there are still a few challenges to be adressed before reaching this objective.

First, an online trust detection implies that our proposed segmentation method should change to be fully automated. Implementing a sliding window for the classification of a sequence is a traditionnal way to do it [6, 43, 74], but requires the study of the optimal window length. Having a fixed window length for the entire sequence might also help finding an optimal input sequence length. Another way to do the segmentation could be to automatically detect speaking turns, since our method yields segments that are close to speaking turns.

Second, the set of features we proposed has to be revised. A new set that is completely automatic should also be chosen, with the added constraint that feature extraction has to run fast enough for the detection to happen online. This implies either having high computing power, or measuring the feature extraction computing time and selecting features according to be able to run on the target robotic platform. It could be interesting to explore a fully automated set that better represent items from TURIN. This could include deictic features such as having arms crossed, touching its face, as studied in [58], the F-formation of the group, or better features for alignment for instance.

Collecting data with a trust-specific scenario

One challenge of this thesis was to choose a multimodal dataset among the publicly available ones that were not made specifically to study trust, or anything related to it. As we explained in Chapter II, there is no publicly available dataset that includes standard trust questionnaires. As far as we know, the MHHRI is the only one that contains a single question relating to trust in its end questionnaire.

The biggest issue that we face during our experiments was the under-representation of the "Mistrusting" class. Creating a scenario dedicated to collecting more data for this class could be beneficial to improve our models performances. Ideally, the scenario would also include trust questionnaires filled by users at the beginning and end of each interaction. This would

help further validate our coding scheme TURIN, and help refine the items from each of its categories.

Refining TURIN and links with other social phenomenons

We created TURIN from a theoretical perspective by leveraging Interactionist Sociology theories. Trust is a multi-faceted construct that can be impacted by many different factors, either before (e.g. user's cultural background, history of interaction with the robot) or during the interaction depending on the robot's actions. We tried to be as exhaustive as possible for items from the "Social Interaction Form". The "Social Interaction Content" category groups concepts that are linked to trust. In particular, our definition and framing of interactional trust shares some link with the phenomenon of engagement. This link could be studied by either collecting data with a specific scenario to this end, or through the many publicly available datasets on engagement in HRI.

We have not had the opportunity to discuss in depth the elements of both the "Benevolence" and "Integrity" in this thesis. These categories were designed from the most common Psychological models of trust, as explained in Chapter II. Their items could benefit from further validation in scenarios where the robot plays a role that impacts its benevolence and integrity, either in a positive or a negative way.

Improving the models of interactional dynamics for trust

In our neuronal architecture, we proposed a module that encodes the intra-group dynamics within a segment, and a module that models the temporal structure of interactions between the robot and the user group. It would be interesting to model intra-group dynamics in a different way, by designing a neuronal graph module. This could better capture the two-way relations between users' behaviors: participants use other members of the group's behaviors as resources to build their own turn, which in turn becomes a resource for other members' turn.

We could also imagine an architecture that takes into account the difference of amount of context needed to detect trust. For instance, users might react negatively to a robot technical fault and thus might be less inclined to trust its technical capabilities in accomplishing a specific task. While they might react negatively, how strongly they will react is influenced by the fault gravity, but it can also be influenced by the amount of previous mistakes. We could model this through a hierarchical architecture, where the lowest level takes the last few segments as input, and the higher level looks deeper in the past to account for this two-level temporal dependency.

Again, our experiment showed the difficulty to choose a single optimal input sequence length. One way to get around this issue would be to study the performance of our model when it takes the entire interaction as input, and tries to classify each segment right after it is being fed to the network. This could also help better model the longer temporal dependencies between behaviors as explained above, although it would require more data. This method could also only be used for interactions that are somewhat short, since temporal relationships might get lost by the model for particularly long interactions.

Bibliography

[1] Morana Alač, Javier Movellan, and Fumihide Tanaka. "When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics". In: *Social Studies of Science* 41.6 (2011), pp. 893–926.

[2] Alexandre Alahi et al. "Social lstm: Human trajectory prediction in crowded spaces". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 961–971.

[3] Nalini Ambady and Max Weisbuch. "Nonverbal Behavior". In: *Handbook of Social Psychology*. John Wiley & Sons, Inc., 2010. Chap. 13, pp. 464–497.

[4] Alexander M Aroyo et al. "Overtrusting robots : Setting a research agenda to mitigate overtrust in automation". In: *Paladyn, Journal of Behavioral Robotics* 12.1 (2021), pp. 423–436.

[5] Alexander Mois Aroyo et al. "Trust and Social Engineering in Human Robot Interaction: Will a Robot Make You Disclose Sensitive Information, Conform to Its Recommendations or Gamble?" In: *IEEE Robotics and Automation Letters* 3.4 (Oct. 2018), pp. 3701–3708.

[6] Asma Atamna and Chloé Clavel. "HRI-RNN: A user-robot dynamics-oriented RNN for engagement decrease detection". In: *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. Vol. 2020-Octob. 2020, pp. 4198–4202.

[7] J Maxwell Atkinson and John Heritage. *Structures of social action*. Cambridge University Press, 1984.

[8] Agnes Axelsson and Gabriel Skantze. "Do You Follow?" In: ACM, Mar. 2023, pp. 102–111.

[9] Franziska Babel et al. "It Will Not Take Long! Longitudinal Effects of Robot Conflict Resolution Strategies on Compliance, Acceptance and Trust". In: *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. HRI '22. Sapporo, Hokkaido, Japan: IEEE Press, 2022, pp. 225–235.

[10] Tadas Baltrusaitis et al. "OpenFace 2.0: Facial Behavior Analysis Toolkit". In: *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2018)*. IEEE, May 2018, pp. 59–66.

[11] Christoph Bartneck et al. "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots". In: *International Journal of Social Robotics* 1.1 (2009), pp. 71–81.

[12] Christoph Bartneck et al. *Human-robot interaction: An introduction*. Cambridge University Press, 2020.

[13] Atef Ben-Youssef et al. "UE-HRI: a new dataset for the study of user engagement in spontaneous human-robot interactions". In: *Proceedings of the 19th ACM international conference on multimodal interaction*. 2017, pp. 464–472.

[14] Timothy Bickmore and Justine Cassell. "Relational Agents: A Model and Implementation of Building User Trust". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '01. Seattle, Washington, USA: Association for Computing Machinery, 2001, pp. 396–403.

[15] David Cameron et al. "Framing factors: The importance of context and the individual in understanding trust in human-robot interaction". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2015*. Hamburg, Germany, Sept. 2015.

[16] Antonio Camurri, Barbara Mazzarino, and Gualtiero Volpe. "Analysis of Expressive Gesture: The EyesWeb Expressive Gesture Processing Library". In: *Gesture-Based Communication in Human-Computer Interaction*. Ed. by Antonio Camurri and Gualtiero Volpe. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 460–467.

[17] Zhe Cao et al. "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 7291–7299.

[18] Justine Cassell et al. "Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents". In: *Embodied conversational agents* 1 (2000), pp. 1–27.

[19] Oya Celiktutan, Efstratios Skordos, and Hatice Gunes. "Multimodal Human-Human-Robot Interactions (MHHRI) Dataset for Studying Personality and Engagement". In: *IEEE Transactions on Affective Computing* 10.4 (2017), pp. 484–497.

[20] Nitesh V Chawla et al. "SMOTE : Synthetic Minority Over-sampling Technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357.

[21] T. Matthew Ciolek and Adam Kendon. "Environment and the Spatial Arrangement of Conversational Encounters". In: *Sociological Inquiry* 50.3-4 (1980), pp. 237–271.

[22] Jacob Cohen. "A coefficient of agreement for nominal scales". In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

[23] Soumia Dermouche and Catherine Pelachaud. "Engagement Modeling in Dyadic Interaction". In: *2019 International Conference on Multimodal Interaction*. Oct. 2019, pp. 440–445.

[24] Laurence Devillers et al. "Multimodal data collection of human-robot humorous interactions in the Joker project". In: *2015 International Conference on Affective Computing and Intelligent Interaction, ACII 2015* SEPTEMBER (2015), pp. 348–354.

[25] Olive Jean Dunn. "Multiple Comparisons Using Rank Sums". In: *Technometrics* 6.3 (1964), pp. 241–252.

[26] Alessandro Duranti. "Ethnography of speaking: Toward a linguistics of the praxis". In: *Intercultural discourse and communication: The essential readings* (2005), pp. 17–35.

[27] Connor Esterwood and Lionel P. Robert. "Having the Right Attitude: How Attitude Impacts Trust Repair in Human—Robot Interaction". In: *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 2022, pp. 332–341.

[28] Florian Eyben, Martin Wöllmer, and Björn Schuller. "Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor". In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: Association for Computing Machinery, 2010, pp. 1459–1462.

[29] Florian Eyben et al. "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing". In: *IEEE Transactions on Affective Computing* 7.2 (2016), pp. 190–202.

[30] Felix Faber et al. "The humanoid museum tour guide Robotinho". In: *Proceedings - IEEE International Workshop on Robot and Human Interactive Communication* September (2009), pp. 891–896.

[31] Cecilia E Ford and Sandra A Thompson. "Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns". In: *Studies in interactional sociolinguistics* 13 (1996), pp. 134–184.

[32] Cecilia E. Ford, Barbara A. Fox, and Sandra A. Thompson. "Practices in the construction of turns: The "TCU" revisited". In: *Pragmatics* 6.3 (1996), pp. 427–454.

[33] Charles O Frake. "How to ask for a drink in Subanun". In: *American Anthropologist* 66.6 (1964), pp. 127–132.

[34] Harold Garfinkel. "Studies in ethnomethodology". In: *Social Theory Re-Wired*. Routledge, 2023, pp. 58–66.

[35] Erving Goffman. "Presentation of Self in Everyday Life". In: *American Journal of Sociology Goffman, Erving* 55 (1959), pp. 17–25.

[36] Erving Goffman. *Forms of talk*. University of Pennsylvania Press, 1981.

[37] Charles Goodwin et al. "Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning". In: *Sociological inquiry* 50.3-4 (1980), pp. 272–302.

[38] Charles Goodwin. "Conversational organization". In: *Interaction between speakers and hearers* (1981).

[39] Charles Goodwin and John Heritage. "Conversation analysis". In: *Annual review of anthropology* 19.1 (1990), pp. 283–307.

[40] James Griffith. "Measurement of group cohesion in US Army units". In: *Basic and applied social psychology* 9.2 (1988), pp. 149–171.

[41] Rebecca Grossman, Sarit B. Friedman, and Suman Kalra. "Teamwork Processes and Emergent States". In: *The Wiley Blackwell Handbook of the Psychology of Team Working and Collaborative Processes*. Wiley, Mar. 2017, pp. 243–269.

[42] Peter A. Hancock et al. "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction". In: *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53.5 (Oct. 2011), pp. 517–527.

[43] Léo Hemamou et al. "HireNet: A Hierarchical Attention Model for the Automatic Analysis of Asynchronous Video Job Interviews". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 573–581.

[44] John C Heritage. "Interactional accountability: a conversation analytic perspective". In: *Réseaux* 8.1 (1990), pp. 23–49.

[45] Marc Hulcelle et al. "TURIN: A coding system for Trust in hUman Robot INteraction". In: *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2021, pp. 1–8.

[46] Hayley Hung and Daniel Gatica-Perez. "Estimating cohesion in small groups using audio-visual nonverbal behavior". In: *IEEE Transactions on Multimedia* 12.6 (2010), pp. 563–575.

[47] Dinesh Babu Jayagopi et al. "The vernissage corpus: A conversational Human-Robot-Interaction dataset". In: *ACM/IEEE International Conference on Human-Robot Interaction* (2013), pp. 149–150.

[48] Xiaoqi Jiao et al. *TinyBERT: Distilling BERT for Natural Language Understanding*. 2020.

[49] Halimahtun M. Khalid et al. "Exploring Psycho-Physiological Correlates to Trust: Implications for Human-Robot-Human Interaction". In: *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 60.1 (2016), pp. 697–701.

[50] Zahra Rezaei Khavas. "A Review on Trust in Human-Robot Interaction". In: *pre-print arXiv.2105.10045* (2021).

[51] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. "Modeling Trust in Human-Robot Interaction: A Survey". In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol. 12483 LNAI. 2020, pp. 529–541.

[52] Cory D Kidd and Cynthia Breazeal. "Effect of a robot on user perceptions". In: *2004 IEEE/RSJ international conference on intelligent robots and systems (IROS)(IEEE Cat. No. 04CH37566)*. Vol. 4. IEEE. 2004, pp. 3559–3564.

[53] Bing Cai Kok and Harold Soh. "Trust in Robots: Challenges and Opportunities". In: *Current Robotics Reports* 1 (4 Dec. 2020), pp. 297–309.

[54] Steve WJ Kozlowski and Katherine J Klein. "A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes." In: *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions*. Ed. by K. J. Klein and S. W. J. Kozlowski. Jossey-Bass/Wiley, 2000, pp. 3–90.

[55] William H. Kruskal and W. Allen Wallis. "Use of Ranks in One-Criterion Variance Analysis". In: *Journal of the American Statistical Association* 47.260 (1952), pp. 583–621.

[56] Cherie Lacey and Catherine Caudwell. "Cuteness as a 'Dark Pattern' in Home Robots". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Mar. 2019, pp. 374–381.

[57] Allison Langer et al. "Trust in socially assistive robots: Considerations for use in rehabilitation". In: *Neuroscience & Biobehavioral Reviews* 104 (2019), pp. 231–239.

[58] Jin Joo Lee et al. "Computationally Modeling Interpersonal Trust". In: *Frontiers in Psychology* 4 (2013).

[59] John D Lee and Katrina A See. "Trust in automation: Designing for appropriate reliance Human Factors". In: *Human Factors* 46.1 (2004), pp. 50–80.

[60] Stephen C Levinson. *Pragmatics*. Cambridge university press, 1983.

[61] Kurt Lewin. "Behavior and development as a function of the total situation." In: *Manual of child psychology*. Ed. by L. Carmichael. John Wiley & Sons Inc, 1946, pp. 791–844.

[62] Scott M Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017.

[63] Merce Mach, Simon Dolan, and Shay Tzafrir. "The differential effect of team members' trust on team performance: The mediation role of team cohesion". In: *Journal of occupational and organizational psychology* 83.3 (2010), pp. 771–794.

[64] Maria Madsen and Shirley Gregor. "Measuring human-computer trust". In: *11th australasian conference on information systems*. Vol. 53. Citeseer. 2000, pp. 6–8.

[65] Navonil Majumder et al. "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (July 2019), pp. 6818–6825.

[66] Lucien Maman et al. "Exploiting the Interplay between Social and Task Dimensions of Cohesion to Predict its Dynamics Leveraging Social Sciences". In: Association for Computing Machinery, 2021, pp. 16–24.

[67] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. "A Temporally Based Framework and Taxonomy of Team Processes". In: *The Academy of Management Review* 26.3 (2001), pp. 356–376.

[68] Michelle A Marks, John E Mathieu, and Stephen J Zaccaro. "A temporally based framework and taxonomy of team processes". In: *Academy of management review* 26.3 (2001), pp. 356–376.

[69] Stephen Paul Marsh. "Formalising trust as a computational concept". In: (1994).

[70] Christoforos Mavrogiannis et al. "Core challenges of social robot navigation: A survey". In: *ACM Transactions on Human-Robot Interaction* 12.3 (2023), pp. 1–39.

[71] Roger C Mayer, James H Davis, and F David Schoorman. "An integrative model of organizational trust". In: *Academy of Management Review* 20.3 (1995), pp. 709–735.

[72] D McNeill. *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL: Univ. of Chicago Press, 1992.

[73] Richard L. Moreland. "Are Dyads Really Groups?" In: *Small Group Research* 41.2 (2010), pp. 251–267.

[74] Louis-philippe Morency. "Modeling Human Communication Dynamics". In: *IEEE Signal Processing Magazine* 27.5 (2010), pp. 112–116.

[75] Masahiro Mori. "The uncanny valley: the original essay by Masahiro Mori". In: *IEEE Spectrum* 6 (1970).

[76] Jonathan Mumm and Bilge Mutlu. "Human-robot proxemics: Physical and psychological distancing in human-robot interaction". In: *HRI 2011 - Proceedings of the 6th ACM/IEEE International Conference on Human-Robot Interaction* (2011), pp. 331–338.

[77] Manisha Natarajan and Matthew Gombolay. "Effects of Anthropomorphism and Accountability on Trust in Human Robot Interaction". In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 33–42.

[78] Mollik Nayyar and Alan R. Wagner. "When Should a Robot Apologize? Understanding How Timing Affects Human-Robot Trust Repair". In: *Social Robotics*. Ed. by Shuzhi Sam Ge et al. Cham: Springer International Publishing, 2018, pp. 265–274.

[79] Tatsuya Nomura, Takayuki Kanda, and Tomohiro Suzuki. "Experimental investigation into influence of negative attitudes toward robots on human-robot interaction". In: *AI and Society* 20.2 (2006), pp. 138–150.

[80] Catharine Oertel et al. "Engagement in Human-Agent Interaction: An Overview". In: *Frontiers in Robotics and AI* 7 (Aug. 2020).

[81] Regina Pally. "Emotional Processing; The mind-body connection". In: *The International journal of psycho-analysis* 79.2 (1998), pp. 349–362.

[82] Harold Pashler and Christine R Harris. "Is the replicability crisis overblown? Three arguments examined". In: *Perspectives on Psychological Science* 7.6 (2012), pp. 531–536.

[83] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035.

[84] Sabine Payr and Robert Trappl. *Agent culture: human-agent interaction in a multicultural world*. CRC Press, 2004.

[85] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[86] Martin J. Pickering and Simon Garrod. "The interactive-alignment model: Developments and refinements". In: *Behavioral and Brain Sciences* 27.2 (2004), pp. 212–225.

[87] Karl Popper. *The logic of scientific discovery*. University Press, 1959.

[88] George Psathas. *Interaction competence*. University Press of Amer, 1990.

[89] George Psathas. "Talk and social structure" and "studies of work". In: *Human studies* 18.2-3 (1995), pp. 139–155.

[90] Tammy Rapp et al. "Team Emergent States: What Has Emerged in The Literature Over 20 Years". In: *Small Group Research* 52 (1 Feb. 2021), pp. 68–102.

[91] Tina L Robbins and Angelo S DeNisi. "A closer look at interpersonal affect as a distinct influence on cognitive processing in performance evaluations." In: *Journal of Applied Psychology* 79.3 (1994), p. 341.

[92] Paul Robinette, Alan R Wagner, and Ayanna M Howard. "Building and maintaining trust between humans and guidance robots in an emergency". In: *AAAI spring symposium* (2013), pp. 78–83.

[93] Paul Robinette et al. "Overtrust of robots in emergency evacuation scenarios". In: *ACM/IEEE International Conference on Human-Robot Interaction* 2016-April (2016), pp. 101–108.

[94] Nicolas Rollet and Chloé Clavel. ""Talk to you later": Doing social robotics with conversation analysis. Towards the development of an automatic system for the prediction of disengagement". In: *Interaction Studies* 21.2 (2020), pp. 268–292.

[95] Nicolas Rollet and Christian Licoppe. "Why (pre)closing matters. The case of human-robot interaction". In: (2019).

[96] Julian B Rotter. "A new scale for the measurement of interpersonal trust." In: *Journal of personality* 35.4 (1967), pp. 651–665.

[97] Denise M Rousseau et al. "Not So Different After All : a Cross-Discipline View of Trust". In: *Academy of Management Review* 23.3 (1998), pp. 393–404.

[98] Damien Rudaz et al. "From Inanimate Object to Agent: Impact of Pre-Beginnings on the Emergence of Greetings with a Robot". In: *J. Hum.-Robot Interact.* 12.3 (Apr. 2023).

[99] J. P. de Ruiter and Saul Albert. "An Appeal for a Methodological Fusion of Conversation Analysis and Experimental Psychology". In: *Research on Language and Social Interaction* 50 (1 Jan. 2017), pp. 90–107.

[100] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. "A simplest systematics for the organization of turn taking for conversation". In: *Studies in the Organization of Conversational Interaction*. Ed. by Jim Schenkein. Academic Press, 1978, pp. 7–55.

[101] Hanan Salam et al. "Fully Automatic Analysis of Engagement and Its Relationship to Personality in Human-Robot Interactions". In: *IEEE Access* 5 (2017), pp. 705–721.

[102] Maha Salem et al. "Towards Safe and Trustworthy Social Robots: Ethical Challenges and Practical Issues". In: 2015, pp. 584–593.

[103] Maha Salem et al. "Would You Trust a (Faulty) Robot?: Effects of Error, Task Type and Personality on Human-Robot Cooperation and Trust". In: *ACM/IEEE International Conference on Human-Robot Interaction*. Vol. 2015-March. 2015, pp. 141–148.

[104] Kristin E. Schaefer. "The Perception and Measurement Of Human-robot Trust". PhD thesis. Orlando, FL, USA: University of Central Florida, 2013.

[105] Kristin E. Schaefer. "Measuring trust in human robot interactions: Development of the "trust perception scale-HRI"". In: *Robust Intelligence and Trust in Autonomous Systems* (Jan. 2016), pp. 191–218.

[106] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. ""I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Mar. 2019, pp. 57–65.

[107] Samuel S Shapiro and RS Francia. "An approximate analysis of variance test for normality". In: *Journal of the American statistical Association* 67.337 (1972), pp. 215–216.

[108] Judy Hanwen Shen, Agata Lapedriza, and Rosalind W. Picard. "Unintentional affective priming during labeling may bias labels". In: *2019 8th International Conference on Affective Computing and Intelligent Interaction, ACII 2019* (2019), pp. 587–593.

[109] Candace L Sidner et al. "Explorations in engagement for humans and robots". In: *Artificial Intelligence* 166.1-2 (2005), pp. 140–164.

[110] Tanya Stivers. "Stance, Alignment, and Affiliation During Storytelling: When Nodding Is a Token of Affiliation". In: *Research on Language and Social Interaction* 41.1 (2008), pp. 31–57.

[111] Tanya Stivers and Jeffrey D. Robinson. "A preference for progressivity in interaction". In: *Language in Society* 35.3 (2006), pp. 367–392.

[112] Randall Stokes and John P. Hewitt. "Aligning Actions". In: *American Sociological Review* 41.5 (Oct. 1976), pp. 838–842.

[113] Dag Sverre Syrdal et al. "The Negative Attitudes towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study". In: *Proceedings of AISB09* (2009), pp. 109–115.

[114] Benedict Tay, Younbo Jung, and Taezoon Park. "When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction". In: *Computers in Human Behavior* 38 (2014), pp. 75–84.

[115] Suzanne Tolmeijer et al. "Taxonomy of Trust-Relevant Failures and Mitigation Strategies". In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 3–12.

[116] Daniel Tozadore et al. "Wizard of Oz vs autonomous: Children's perception changes according to robot's operation condition". In: *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2017, pp. 664–669.

[117] Federico Visi, Rodrigo Schramm, and Eduardo Miranda. "Use of Body Motion to Enhance Traditional Musical Instruments: A Multimodal Embodied Approach to Gesture Mapping , Composition and Performance". In: July 2014.

[118] Alan R Wagner and Ronald C Arkin. "Recognizing situations that demand trust". In: *2011 RO-MAN*. IEEE. 2011, pp. 7–14.

[119] Sheida White. "Backchannels across cultures: A study of Americans and Japanese1". In: *Language in society* 18.1 (1989), pp. 59–76.

[120] Frank Wilcoxon. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1.6 (1945), pp. 80–83.

[121] Dennis L Wilson. "Asymptotic properties of nearest neighbor rules using edited data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 3 (1972), pp. 408–421.

[122] Peter Wittenburg et al. "ELAN: A professional framework for multimodality research". In: *5th international conference on language resources and evaluation (LREC 2006)*. 2006, pp. 1556–1559.

[123] Anqi Xu and Gregory Dudek. "OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations". In: vol. 2015-March. IEEE Computer Society, Mar. 2015, pp. 221–228.

[124] Rosemarie E. Yagoda and Douglas J. Gillan. "You Want Me to Trust a ROBOT? The Development of a Human-Robot Interaction Trust Scale". In: *International Journal of Social Robotics* 4 (3 2012), pp. 235–248.

[125] Baosong Yang et al. "Context-Aware Self-Attention Networks". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.01 (July 2019), pp. 387–394.

[126] Fangkai Yang et al. "A dataset of human and robot approach behaviors into small free-standing conversational groups". In: *PLoS ONE* 16.2 February (2021), pp. 1–24.

[127] Victor H Yngve. "On getting a word in edgewise". In: *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*. 1970, pp. 567–578.

[128] Jakub Złotowski et al. "Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction". In: *International Journal of Social Robotics* 7 (3 June 2015), pp. 347–360.

[129] Dominik Zunt. "Who did actually invent the word "robot" and what does it mean?" In: *The Karel Čapek website* (2013).

ECOLE
DOCTORALE

Titre : Analyse automatique de la confiance au cours d'une interaction homme-robot par descripteurs multimodaux et architectures neuronales récurrentes

Mots clés : Confiance, Réseaux Neuronaux Récurrents, Apprentissage Statistique, Mutlimodal, Sociologie Interactioniste

Résumé : La confiance est une notion importante en interaction humain-robot puisqu'elle impacte la qualité des relations entre les partenaires d'interaction et ainsi les performances de la tâche en cours. Les recherches autour de la confiance se sont essentiellement circonscrites autour des analyses des effets socio-psychologiques sur l'utilisateur du design du robot, ou de son comportement. Les mesures de la confiance se font généralement au début et fin de l'interaction par des questionnaires remplis par les utilisateurs eux-mêmes. Dans cette thèse, nous nous intéressons à une analyse de la dynamique de la confiance conduite régulièrement tout au long de l'interaction. Comme les approches usuelles de Psychologie dites mentalistes ne nous permettent pas de faire ceci, nous faisons appel aux théories de la Sociologie Interactioniste afin d'établir un schéma de codage TURIN (Trust in hUman Robot INteraction) dédié à cela. Ensuite, nous utilisons des outils de Machine Learning afin de développer des modèles d'analyse automatique de la confiance. Nous proposons une nouvelle méthodologie permettant de conduire l'analyse au cours de l'interaction, en s'appuyant sur des approches simples dans un premier temps, puis sur une nouvelle architecture neuronale récurrente dans un deuxième temps. Nous analysons ensuite nos modèles afin de déterminer les indices comportementaux les plus pertinents et comprendre les types d'erreur que ceux-ci commettent.

Title : Automatic analysis of trust over the course of a human-robot interaction using multimodal features and recurrent neural architectures

Keywords : Trust, Recurrent Neural Network, Machine Learning, Multimodal, Interactionist Sociology

Abstract : Trust is an important psychological construct in HRI as it mitigates the relationship qualities between partners of an interaction, as well as the performance of the interaction task. Research on trust were essentially organized around the study of socio-psychological effects of the robot design and behavior on users. Trust is usually measured through questionnaires filled by users themselves at the beginning and end of the interaction. In this thesis, we tackle the issue of automatic analysis of trust dynamics during the course of interaction. The standard Psychological approaches used in HRI to study, coming from a mentalist perspective, do not currently allow such analysis. We thus leverage Interactionist Sociology theories to create a coding scheme named TURIN (Trust in hUman Robot INteraction) dedicated to this task. From there, we use Machine Learning tools to develop multimodal models of trust. We propose a new methodology that allows to conduct the analysis over the course of the interaction, first through simple models, then by the design of a specific recurrent neural architecture. We finish by an analysis of ours models to determine which behaviors are the most indicative of trust and understand the types of errors that they make.

Institut Polytechnique de Paris
91120 Palaiseau, France