Towards detection of nudges in Human-Human and Human-Machine interactions

Natalia Kalashnikova

université PARIS-SACLAY

Towards detection of nudges in Human-Human and Human-Machine interactions

Vers la détection des nudges dans les interactions Humain-Humain et Humain-Machine

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 580, Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat: Informatique
Graduate School : Informatique et Sciences du Numérique.
Référent : Faculté des Sciences d'Orsay.

Thèse préparée dans l'unité de recherche **Laboratoire Interdisciplinaire des Sciences du Numérique** (Université Paris-Saclay, CNRS), sous la direction de **Laurence DEVILLERS**, Professeure des universités, le co-encadrement de **Ioana VASILESCU**, Directrice de recherche

Thèse soutenue à Paris-Saclay, le 28 juin 2024, par

Natalia KALASHNIKOVA

Composition du jury

Membres du jury avec voix délibérative

Pierre ZWEIGENBAUM Directeur de recherche, Université Paris-Saclay, LISN	Président
Patrice BELLOT Professeur des universités, Aix-Marseille Université, LIS-CNRS	Rapporteur & Examinateur
Marie TAHON Professeure des universités, Le Mans Université, LIUM	Rapporteure & Examinatrice
Nicolas AUDIBERT Maître de Conférences, Université Nouvelle Sorbonne	Examinateur
Alexandre PAUCHET Maître de Conférences HDR, Université de Rouen, INSA	Examinateur
Laura SPINU Professeur, Kingsborough Community College	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASG031

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre: Vers la détection des nudges dans les interaction Humain-Humain et Humain-Machine

Mots clés: nudge, alignement, création de corpus, prédiction automatique

Résumé: Les techniques qui influencent indirectement la prise de décision des humains, connues sous le nom de "nudge" (Thaler and Sunstein, 2008), sont peu étudiées dans les interactions parlées. Les nudges linguistiques sont des techniques de manipulation douce fondées sur les biais cognitifs et utilisent les moyens linguistiques pour encourager les changements dans la prise de décision des humains sans aucune restrictions ou sanctions pour leur choix. Addressées directement au destinataire (par exemple, sous forme de lettre ou de note), ces techniques ont prouvé leur efficacité dans plusieurs domaines. Néanmoins, avec la présence de plus en plus répandu des agents conversationnels au quotidien, plusieurs questions se posent sur l'impact du type de l'interlocuteur et la réaction de différents types de public aux nudges. En tenant compte de cette connaissance préalable, nous étudions plusieurs descripteurs linguistiques et paralinguistiques et posons la question de la pertinence d'un modèle qui prédit si quelqu'un a été verbalement manipulé. Les recherches dans ce domaine en sont encore à leurs débuts, nous proposons donc d'abord une méthodologie innovative pour la collection de données dans le but d'estimer la propension des participants à être nudgés (influencés). Nous avons testé deux types de publics : les enfants et les adultes. Le protocole compare les interactions contenant une intervention qui influence le choix (nudge) avec trois agents conversationnels (robot Pepper, enceinte Google Home, humain). Dans l'expérience avec les adultes, nous avons comparé les scores des participants quant à leur volonté d'adopter des habitudes écologiques après le nudge avec leurs scores de base afin de mesurer l'influence des nudges. Dans l'expérience avec les enfants, nous avons comparé le nombre de billes qu'ils étaient prêts à garder pour eux après le nudge avec le nombre de billes qu'ils voulaient garder avant le nudge pendant le jeu. En utilisant cette méthodologie, nous avons enregistré 22 heures d'échanges entre des adultes et trois agents conversationnels (le robot Pepper, le haut-parleur Google Home et un humain) et 10 heures d'échanges entre des enfants et les mêmes agents conversationnels. Dans un premier temps, ces données ont été transcrites manuellement et segmentées en tours de parole, puis annotées à différents niveaux affectifs. Deuxièmement, pour mesurer la capacité des différents agents conversationnels à donner des nudges de manière efficace, nous avons analysé la prise de décision des participants en fonction de l'interlocuteur et du type de nudge. Plus précisément, nous avons étudié la corrélation entre les états émotionnels des participants et leurs réponses aux nudges et aux agents conversationnels. Troisièmement, pour mieux comprendre comment l'incarnation d'un agent conversationnel peut influencer la propension d'un participant à recevoir des encouragements, nous avons proposé une comparaison de certains éléments paralinguistiques, lexicaux et discursifs pertinents des participants selon le type d'agent conversationnel. Enfin, nous avons utilisé différentes combinaisons d'annotations émotionnelles, de transcriptions et de données audio provenant des expériences enregistrées pour construire un modèle d'apprentissage profond basé sur des caractéristiques acoustiques, textuelles et des états émotionnels afin de prédire si le participant a été nudgé. Les principaux résultats soulignent que nos participants ont été nudgés quel que soit leur groupe d'âge, avec un effet plus important sur les adultes.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title: Towards detection of nudges in Human-Human and Human-Machine Interactions

Keywords: nudge, alignment, data acquisition, automatic prediction

Abstract: Nudges, techniques that indirectly influence human decision-making (Thaler and Sunstein, 2008), are little studied in spoken interactions. Linguistic nudges are techniques of latent manipulation based on cognitive biases that use linguistic means to encourage changes in human decision-making without any restrictions or penalties for their choices. Addressed directly to the recipient (e.g., in the form of a letter or a note), these techniques have proven their effectiveness in many domains. However, with the growing presence of conversational agents in everyday life, several questions have been raised about the impact of the type of interlocutor and the reaction of different types of public to nudges. With this prior knowledge in mind, we study several paralinguistic and linguistic features and question the relevance of a model that predicts whether someone has been verbally influenced. This domain is in its early stages; thus, we first propose an innovative methodology for data collection with the goal of estimating participants' propensity to be nudged. We tested two populations: children and adults. The protocol compares nudging interaction with three conversational agents (robot Pepper, smart-speaker Google Home, and human). In the experiment with adults, we compared the participants' scores of willingness to adopt selected ecological habits after the nudge with their baseline scores to measure the influence of nudges. In the experiment with children, we compared the number of little balls they were willing to keep for themselves after the nudge with the number of balls they wanted to keep before the nudge during the game. Using this methodology, we recorded 22 hours of exchanges of adults with three conversational agents (robot Pepper, smart-speaker Google Home, and human) and 10 hours of exchanges of children with the same conversational agents. Firstly, these data were manually transcribed and segmented into speaking turns and then annotated on different affective levels. Secondly, to measure the ability of the various conversational agents to nudge effectively, we analyzed the participants' decision-making according to the interlocutor and the type of nudges. Specifically, we studied the correlation between participants' emotional states and their answers to nudges and conversational agents. Thirdly, to better understand how the embodiment of a conversational agent could influence a participant's propensity to be nudged, we proposed a comparison of some relevant paralinguistic, lexical, and discursive cues of participants regarding the type of conversational agent. Finally, we used different combinations of emotional annotations, transcriptions, and audio data from the recorded experiments to build a deep-learning model based on acoustic, textual features, and emotional states to predict whether the participant was nudged. The main results underline that our participants were nudged regardless of their age group, with a more significant impact on adults.

I would just like to take this moment to say to all the young girls out there who dream about science as a profession: go for it. It is the greatest job in the world. And if anybody tells you you can't, don't listen.

Amy Farrah Fowler, TBBT, season 12, episode 24.

Acknowledgements

Je veux remercier...

les membres de mon jury d'avoir pris le temps pour lire mon travail en fin de l'année scolaire, montrer l'intérêt et le regard critique. Mes directrices de thèse de m'avoir donné la chance de faire une thèse sur un sujet passionnant.

L'équipe du Collège des Bernardins et du laboratoire RITM ainsi que tous les participants de nos expériences, les bénévoles du collège des Bernardins et mes collègues du labo qui venaient nous aider. Même si à la fin j'en pouvais plus de porter les robots, ce sont mes moments préférés du travail.

Sandy pour être là malgré les changements dans ma vie et dans ce monde, et de nos moments de petites parisiennes qui me donnaient la force pour continuer.

Ania et Denis pour le soutien et la compréhension absolue, et peu importe à Paris ou à Reims votre chez vous c'est notre petit coin de la mère patrie.

Natasha et Tena de "Chitalochka" pour nos discussions les plus sincères et profondes.

Liza, pour m'encourager de me battre pour moi.

Tania, pour ta gentillesse et le coeur ouvert qui m'accompagnent depuis le fameux "Moi je veux vivre, aller haut, pouvoir me dire, que c'est beau".

Votre amour est avec moi tous les jours.

ma famille en Russie pour m'avoir donné depuis mon enfance la chose la plus chère qui est la liberté.

Anne et Olivier pour l'encouragement, le soutien et le sentiment de faire partie

d'un foyer.

Mathilde de notre travail ensemble, de m'avoir donné des conseils et soutient émotionnel, tu es ma grande sœur de recherche.

Francesca, Pauline, et Tom d'avoir toujours été réactifs et intéressés par mon travail, d'avoir travaillé plus qu'il fallait pour que je puisse avancer.

tout le monde qui était à un moment au 2ème étage du bâtiment 507 pour l'ambiance qui donnait envie de venir à bout de l'île de France : Sofiya, Yajing, Lufei. Elise, pour la course au foie gras "ou quoi", Aina pour le cours de posters, Maxime pour le vaccin contre Hanabi, David pour l'attention qu'il porte aux autres, François pour la gentillesse et la place au bureau.

Docteur Hugues Ali Mehenni, Docteur Paul Lerner, mon collègue Docteur Alban Petit et Docteur Shu Okabe pour m'avoir accepté comme je suis, rigoler de mes histoires et mes blagues, expliquer, accompagner, faire les tours de manège, prendre des verres et jouer. Grâce à vous, je me suis sentie à ma place pour la première fois dans ma vie. Passer ces années avec vous était un grand cadeau et plaisir que je n'ai pas pu imaginer. J'espère que maintenant j'aurai le droit de m'installer à votre table des docteurs aux restos.

Alexandre, pour tout ce qu'on a partagé et de m'avoir accompagné sur le chemin de devenir adultes ensemble. J'y suis arrivée parce que tu pensais m'être capable.

mon plus que collègue mais partenaire Alban de partager avec moi les hauts et les bas et de m'attraper quand je suis en chute libre. Grâce à la brillance de tes yeux quand tu me regardes les choses sont moins lourdes.

Je finis le chapitre de l'expérience que je voulais vivre depuis mon adolescence, merci encore à vous d'en faire partie. Les autres aventures nous attendent.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context

50 people are currently looking at this item and there are only 10 left with 50% off!

Did you feel the urge to buy this item and be among the luckiest ten? It is because of nudge - a strategy that influences our behavior using cognitive biases, such as aversion loss in the example. Nudges are all over us, but how are they studied? Our example shows that the concept of nudges is an interdisciplinary subject and lies between economics, sociology, linguistics, psychology, etc.

In 2008, Thaler and Sunstein (2008) highlighted the concept of nudges in the domain of behavioral economics, defining them as *"any aspect of the choice architecture that alters people's behavior predictably without forbidding any options or significantly changing their economic incentives. The intervention must be easy and cheap to avoid to count as a mere nudge."* They argued that to be efficient, nudges rely on cognitive biases that make our decisions fast, effortless, and unconscious.

Since then, conventional (not linguistic) nudges have been effectively applied in different domains to predictably change the users' environment for their better choices. For example, arranging the placement of fruits at eye level in the cafeteria

for a healthier dessert choice (Mulderrig, 2018).

As for linguistic nudges, in many cases, they appear in the written form, like notes or reminders. For example, letters using peer-comparison bias (the choice of the user is compared with the choice of other users) to steer the population in receiving the COVID-19 vaccine (Sasaki et al., 2022). However, linguistic nudges have been little studied in spoken interactions. Rare studies in this domain have only analyzed the effectiveness of nudges on human decision-making without investigating how the effectiveness of nudges depends on the character of who nudges and to whom nudges are applied (Kawano et al., 2022).

Furthermore, in spoken interactions, linguistic nudges can be used by conversational agents and connected objects to simplify people's lives. The common use of connected devices and the growing capacity to gather and analyze data allow the choice architect to create dynamically personalized nudges, which promise to be even more powerful in changing someone's opinions (Bergram et al., 2022). However, a few limits of ethical norms exist for a conversational agent to enter a more private zone, influencing opinions or purchases. Therefore, one can question the propensity of humans to be nudged by machines and the contexts that trigger successful nudging.

The particularity of this thesis is nudging strategies in spoken interactions, we, thus, need to analyze them from an informational level of abstraction. This approach proposes to analyze how the information was presented, the structure of the nudging strategies, and the relationship between the user and the system.

Therefore, another issue when it comes to nudging through interactions is the alignment (mimicry, entrainment, adjustment) between interlocutors. In the same vein, as the more general issue of verbal nudging, one can question the alignment in both linguistic and emotional dimensions as a factor that could contribute to successful nudging.

The alignment between interlocutors conditions human-human and human-machine

interactions. It is well known that interlocutors adjust their communicative behavior during a successful interaction by being more similar at different linguistic and paralinguistic levels (Bonin et al., 2012). However, the alignment is also realized at the emotional level, characterized by two factors. Humans tend to stick with the same emotional state during the conversation, and at the same time, we are influenced by the emotional state of our interlocutor. Previous research showed that the cues of linguistic, paralinguistic, and emotional alignment allowed to predict the outcome of marriage (Nasir et al., 2015) and increased the willingness to speak about personal affairs to reduce a feeling of loneliness with a machine (Sabelli et al., 2011).

Given all these relevant aspects highlighted above and the lack of studies, we hypothesize that the study of nudges in spoken interactions through the analysis of linguistic, paralinguistic, and emotional levels of alignment would lead to a better understanding of the relationships between the nudgee, the nudger, and the nudge.

This PhD thesis is supported by Chair AI HUMAAINE directed by Laurence Devillers. This project studies the nudges in social interactions regarding multiple factors, such as the interlocutor, the audience, ethical norms, etc.

1.2 Research questions

This thesis aims to propose an approach to detect if the speaker was nudged regarding their linguistic, paralinguistic, and emotional cues during an oral interaction. This approach is applied to two different audiences: children and adults. The conversation is characterized by inter-personal emotional dependencies, which are modelized with context-aware emotion classification. We adopt the approach of context-aware emotion recognition to keep track of the speaker's contextual information which is modelized with linguistic, paralinguistic, and emotional features.

However, this aim could not be achieved without a better understanding of the relationship between the nudgee, the nudger, and the nudge. We, therefore, address

the following research questions:

1. Are nudges efficient in spoken interactions?
2. If so, how can it be measured?
3. Is it possible to change someone's opinion against mainstream ideas?
4. How do linguistic and paralinguistic behavior, as well as emotional states, change during nudging interactions regarding the following three research axes:
    - the character of the nudgee - an adult or a child;
    - their propensity to be nudged - if they have changed their opinions or not;
    - the type of their interlocutor (nudger) - a human, or a smart-speaker Google Home, or a robot Pepper.

To answer these research questions, we first propose a similar methodology of data acquisition for adults and children, where we measure the baseline score and then how it changes after nudging.

Secondly, we propose measures to evaluate the effectiveness of nudges in the recorded data and realize a statistical analysis to define the participants' propensity to be influenced. We also statistically correlate the emotional labels used for the annotation, the type of participant's interlocutor, and their propensity to be nudged.

Thirdly, we analyze how linguistic and paralinguistic parameters change among participants regarding their propensity to be nudged and the type of their interlocutor and if these changes follow the same pattern between adults and children.

Finally, these analyses allow us to propose a multimodal deep-learning system based on linguistic, paralinguistic, and emotional features that predicts whether a speaker was nudged or not. A supplementary challenge to this task is to adapt the system to the small size of our corpus.

1.3 Outline

This thesis is organized according to the research questions.

In Chapter 2 we introduce the fundamental principles of the theory of nudges and discuss its main ethical issues. We present examples of studies of linguistic nudges. Then we describe linguistic and paralinguistic cues analyzed in previous research to describe speech addressed to different conversational agents. In the last part of this chapter, we review the most common techniques of textual and acoustic data preprocessing and the context-aware emotion recognition architectures mainly used as a baseline in recent studies and the current state-of-art.

We explain in detail the methodology of data acquisition in Chapter 3. It contains a description of the physical procedure and examples of the exchanges. The annotation guide and the description of the collected data follow it. The first part of this chapter presents the data acquisition in the experiment with adults and the second one presents the data acquisition in the experiment with children.

The first part of Chapter 4 explains the proposed measures to evaluate the effectiveness of nudges, as well as the statistical results that compare the effectiveness of nudges and 1) in general, regarding 2) the type of the conversational agent, 3) the type of nudge applied to participants, 3) the type of participants, and 4) their personal traits measured with Big Five personality test. We also present the correlation between participants' metadata and their propensity to be nudged. The second part analyzes the correlation between the discussed factors and participants' emotional states.

We describe the differences in linguistic and paralinguistic cues of adults and children in Chapter 5. Here, we follow the same axes of analysis: 1) how they differ regarding the type of conversational agent, 2) the propensity to be nudged, and 3) the type of participant.

Our last contribution is presented in Chapter 6. First, we explain the preprocessing step for acoustic and textual data. Secondly, we propose a system that

models the contextual information of utterances based on linguistic, paralinguistic, and emotional features to predict whether the participant was nudged.

The thesis concludes with a global overview of our work and the axes of future research.

1.4 Publications

During the thesis, we published the following articles.

Linguistic Nudges and Verbal Interaction with Robots, Smart-Speakers, and Humans

Authors: Natalia Kalashnikova, Ioana Vasilescu, Laurence Devillers

Year: 2024

Conference: Language Resources and Evaluation Conference (LREC'24)

Abstract: This paper describes a data collection methodology and emotion annotation of dyadic interactions between a human, a Pepper robot, a Google Home smart-speaker, or another human. The collected 16 hours of audio recordings were used to analyze the propensity to change someone's opinions about ecological behavior regarding the type of conversational agent, the kind of nudges, and the speaker's emotional state. We describe the statistics of data collection and annotation. We also report the first results, which showed that humans change their opinions on more questions with a human than with a device, even against mainstream ideas. We observe a correlation between a certain emotional state and the interlocutor and a human's propensity to be influenced. We also reported the results of the studies that investigated the effect of human likeness on speech using our data.

Do We Speak to Robots Looking Like Humans As We Speak to Humans? A Study of Pitch in French Human-Machine and Human-Human Interactions

Authors: Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence

Devillers

Year: 2023

Conference: Companion Publication of the 25th International Conference on Multimodal Interaction ((ICMI '23 Companion)

Abstract: Robot-directed speech refers to speech to a robotic device (speakers, computers, etc.). Studies have investigated the phonetic and linguistic properties of this type of speech and shown that, humans tend to change their pitch when talking to a robot *vs* to a human. Parallelly, it has shown that the anthropomorphism of the devices affects the social aspect of interaction. However, none have investigated the effect of the device's human-likeliness on linguistic realizations. This study proposes to fill this gap by comparing the effect of anthropomorphism in speech directed at a speaker *vs* a humanoid robot *vs* a human by analyzing the F0 values and range in the three conditions, and how these parameters change throughout the conversation. The data from 52 native speakers of French show that robot-directed speech shares several pitch tendencies with speaker-directed speech, which in its turn is situated between human- and robot-directed speech.

The Effect of Human-Likeliness in French Robot-Directed Speech: A Study of Speech Rate and Fluency

Authors: Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers

Year: 2023

Conference: Ekštein, K., Pártl, F., Konopík, M. (eds) Text, Speech, and Dialogue. TSD 2023.

Abstract: Robot-directed speech refers to speech to a robotic device, ranging from small home smart speakers to full-size humanoid robots. Studies have investigated the phonetic and linguistic properties of this type of speech or the effect of anthropomorphism of the devices on the social aspect of interaction. However, none have investigated the effect of the device's human-likeliness on linguistic re-

alizations. This preliminary study proposes to fill this gap by investigating one phonetic parameter (speech rate) and one linguistic parameter (use of filled pauses) in speech directed at a home speaker *vs* a humanoid robot *vs* a human. The data from 71 native speakers of French indicate that human-directed speech shows longer utterances at a faster speech rate and more filled pauses than speech directed at a home speaker and a robot. Speaker- and robot-directed speech is significantly different from human-directed speech, but not from each other, indicating a unique device-directed type of speech.

Effet de l'anthropomorphisme des machines sur le français adressé aux robots: Étude du débit de parole et de la fluence

Authors: Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers

Year: 2023

Conference: 18e Conférence en Recherche d'Information et Applications – 16e Rencontres Jeunes Chercheurs en RI – 30e Conférence sur le Traitement Automatique des Langues Naturelles – 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues

Abstract: "Robot-directed speech" désigne la parole adressée à un appareil robotique, des petites enceintes domestiques aux robots humanoïdes grandeur-nature. Les études passées ont analysé les propriétés phonétiques et linguistiques de ce type de parole ou encore l'effet de l'anthropomorphisme des appareils sur la sociabilité des interactions, mais l'effet de l'anthropomorphisme sur les réalisations linguistiques n'a encore jamais été exploré. Notre étude propose de combler ce manque avec l'analyse d'un paramètre phonétique (débit de parole) et d'un paramètre linguistique (fréquence des pauses remplies) sur la parole adressée à l'enceinte *vs* au robot humanoïde *vs* à l'humain. Les données de 71 francophones natifs indiquent que les énoncés adressés aux humains sont plus longs, plus rapides et plus dysfluents que ceux adressés à l'enceinte et au robot. La parole adressée à l'enceinte et

au robot est significativement différente de la parole adressée à l'humain, mais pas l'une de l'autre, indiquant l'existence d'un type particulier de la parole adressée aux machines.

Corpus Design for Studying Linguistic Nudges in Human-Computer Spoken Interactions

Authors: Natalia Kalashnikova, Serge Pajak, Fabrice Le Guel, Ioana Vasilescu, Gemma Serrano, and Laurence Devillers

Year: 2022

Conference: Language Resources and Evaluation Conference (LREC 2022)

Abstract: In this paper, we present the methodology of corpus design that will be used to study the comparison of influence between linguistic nudges with positive or negative influences and three conversational agents: robot, smart speaker, and human. We recruited forty-nine participants to form six groups. The conversational agents first asked the participants about their willingness to adopt five ecological habits and invest time and money in ecological problems. The participants were then asked the same questions but preceded by one linguistic nudge, with positive or negative influence. The comparison of standard deviation and mean metrics of differences between these two notes (before the nudge and after) showed that participants were mainly affected by nudges with positive influence, even though several nudges with negative influence decreased the average note. In addition, participants from all groups were willing to spend more money than time on ecological problems. In general, our experiment's early results suggest that a machine agent can influence participants to the same degree as a human agent. A better understanding of the power of influence of different conversational machines and the potential of influence of nudges of different polarities will lead to the development of ethical norms of human-computer interactions.

Detection of Nudges and Measuring of Alignment in Spoken Interactions

Authors: Natalia Kalashnikova

Year: 2021

Conference: Doctoral Consortium of International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)

Abstract: Nudges, techniques that indirectly influence human decision-making, are little studied in spoken interactions. However, the limits of human-computer spoken interactions are not controlled, allowing machines to realize bad nudges. In this context, a framework for detecting nudges is needed to enhance the ethics of HCI. The work proposed in this PhD thesis is based on the hypothesis that detecting nudges lies in measuring linguistic, paralinguistic, and emotional alignments between interlocutors. Therefore, this PhD thesis aims to answer two research questions. First, does a high linguistic and paralinguistic alignment influence a human's potential to be nudged? Second, if a person resists others' emotions, is she or he less sensible to be nudged? To better understand the correlation between alignment and nudges, as well as a human's potential to be nudged knowing their level of alignment, we will conduct a series of experiments.

Chapter 2

State of the art

This chapter introduces the previous research related to the contributions of this thesis. The first part of this chapter focuses on the theory of nudges. We explain the cognitive context of the theory and several frameworks on different cognitive processes and designed for the nudges' classification. We also address the main ethical issues that the nudge theory raises. This part is completed by describing the previous research on linguistic nudges. In the second part, we review the theoretical frameworks of alignment in human-human and human-machine spoken interactions. A description of paralinguistic and lexical characteristics provides the evidence for these frameworks. The third part of the chapter introduces the emotional component of the communicative alignment and is analyzed through context-aware emotion recognition. Thus, we describe the main pre-processing techniques for auditive and textual data, as well as baseline and state-of-the-art systems of context-aware emotion recognition in conversations. The last part of the chapter presents the main datasets used in studies of emotion classification in conversations and communicative alignment.

2.1 Theory of nudges

The theory of nudges comes from the domain of behavioral economics. In 2008, Thaler and Sunstein (2008) highlighted the concept of nudges, defining them as

"any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives. To count as a mere nudge, the intervention must be easy and cheap to avoid."

Nudges come from the idea of libertarian paternalism, which is composed of two concepts. Paternalism is considered in the sense that the intervention aims to benefit from the suggested choice, and the concept of libertarian gives people the freedom to change the suggested option (Thaler and Sunstein, 2008).

Another concept of the nudging theory that needs to be defined is the notion of sludges. There are two visions of the relationship between these two concepts.

Typically, sludges are considered actions that make good (beneficial for users) decisions harder (Sunstein, 2020). This definition suggests that sludges have bad intentions while nudges have good intentions (Thaler, 2018). In other works within this approach, sludge is defined as a "nudge for evil" - a particular kind of nudge that serves bad intentions. Furthermore, researchers of this approach suggest that if the ethical criteria (such as transparency, autonomy of the nudgee, etc.) of a nudge are not included in its definition, it should be called a sludge (Thaler, 2018; Lades and Delaney, 2020).

However, with the growing use of choice architectures based on nudges, Sunstein (2020) offers another definition of sludges. In this second approach, the notion of friction is introduced to distinguish the concepts of nudge and sludge. Thus, a sludge is considered an action that increases frictions to make a decision, while a nudge decreases frictions to make a decision. Consequently, concepts of nudge and sludge can be used for good and bad purposes.

From now on, we will use the term "nudge" to refer to both concepts as a more common and general word to refer to an action that motivates someone to change

their choices without any restriction to this choice. Otherwise, we will specifically apply the word "sludge".

Another aspect that must be clarified is the criteria of "good" and "bad" intentions. For example, if a commercial contract is designed so that the company will benefit from the outcome of nudging action and not the end users, it could be considered a "good" intention from the company's point of view but a "bad" intention from the consumer's point of view. Beggs (2016) addresses this issue by proposing "Pareto" and "rent-seeking" nudges. Pareto nudges allow both the choice architect (the person who designs the environment where the nudge is applied) and the end-user to benefit from the nudging action. Rent-seeking nudges allow only the choice architect to benefit from the nudging action. Another more common terminology proposes to distinguish two kinds of nudges: *"nudge for good"* and *"nudge for evil"* (Sunstein, 2020). *"Nudges for good"* make a choice easier (decrease frictions) for personal or societal advantage, such as an opt-out donation program, which presumes that residents of a country are willing organ donors (Etheredge, 2021). *"Nudges for evil"* serve others' interests. For example, an easy and quick check-out at e-commerce sites (Sunstein, 2020). In the case of sludges, i.e., increasing frictions of an action, Sunstein (2020) illustrates "sludge for good" by the waiting period for firearm licensing and "sludge for evil" by a complicated form-filling.

Thaler and Sunstein (2008) highlighted that the principle of nudges is based on the assumption that humans possess two cognitive "modes" of thinking: "System 1" and "System 2" (Kahneman, 2011). The research of Kahneman (2011) in behavioral economics proposed that System 1 is an unreflective, automatic "mode" that is fast, effortless, and mostly based on cognitive biases and emotions. In contrast, system 2 demands effort for reflection and leads to conscious thinking. Therefore, regarding the "mode" of thinking, we can distinguish two types of interventions: those based on emotions and those based on reflection (Hansen and Jespersen, 2013). System 1 is tightly linked to cognitive biases or heuristics (Kahneman, 2011). However,

another approach (Conway-Smith and West, 2022) suggests that Systems 1 and 2 are the edges of a spectrum of cognitive processes.

2.1.1 Classification of nudges based on cognitive factors

The reviewed studies presented in this subsection proposed frameworks for the classification of nudges. These frameworks are based on the cognitive processes, which make nudges particularly effective in changing human's behavior.

Thus, Dolan et al. (2012) reviewed the nine most robust effects appealing to System 1 that impact our behavior. Their classification is MINDSPACE and stands for **M**essenger, **I**ncentives, **N**orms, **D**efaults, **S**alience, **P**riming, **A**ffect, **C**ommitment, and **E**go.

- **M**essenger: The authors argued that factors describing who sends information impact our behavior. Among these factors, they enumerated the following: whether the messenger is an authority figure, has characteristics similar to ourselves, is an expert, and what feelings we have for them. For example, companies advertise their products through social media influencers as they have large communities that relate to them.

- **I**ncentives: The impact of incentives depends on several factors, such as the reference point (e.g., if the change is perceived as big or small, humans dislike losses more than they like gains of the same amount, etc.), overestimation of unlikely but easy to recall events and preference for smaller but immediate payoffs. For example, a mandatory charge for the plastic bags in stores.

- **N**orms: Humans tend to respect and conform to social and cultural norms to be part of the group. The more a norm is followed, the more people will want to respect it. Vice versa, when people know they are doing better than the norm, their desire for the "good" behavior will decrease. For example,

telling people that they consume more energy than others will reduce their consumption.

- **D**efaults: A default option is proposed to avoid active choice. However, this effect raises many ethical questions, such as the lifetime utility and to what extent the default options should be used, etc. For example, opt-out donor system.

- **S**alience: The authors of MINDSPACE highlighted that information was considered only if it was salient. To be salient, stimuli that attract our attention should be novel (e.g., in flashing lights), easily accessible (products on sale near the cash register), and simple (something that we can understand). For example, big fonts and bright colors.

- **P**riming is any activation (such as words, sights, smells, etc.) of knowledge in memory, making it more accessible. For example, when indicating the likelihood of flossing the teeth in the coming week, participants of the study by Levav and Fitzsimons (2006) significantly increased the frequency of this action during the indicated period. For example, people will eat less if food is served on a smaller plate.

- **A**ffect: Reactions provoked by emotions are one of the most rapid and automatic. The authors of MINDSPACE declare that affect influences even more decision-making than financial incentives. For example, the study of Scott et al. (2007) showed that people washed their hands when provoked to feel disgust rather than when they were presented with the benefits of a soap.

- **C**ommitment: Declaring a commitment increases the likelihood that the action will be fulfilled since breaking the commitment leads to an undesirable image of ourselves. For example, signing a contract or a promise.

- **E**go: This desire for a positive self-image motivates us to compare ourselves

with others. Even if we want to be similar to our groups, our ego wants to be better than others. Moreover, humans consider themselves self-consistent, meaning small changes lead to greater ones without being noticed. For example, telling people that responsible citizens use less energy.

A similar framework that overlaps with the MINDSPACE framework was proposed by Luo et al. (2022). The authors considered three dimensions in the proposed framework. The first one is the type of intervention - whether it reduces the friction of an action (type nudge) or increases it (type sludge). The second one is the direction of intervention - whether it benefits people (nudge/sludge for good) or harms them (nudge/sludge for evil). The third one is the cognitive processes used to motivate behavioral change.

The idea behind the proposed six cognitive processes is the assumption that "cognitive psychology serves as a foundation for decision-making research" (Luo et al. (2022), p.4).

- **Attention**: using the external stimuli to make the desired option more noticeable. The users can be conscious of the stimuli, but it can also be subliminal.
  - *Nudge for good*: Highlighting;
  - *Nudge for evil*: Flashing lights in casinos;
  - *Sludge for good*: Increased font size of calories label;
  - *Sludge for evil*: Reduced font size.
- **Perception**: organizing the information in a way to change the mental representation of it.
  - *Nudge for good*: Availability, assortment size;
  - *Nudge for evil*: Bundle pricing;
  - *Sludge for good*: Smaller portions;

  - *Sludge for evil*: Price partitioning.

- **Memory**: encoding cues to alter the behavior.
  - *Nudge for good*: Anchoring by suggesting an amount of money (e.g., for a donation);
  - *Nudge for evil*: Anchoring by suggesting an amount of money (e.g., maximum deposit);
  - *Sludge for good*: Reminder;
  - *Sludge for evil*: Absence of reminder at the end of the trial periods.

- **Effort**: changing the perception of an effort that needs to be applied for an action.
  - *Nudge for good*: Simplification;
  - *Nudge for evil*: Active choice;
  - *Sludge for good*: Inconvenience;
  - *Sludge for evil*: Complex cancellation processes.

- **Intrinsic motivation**: increasing inherent motivation for the behavior.
  - *Nudge for good*: Goal setting;
  - *Nudge for evil*: Vaping norms for non-smokers;
  - *Sludge for good*: Social norms;
  - *Sludge for evil*: Vaping norms for smokers who want to quit.

- **Extrinsic motivation**: increasing external motivation for the behavior.
  - *Nudge for good*: Financial incentives;
  - *Nudge for evil*: Micro-incentives to gamble;
  - *Sludge for good*: Small fees for no-shows;

– *Sludge for evil*: Membership fees.

The authors also realized a meta-analysis, which highlighted that nudges and sludges had similar effectiveness in the reviewed studies. However, most of the studies concentrated on nudges rather than sludges.

Another meta-analysis realized by Caraban et al. (2019) reviewed multiple studies of nudges to distinguish 23 different ways to influence someone's opinion. These techniques were grouped into 6 categories based on different cognitive biases:

- **Facilitate** (*status-quo bias*) — decrease someone's effort,
- **Confront** (*regret aversion bias*) — create a doubt to encourage a reflective choice,
- **Deceive** (e.g., *decoy effect*, or *peak-end rule*) — affect the perception of alternative choices using deception for usual behavior,
- **Social influence** (e.g., *spotlight effect*, or *herd instinct bias*) — confirm people's desire to correspond to social standards,
- **Fear** (*scarcity bias*) — evoke a sentiment of fear to continue an activity,
- **Reinforce** (*affect heuristic*) — increase the presence of a desired behavior in someone's mind.

We used this framework to categorize nudges used in our study in Chapter 3.

The classes of nudges overlap between classifications, but the reviewed frameworks distinguishing different types of nudges are all based on the weaknesses of human reasoning - systematic patterns of deviation from norm or rationality in judgment (Haselton et al., 2015), i.e., cognitive biases. Recent studies suggest that these psychological mechanisms should be explained to the end users of the actions where nudges are intervened so that they can detect the nudging intervention. Nudges

are often used to prevent users from undesirable actions (e.g., Kostick-Quenet and Gerke (2022)). Still, no system prevents users from nudging intervention.

As mentioned above, more and more choice architects are based on nudging strategies. We presented some conventional examples, but the widespread use of digital tools creates another kind of nudge - digital. "Digital nudging is the use of user-interface design elements to guide people's behavior in digital choice environments" (Weinmann et al., 2016). The authors also highlighted that even though digital nudges occur in digital environments, they are increasingly used to change physical behavior. Digital nudges differ from conventional nudges by two characteristics: personalization and interconnectedness (Bergram et al., 2022). The common use of connected devices and the growing capacity to gather and analyze data allow the choice architect to create dynamically personalized nudges. Moreover, the choice architecture might be constructed in a way where the user knows the choices of others, and in turn, the choice of the user could change the choice architecture for others. Most of the digital nudges are applied in the domain of privacy/security and e-commerce/marketing. The effectiveness of digital nudges was mostly evaluated in online experiments (Bergram et al., 2022).

Digital nudges are at the core of our experiments (described in Chapter 3. As proposed by the framework MINDSPACE (Dolan et al., 2012) of the influence of messenger on the perception of the information, we investigated the influence of the embodiment of an agent.

Nudges and ethics

The main ethical issue of the theory of nudges is who decides what is good and bad for others. The notions of good and bad intentions introduce subjectivity since what is good for someone could be bad for somebody else (Thaler and Sunstein, 2008). Moreover, the choice architects themselves suffer from cognitive and other biases (Rebonato, 2013). In this vein, Panai and Devillers (2023) proposed that multidisciplinary experts should design choice architectures.

As discussed earlier, a common definition of nudges suggests that it is expected to affect end-users positively (in the framework of libertarian paternalism (Thaler and Sunstein, 2008)). However, nudges are mostly designed to influence an average person, so some heterogeneous people can be affected negatively by being nudged (Sunstein, 2013).

Hertwig and Grüne-Yanoff (2017) argue that "obscurantism" is one of the features of nudges. They suggest that the nudge based on status-quo bias (default option) may be effective because end-users fail to understand their choices. Moreover, if the default option does not satisfy somebody, this person should search for other options, which are considered sludges.

Theoretically, nudges leave humans the freedom to choose another option than the one towards which they were nudged, but practically, the intervention does not provide any other options (Meske and Amojo, 2020; Thaler and Sunstein, 2008). Thus, Saghai (2013) proposes that the choice architect should provide options to resist nudges. The possibility to resist occurs when the nudgee knows their behavior is steering towards a certain choice.

Nudges are mostly studied for one-time changes, and the long-term effects are still understudied (Caraban et al., 2019).

Since most of the studies on nudges were conducted in Europe and North America, we still know little about whether cultural peculiarities condition the effectiveness of nudges (especially the digital ones) (Bergram et al., 2022).

Currently, two research teams are working on the standardization of the nudging strategies for the manufacturers in the EU. These norms will make the environment where nudges are applied safer for users and companies (Panai and Devillers, 2023).

2.1.2 Linguistic nudges

We consider linguistic nudges techniques that use linguistic and paralinguistic methods to influence someone's choice. Most of the studies of linguistic nudges are real-

ized using textual modality.

The research of Gohsen et al. (2023) studied how different syntactic (e.g., placing the target information at the end of the utterance) and auditive (e.g., placing the pause before and after the target information) modifications in spoken interactions between a human and a voice-based conversational system nudge participants to ask more questions about specific topics. They found that the auditive nudging techniques made the interaction less natural. As for syntax, nudging techniques were less efficient than direct suggestions to ask about a certain topic, but they were less obtrusive.

Scarcity cues are used in marketing to create nudges that steer users to compulsive purchases. These cues can be supply-based: "Hurry! Only a few seats left" on airline websites (Fenko et al., 2017), or popularity-based: "Over 350 people are currently looking at this!" on e-commerce websites (Teubner and Graul, 2020). Teubner and Graul (2020) studied the effectiveness of scarcity-based nudges in hospitality platforms. They found that both supply-based and popularity-based nudges increased booking intentions, but supply-based nudges were more efficient.

Linguistic nudges are often used in education. For example, OldenBeek et al. (2019) tested how personalized emails providing feedback influenced the learning progress in online courses. The learners received their feedback emails on a weekly and daily basis. The authors measured the effect of nudges on whether the learners viewed videos (extent) and more video minutes (intensity). They reported that nudged learners were 1.5 times more likely to view videos from the online course and spent 15% more time. Moreover, male learners were more susceptible to nudges than female ones. However, the effect of nudge decreased over time.

The medical domain is one of the main domains where linguistic nudges are applied. Sacarny et al. (2018) realized a randomized clinical trial on care prescribers where they sent three letters including nudges based on peer comparison cognitive bias to reduce prescriptions of antipsychotic agents to raise clinical quality. There

was no difference in mortality and hospital use with the control group; the authors declared that nudge resulted in substantial and durable reductions in quetiapine prescription without negative consequences for patients.

Caris et al. (2017) displayed posters with nudging slogans and images at hospital wards' entrances. All nudges increased the use of alcohol-based hand rub. However, the authors did not specify whether there were any differences in the effectiveness of different nudges. Moreover, it is impossible to distinguish the propensity to nudge between image and text.

We hypothesize that a nudge designed as a picture can more spontaneously affect the users than a textual nudge. However, if a reaction to a visual nudge is immediate and does not require a conscious reflection, does it last? We hypothesize that a textual nudge might have more impact in terms of duration since it requires to be processed and analyzed.

With the emergence of the COVID-19 pandemic, a significant interest in linguistic nudges appeared in the medical domain. The studies proposed linguistic nudges to respect new norms of social distance. For example, Ervas et al. (2022) proposed a nudge based on emotions to respect Covid-19 social norms. They presented COVID-19 as a fire metaphor, where the illness was associated with fire, and people were presented as matches that should stay away from the fire to prevent its spreading. The metaphor was presented in verbal messages and visual modes.

Dai et al. (2021) tested different types of text-based reminders that presented vaccination as salient and easy. They designed a basic reminder and a reminder that made participants feel to have ownership of the vaccine. Both reminders were designed with and without video-based information. The study showed that the "ownership reminder" had the greatest effect on the participants. Similar results were observed in the study of Sasaki et al. (2022). The authors analyzed how different types of textual nudges influence people's intention to receive the COVID-19 vaccine regarding different social groups. They found that nudging was efficient if it

gave the impression that the vaccination was voluntary. Moreover, older responders were more susceptible than younger ones.

Regardless of the effectiveness of linguistic nudges in textual modality, they are still understudied in spoken interactions. One of the rare studies that analyzed linguistic nudges in spoken interactions with a teleoperated Android ERICA introduced "persuasion strategies" for Japanese participants. These strategies are close to the strategies of nudges since this study's participants did not have any restrictions or penalties for their choices (Kawano et al., 2022). The authors used 9 techniques, e.g., "actual information," which provided specific information about target tasks. These techniques were studied for three goals: encouraging sports activity, reducing internet consumption, and encouraging charity, with one goal per participant. The authors reported that the persuasion techniques influenced participants. However, the actual change in their behavior or awareness could not be analyzed. They also pointed out that it seemed difficult to predict the persuasion outcome only from linguistic features. They predicted the binary outcome of the persuasion strategies using a Support Vector Machine (SVM) with the following features:

- Personality: results of the Big-five personality test;
- Impression: results of the survey about the robot's impressions;
- Emotion: results of annotation of facial expressions realized during the experiment;
- Action Unit: average and variance of the muscle components of facial expressions;
- Dialogue Act: frequency of labels indicating the type of the utterance: e.g., information seeking.

The best prediction was made based on all features (87.2 % in accuracy). Concerning the Big-Five personality test, it was discovered that the traits of extraversion

and conscientiousness were related to the propensity to be influenced. Successful persuasion was realized when using such strategies as logical persuasion, appealing to credibility, and providing specific information.

These studies focused on the effectiveness of nudges in a specific domain without investigating the relationship between interlocutors. To the best of our knowledge, only the work of Mehenni et al. (2020) addressed the question of the speaker's propensity to influence someone's choice. The preliminary results showed that a robot and a smart-speaker had more impact on children's decisions during a game than a human interlocutor. Nevertheless, the experiment was not replicated with adults in domains where nudges are susceptible to occur and have an impact, such as ecology.

We proposed a methodology of data acquisition from adults and children during spoken interactions in Chapter 3 and analyzed the effectiveness of nudges in spoken interactions in Chapter 4.

2.2 Alignment in dialogs

The term alignment describes linguistic behavior in which interlocutors converge to the use of the same linguistic patterns (e.g., phonetic realizations of repeated words (Pardo, 2006), grammatical structures (Branigan et al., 2000), lexical choices (Garrod and Anderson, 1987)). Pickering and Garrod (2004) declared that alignment on one level leads to alignment on other linguistic levels. Alignment, therefore, plays a crucial role in establishing understanding between interlocutors and their successful communication. However, other studies (Bonin et al., 2013) found the opposite tendencies: alignment did not necessarily manifest at several linguistic levels simultaneously.

Alignment in dialogues can be studied from another perspective, other than developing the same linguistic patterns between two parties. In dialogues where one

of the interlocutors always uses the same linguistic characteristics (e.g., robots or smart-speakers), we can study how someone's speech changes when speaking with conversational agents of a different nature (device *vs* human). We analyzed this aspect in Chapter 5.

The presence of conversational devices in everyday life is growing constantly (e.g., voice control cars, virtual voice assistants to control home automation devices, etc.) The analysis of linguistic differences between human-human interactions and human-machine interactions leads to the development of more naturalistic artificial conversational systems (Branigan et al., 2010).

2.2.1 Human-Likeness

Several studies showed that humans transferred their behavior in human-human interactions into their human-machine interactions (Nass and Moon, 2000), e.g., by reacting in the same manner to machine's behavior as they reacted to human's behavior (Nass and Brave, 2005). Thus, they attribute to machines such characteristics as intentionality (Ju and Takayama, 2009), gender (Nass and Brave, 2005), and ethnicity (Pratt et al., 2007). Moreover, humans attribute "female" behavioral characteristics to computers with a synthesized female voice (Nass and Brave, 2005).

Nass (2004) explained this communicative behavior by the mechanisms of evolutionary psychology. Researchers supported their hypothesis by describing an experiment where participants evaluate a computer's performance after a tutoring session using two computers. The notes of the computer's performance are higher when the evaluation form is filled out on the same computer. According to Nass (2004), the experiment's results represented an example of mindless transfer, since in human-human interactions, it is impolite to tell another person directly that they are not up to our expectations. They observed similar behavior in experiments involving other social situations (Nass and Brave, 2005; Reeves and Nass, 1996; Fogg and Nass, 1997). In a similar study conducted by Aharoni and Fridlund (2007), the

authors compare the verbal and non-verbal measures (smiles, filled pauses, frowns, etc.) between participants believing to be talking to a human and participants believing to be communicating with a computer during a job interview. The nature of the conversational agent influenced the resentment of the job interview's outcome (if they are rejected or accepted for the offer).

Gong (2008) added that the transfer degree is correlated to the degree of the conversational agent's anthropomorphism. In her experiment of choice dilemma, more anthropomorphic agents received more social responses in terms of social judgment (at what point the agent was perceived to be personal, sociable, sensitive, warm), homophily (at what point the agent thought, behaved, was like a participant), social influence, competency (at what point the agent was perceived to be intelligent, informed, competent, experienced, etc.), and trustworthiness (at what point the agent was perceived to be reliable, sincere, respectful, etc.). The functional magnetic resonance imaging study by Krach et al. (2008) confirmed this idea by showing that humans activated the same brain regions that were responsible for taking account of their human partner's intentions when talking to robots. They observed that the intensity of brain activation was associated with the degree of a robot's human likeness.

Furthermore, the degree of accommodation depends on the speakers' beliefs about the interlocutor's capabilities (Branigan et al., 2003). For example, Pearson et al. (2006) conducted a study where they presented two interactive interfaces with the only difference of start-up screen to the participants. In the first condition, the start-up screen presented the interactive interface as a basic, simple version, and in the second condition, the start-up screen described the same interactive interface as an advanced, proficient version. The results showed that participants rated the second version more competent than the first one. In the same vein, native speakers align more with non-native speakers (Bortfeld and Brennan, 1997). Moreover, more anthropomorphized agents seem more capable and intelligent (Cowan et al., 2015).

Agents with more anthropomorphized voices were perceived as more competent and flexible (Nowak and Biocca, 2003). However, Cowan and Branigan (2015) did not find any impact of the kind of the agent on lexical alignment: in their study, participants aligned lexically to the same degree to a human, a computer with a robotic voice, and a computer with an anthropomorphized voice.

Another hypothesis, however, suggests that other more usual interactions shape Human-Computer interactions and communication with children is considered a prototype (Fischer, 2011). This hypothesis lies in the idea that speech addressed to a computer and speech addressed to a child represent examples of simplified linguistic registers (Ferguson, 1982; DePaulo and Coleman, 1986).

Other studies showed that these two hypotheses (supporting the idea of mindless transfer of communication with adults or imitating the infant-directed speech when speaking to machines) cannot always explain human's linguistic behavior: some people do not treat robots as social actors (Fischer, 2011), or people adapt their linguistic behavior regarding the feedback received from the robot (Fischer et al., 2011). Therefore, the question of the human's register when speaking to different kinds of machines needs to be studied (described in Chapter 5).

In the following, we review the most studied paralinguistic and lexical characteristics to describe speech addressed to machines.

Paralinguistic characteristics

One of the first studies was realized by Amalberti et al. (1993). One of the groups was told to speak to a computer, and the subjects of another believed in speaking to a human with the aim of finding out more information about air travel. However, both groups of participants communicated with the same human. Statistically significant differences were found for the mean number of words per dialogue and filled pauses. Participants used more words and fewer fillers when they believed to be talking to a computer. Moreover, the authors found that significant differences occurred at the beginning of the exchange, but both types of speech became similar

over time. In another comparison of human-directed speech and computer-directed speech, Burnham et al. (2010) found no differences in pitch during the listening error task.

During the interactions with a robot, Kriz et al. (2010) showed that first-time users spoke louder, raised their pitch, and hyperarticulated compared to the participants who addressed another human being. In contrast, in a task that consisted in learning new words, Kudera et al. (2023) did not find any significant evidence of a hyperarticulation in the speech of a participant. When comparing robot-directed speech with infant-directed speech, Kriz et al. (2009) found that participants showed more intra-speaker variation in robot-directed speech. The authors observed that participants were aware of the robot's low linguistic competence but relied on its strong information capabilities.

A smart-speaker is often used to describe paralinguistic characteristics of machine-addressed speech. Amazon Alexa was used in the research of Raveh et al. (2019). This study analyzed speech changes at the conversational level in human-human-computer interactions, where an Amazon Alexa device represented the computer interlocutor. The study was realized on the Voice Assistant Conversation Corpus (VACC). The features were analyzed on slices of at most 2 seconds of a single speaker. This research declared that participants had a higher pitch and a higher intensity but almost no differences in speech rate when talking to a smart-speaker than to a human. However, in this study device's voice was set to a default Alexa's female voice, and a human interlocutor was a male. Thus, the results might be influenced by the bias of the device type and the interlocutor's pitch (high-pitched voice for Alexa and low-pitched voice for a human interlocutor).

Cohn and Zellou (2021) proposed a study with a listening error task. They compared phonetic features such as intensity, mean pitch, and range pitch and how these factors change throughout the exchange between human-directed speech and smart-speaker-directed speech (Apple's Siri). The phonetic analysis was done on

recordings first ranked of the perceptual degree of human-likeness. Thus, Siri's utterances were judged much less human-liked than human utterances. Phonetic measures were taken at the sentence level and then analyzed using separate linear mixed effects models. Across different conditions of listening errors, the authors found that smart-speaker-directed speech was characterized by being louder (similar results of intensity were observed by Lunsford et al. (2006); Siegert and Krüger (2021), with a lower mean pitch, and a smaller pitch range (as in the study of Mayo et al. (2012)) comparing with human-directed speech. Interestingly, the pitch range increased over the conversation to approach the levels in human-directed speech (Cohn et al., 2022).

In another study, Cohn et al. (2022) analyzed if the emotional expressiveness of the conversational agent influences the participant's speech. The contrary results were reported: smart-speaker-directed speech was characterized by a slower speech rate, a higher mean pitch, and a greater pitch variation compared to human-directed speech. However, emotional expressiveness did not influence the participant's speech (Cohn and Zellou, 2021).

Lexical characteristics

When communicating with a computer, Brennan (1991) found that parties used fewer pronouns and acknowledgments in the context of the Wizard-of-Oz paradigm. Moreover, when interacting with a computer, speakers adapted their linguistic behavior to a greater degree at the syntactic and lexical levels (Branigan et al., 2003).

In the framework comparing the computer-directed speech and infant-directed speech, Fischer et al. (2011) compared parents' speech when explaining the functionality of simple objects (e.g., how to ring a bell, how to switch on the light) to the children and adults' speech when explaining the same things to the infant avatar. The linguistic analysis contains measures such as verbosity and complexity of utterances. Verbosity is measured by the total number of words in the whole corpus and the number of unique words per speaker. This parameter shows the quantity

of information that participants think they need to share to be understood. It gives us indirect information about how participants perceive their partners. One of the measures to estimate the complexity of utterances is the mean length of utterance (MLU) proposed by Snow (1977). The MLU is calculated by dividing the number of words per speaker by the number of utterances of the same speaker. Researchers report the following results:

1. no significant difference in the number of turns;
2. significantly more unique words in human-machine interactions;
3. significantly longer utterances in human-machine interactions;
4. significantly more frequently use abstract nouns human-machine interactions.

These studies did not specify the approach used for the calculation, e.g., if the lemmatization process was applied.

We evaluated the reviewed measures in our data in Chapter 5.

2.2.2 Engagement

Alignment in spoken interactions is also characterized by the speaker's engagement. Pellet-Rostaing et al. (2023) described several characteristics of engagement that should be taken into account when defining engagement:

- Engagement is considered as a varying process over the conversation that needs a measure to observe changes in its level.
- Engagement is considered as an interdependent process, so we should measure the level of engagement of an interlocutor in the context of the level of engagement of their interlocutors. In that manner, engagement is a bivalent notion since it describes the states of both the interlocutor and their conversational partners.

- Engagement has a property of willingness to contribute efforts in the conversation. Thus, engagement is correlated to the emotional state of the interlocutor.
- Engagement depends on interlocutors' conversational goals.

The authors add that engagement is often associated with the notion of interest. Taking into account previous points, Pellet-Rostaing et al. (2023) proposes the following definition of engagement: *"a state of attentional and emotional investment in contributing to the conversation by processing partner's multimodal behaviors and grounding new information."* Researchers also consider engagement as a mechanism that is closely related to motivation: to the theme of the conversation, and/or the interlocutor (Philp and Duchesne, 2016).

Jacques (1996) described engagement as a complex notion with a set of attributes, such as attention, motivation, etc. He argued that these attributes may take values on the scale from positive to negative, and positive values do not also mean that a user is engaged. Thus, we can distinguish negative engagement, which is not disengagement. Studies (Trowler, 2010; Chipchase et al., 2017) define disengagement as an absence of any engagement and is placed between positive and negative engagement. In this manner, positive engagement is associated with interest and motivation, negative engagement is linked with rejection, and disengagement is manifested by a lack of interest and signs of boredom (O'Brien et al., 2022).

However, in human-machine interactions, engagement is not considered as a mutual process, but as a cue to improve the quality of conversational agents' communication (Pellet-Rostaing et al., 2023). Another characteristic that describes the engagement in human-machine interactions is the subject of the user's engagement, which could be an agent itself, a task of the interaction, or both (Oertel et al., 2011).

Engagement represents the internal state of a speaker, so it is difficult to be annotated. Bonin et al. (2012) studied participants' behavior within a group. Authors supposed that annotating engagement based on annotators' intuition by indicating the moment in a conversation when the level of engagement changed would allow

them to gather more cues to different types of engagement. Among the five labelers, authors found a set of common cues, which were: whether a participant was speaking or not, leaning backward was associated with not-involvement, nodding, and gestures. It was also shown that participants' activity was correlated to the group activity perception. Thus, when some participants were not engaged in the conversation, the global perception of group engagement was considered to be low.

On the contrary, Pellet-Rostaing et al. (2023) proposed to annotate engagement degree at the turn-level at a scale from 1 (strongly disengaged) to 5 (strongly engaged) in human-human interactions. The authors explained their choice by the fact that a speaking turn is a coherent unit from semantic, lexical, and syntactic points of view. Plus, it allowed authors to combine multimodal features with annotation of engagement. In this research, it was hypothesized that engagement is associated with a higher speech rate, pitch, and intensity, a smaller pause after the end of an interlocutor's turn, and a more complex discourse structure. Features were selected according to their correlation with the degree of engagement which was measured by the Pearson correlation coefficient. The selected feature set was used to test the performance of 7 classifiers: Logistic Regression, SVM, K-Nearest neighbors, AdaBoost, Naive Bayes, Random Forest, and Multilayer Perceptron. A combination of different modalities obtained the best results. However, linguistic features (describing a complex discourse structure) did not improve the performance of classifiers. These results highlighted the importance of prosody when expressing engagement.

In human-human interactions, Oertel et al. (2011) found that engagement is described by high pitch and high intensity. Charfuelan et al. (2010) investigated the prosody (pitch, intensity, and voicing rate) and voice quality of dominance in scenarios of professional meetings using Principal Component Analysis (PCA). Participant's involvement was ranked from lowest to highest according to the degree of perceived dominance. The analysis showed that pitch, voicing rate and intensity are higher for the highest level of dominance and lower for the lowest level of dominance.

2.3 Context-aware emotion classification

Other than at the linguistic and paralinguistic levels, alignment in dialogs is also manifested at the emotional level. Human alignment at the emotional level is characterized by its emotional dynamics, which are described by two aspects. The first one is *Emotional inertia*, which is the speaker's emotional state and human's tendency to stick with it regardless of external stimuli. Simultaneously, humans are affected by the emotional states of their interlocutors and tend to align with them. This tendency to mirror partners' emotions can provoke an *emotion shift*. That constitutes the second aspect of emotional dynamics in dialogs, which is *inter-personal dependencies*. This double character of emotional alignment in dialogs constitutes one of the main challenges for emotion classification (Poria et al., 2019b).

Traditional machine learning algorithms were the first approaches used on the multimodal emotion recognition task. However, they were either mainly based on only one modality (Lin and Wei, 2005; Bhavan et al., 2019) or analyzed an utterance using a multimodal approach but regardless of its place in the dialog (Rozgić et al., 2012).

Considering the speaker's emotional inertia and interlocutor stimulation is impossible without considering the "context" of the conversation, i.e., the preceding utterances and their temporality. Previous research had proved that the performance of vanilla emotion recognition approaches (Colnerič and Demšar, 2020; Kratzwald et al., 2018; Hsu et al., 2018; Zhou et al., 2018) which did not consider the context of the conversation was not as high as on prediction of the emotion of separate utterances. Previous studies have also shown that the performance of the emotion classification task was improved if it was based on a multimodal approach (Ghosal et al., 2019; Majumder et al., 2019; Lerch et al., 2024).

The following presents the frequent feature extraction techniques of audio and textual data and deep learning algorithms.

2.3.1 Feature extraction

Most studies cited in this section used three modalities: text, audio, and video to train the models. However, in this PhD thesis, we did not use video for the algorithm, we, therefore, introduced preprocessing methods only for text and audio data.

Text. The most common feature extraction techniques used for deep learning models can be roughly divided into three methods: word embedding technology, word embedding technology followed by a neural network as a feature extractor, and pre-trained language models.

Wang et al. (2016) used Word2vec (Mikolov et al., 2013) dictionary to create word embeddings, which were concatenated to represent an utterance's vector. Poria et al. (2015, 2017) also used Word2vec to create word embeddings, which were used as input for a feature extractor, a 1-D CNN. It extracts local features and combines them into a global feature vector to hierarchically represent a larger unit (e.g., utterance), representing the temporal aspect. Word2vec captured the semantic similarity between words, but it required a lot of data to be trained. Hazarika et al. (2018a) also used a 1-D CNN to extract features obtained with FastText (Bojanowski et al., 2016).

To handle this problem, pre-trained word embedding technology, such as GloVe (Pennington et al., 2014), can be used. Thus, Zadeh et al. (2018a,b); Wang et al. (2019); Liu et al. (2018) used it to create word embeddings in the same way as previously cited Wang et al. (2016) used word2vec. In their other work, Zadeh et al. (2017) used an LSTM-network as a feature extractor to learn time-dependent language representations.

More recent studies (Mai et al., 2021; Lin et al., 2022; Li et al., 2022b; Wu et al., 2022; Macary et al., 2021) used a pre-trained language model BERT (Devlin et al., 2019) (or one of its variants: RoBERTa (Liu et al., 2019), CamemBERT (Martin et al., 2020)) to extract text feature vectors. This method offers rich contextual

word representation. Combining the functions of word embedding creation and contextualizing within it simplifies the training pipeline (Sun et al., 2019). It also allows the grammatical and semantic features of the utterance to be represented (Li et al., 2022a).

BERT is the most common approach in sentence encoding algorithms.

Audio.

Among various methods to extract audio features for emotion recognition, two of them are mainly used in the domain of emotion recognition in conversation: 1) toolkits such as COVAREP Degottex et al. (2014) and openSMILE Eyben et al. (2010), and 2) visual representation of audio with spectrograms and Mel-spectrograms.

COVAREP is a freely available repository that extracts numerous audio features associated with expressing emotions in speech. Studies (e.g., Zadeh et al. (2017, 2018a,b); Tsai et al. (2019)) extracted 12 Mel-frequency cepstral coefficients, pitch tracking, voiced / unvoiced segmenting features, glottal source parameters, peak slope parameters, and maxima dispersion quotients. These features form a vector of size 74 that represents different voice characteristics and is related to emotions (Ghosh et al., 2016).

Like COVAREP, openSMILE is a freely available repository that extracts speech-related features. Moreover, it proposes predefined sets for extracting acoustic features for emotion recognition. Thus, most studies (Hazarika et al., 2018b,a; Poria et al., 2017) used the 2013 ComParE feature set, which contains 6373 features. These features represented many generic acoustic descriptors and their statistical functionals. A large number of features allow this extraction set to be used in many paralinguistics domains, making this set so commonly used. Another commonly used set, GeMAPS (Eyben et al., 2016), proposes a minimalistic parameter set of 62 features related to frequency, energy, and spectrum. Other studies used openSMILE to extract selected features requiring expert knowledge. Thus, Wang et al. (2016) extracted low-level descriptors, such as Mel-frequency cepstral coefficients, pitch,

and voice quality for each utterance. Similarly, Zhang and Chai (2021) extracted 39 features of Mel-frequency cepstral coefficients and pitch, to which they applied Z-normalization.

The success of convolutional neural network (CNN) in image classification motivated to test them on acoustic data with spectrograms. Spectrograms show frequency evolution over time and are useful for capturing changes in tone and pitch associated with emotional state. Another advantage of CNN is its capacity to identify low-level descriptors (e.g., pitch) and high-level features (e.g., intonation patterns). Tursunov et al. (2020) used CNN to extract features from spectrograms and train the model. Issa et al. (2020) proposed using Mel-spectrogram to represent different sound characteristics better and obtain a rich description of an audio.

However, these studies applied CNN to spectrograms without considering the context of the conversation. To handle this problem, CNN can be combined with such methods as Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) neural networks. This issue is discussed in the following subsection.

2.3.2 Deep Learning algorithms

The context-aware emotion recognition approach considers that surrounding utterances influence each utterance and aims to combine their representations of features. Three deep-learning architectures allow the temporal dependencies to be considered: Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and Transformer. Numerous studies propose different architectures using these three blocks, and it seems impossible to list them all. In this subsection, we concentrated on presenting architectures often used as baselines to evaluate new models and the most recent models that achieved state-of-the-art results on context-aware emotion recognition.

LSTM (Hochreiter and Schmidhuber, 1997) is a feed-forward recurrent neural network capable of handling long-distance dependencies. An input gate, an output gate, and a forget gate control the information flow within the network's cells. When

contextual information is necessary to the classification problem, LSTM is used as a context-dependent feature extraction.

One of the first works that used LSTM for context-aware emotion recognition in conversation was the study of Poria et al. (2017). The proposed system applies LSTM to unimodal features to extract context-sensitive unimodal representations of the utterance. Its outputs are then concatenated and fed into bi-LSTM that models long-distance dependencies of utterances in the conversation. The authors noted that using GRU did not improve the model's performance.

GRU (Cho et al., 2014) is a similar mechanism that proposes a similar performance with a simpler computation. It uses reset gate and update gate to handle the contextual information.

The studies of Hazarika et al. (2018b) and Majumder et al. (2019) both used GRU structure and the same features for training. The work of Hazarika et al. (2018b) proposes a model that concatenates the multimodal feature vectors of the 4 previous utterances. This combined vector is then modeled with GRU separately for both speakers. An attention mechanism is applied to extract the most relevant information of the history of 4 utterances. The output of this procedure is then merged with the feature vector of the utterance. The process is repeated multiple times, and the final output is used to classify the emotion category of the utterance.

Compared to Hazarika et al. (2018b), Majumder et al. (2019) proposed a speaker-dependent architecture, which models separately the emotional state of each speaker as well as the global state of the conversation using GRU. The authors proposed several variants of this model, which differed in the use of attention and/or bidirectional variant of GRU. The best performance was observed for the setting, combining both attention and bi-GRU. This variant models the three factors (speaker, context, and emotions) by capturing context from past and future utterances and calculating the attention score.

Contrary to many studies that mainly focused on contextual modeling, a re-

cent work of Li et al. (2022a) concentrated on feature extraction and proposed a new extraction method based on Transformer (without the Decoder structure). The proposed system is called Emocaps. After extracting emotional content from acoustic, visual, and textual modalities, this structure merges it with the sentence vector obtained with BERT. They refer to this structure as an "emotional capsule". These capsules are given to a bi-LSTM layer to produce a contextual representation, which is then used to classify emotions. The model was performed on two datasets: IEMOCAP (Busso et al., 2008) and MELD (Poria et al., 2019a). The average F1 is 71.77 for IEMOCAP and 64.0 for MELD. We tested this system as a baseline on our data (see Chapter 6).

The Transformer modelizes long-distance conversational context without using sequence (Vaswani et al., 2017). However, it is combined with GRU and LSTM.

The reviewed systems were trained on publicly available corpora, such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), etc., which contain acted speech. The performance of approaches trained on acted speech is limited on the real-life data (Ringeval et al., 2014). For example, as was shown by Tahon and Devillers (2010), the realization of anger is different in acted and spontaneous datasets, and its patterns detected in an acted corpus were not found in a corpus of spontaneous speech. Our lab contributes to the creation of more naturalistic datasets, such as the work of Pandey et al. (2014) with a robot, Delaborde et al. (2009) with children, and the datasets presented in this thesis.

2.4 Discussion

In this chapter, we reviewed theories defining the notions of nudge and sludge and their dimension of purpose - whether an intervention serves the interest of the end-user or the creator of the nudge. However, to avoid any misunderstanding between these two approaches, we proposed that in this thesis we use the term "nudge" to

describe any intervention influencing choices. We also presented different techniques based on cognitive processes that make nudges efficient in influencing our choices. We will return to these frameworks in Chapter 3 when classifying nudges created for our data acquisition.

Concerning data acquisition, the literature review on linguistic nudges showed that most linguistic nudges are presented in textual modality. Research on linguistic nudges in spoken interactions is understudied and has several gaps, that we propose to fill in this thesis.

We propose that the detection of linguistic nudges in spoken interactions lies in the study of communicative alignment, which is considered from two aspects: 1) linguistic and paralinguistic, and 2) emotional. We reviewed that speech addressed to machines has its own characteristics compared to speech addressed to a human, but these characteristics are often contradictory regarding the experimental setup. Thus, we believe that the experimental setup with nudging intervention may influence the participants' speech.

We propose to analyze the emotional alignment using context-aware emotion recognition. The best performance in the domain of emotion recognition in conversation is indeed obtained by the systems that consider the emotional context of an utterance.

Thus, in this thesis, we propose the following contributions:

- To the best of our knowledge, only one study (Ali Mehenni, 2023) investigated how the type of the interlocutor influences the effectiveness of nudges in spoken interactions. However, this study was addressed to only one audience - children. Therefore, we propose to enrich the knowledge of this topic with a study applied to another type of audience - adults - to generalize the understanding of nudges in spoken interactions (discussed in Chapter 3).

- Since linguistic and paralinguistic cues describe speech addressed to machines reached opposing conclusions (e.g., smart-speaker-directed speech is charac-

terized by a higher pitch in the study of Raveh et al. (2019), and by a lower pitch in the study of Cohn and Zellou (2021)), we propose to investigate the difference in paralinguistic and lexical features of speech addressed to different types of conversational agents in spoken interactions, adding the aspect of nudging to these interactions (described in Chapter 5).

- Moreover, the linguistic and paralinguistic description of speech allow us to analyze whether the propensity to be nudged influences the speech.

- Apart from the linguistic and paralinguistic alignment, spoken interactions are also characterized by emotional interdependencies and speaker's engagement in conversation. We are inspired by the latest outbreak (Li et al., 2022a) in the domain of emotion recognition in dialogs to propose an architecture based on auditive, textual, and affective cues to predict the outcome of spoken nudging interactions (presented in Chapter 6).

Chapter 3

Data Acquisition and Annotation

This chapter presents the experimental protocol for data acquisition and the annotation strategy. It is divided into two parts: the first introduces the experiment with adult participants, and the second describes the experiment with children. In Sections 3.1.1 and 3.2.1, we discuss the theoretical motivation and the different experimental phases. Then, we describe the experimental procedure in Sections 3.1.2 and 3.2.2. The strategy adopted to annotate the data is provided in Sections 3.1.3 and 3.2.3. Finally, the description of the collected data and the annotation results are presented in Sections 3.1.4 and 3.2.4.

The research described in this chapter focused on two goals:

1. Acquire new data from two audiences - adults and children - who were nudged by different conversational agents - a human / a robot Pepper / a smart-speaker Google Home. The experiment should be framed to measure the changes induced by nudges.

   The experiment with adults was conducted at the Collège des Bernardins, a research center of theology, which also organizes public events in philosophical and cultural domains. We recruited participants from attendees and employees of the research center, as well as visitors to exhibitions that were taking place at this moment. This experiment was realized in collaboration with Col-

lège des Bernardins, LISN lab (Université Paris-Saclay), and a research team of behavioral economists from RITM lab (Université Paris-Saclay). Volunteers from Collège des Bernardins and LISN lab were involved to ensure the recording process.

The experiment with children took place in the outdoor centers of four public schools in the city of Sceaux in the Paris region. This experiment was also realized as a part of a collaboration with a research team of behavioral economists from IRIT lab (Université Paris-Saclay).

The research center's ethics committee of the Université Paris-Saclay approved the experimental procedures for both experiments. Adult participants or parents of children participants signed the consent notice to use their data.

We concentrated on the paralinguistic and linguistic content of the experiment, so the robot Pepper did not provide any gestural or facial expressions. Similarly, a human agent was asked to stay as neutral as possible.

2. Transcribe and annotate recorded data on different affective levels.

3.1 Experiment with adults

3.1.1 Methodology

In Chapter 2 we reviewed different approaches for the notions of nudges and sludges for good and for evil. The intervention techniques proposed by this thesis can be considered both a nudge for evil and a sludge for good, depending on the personal opinion of the nudger and the nudgee. To avoid any misunderstanding between these approaches, we adopt the general term "nudge", defining it as a gentle push toward one particular decision but without any consequences or obstacles.

Using the framework described by Caraban et al. (2019) and reviewed in Chapter 2, we created two groups of nudges: those based on reflection and those based

on emotions.

Nudges based on reflection take a scientifically proven piece of information about one ecological habit and explain its outcomes for the environment. *Nudges based on emotions* speculate on the nudge's sentimental message (e.g., evoking fear or pride).

Within these two groups, we distinguish *nudges with positive influence* and *nudges with negative influence*. The terms "positive" and "negative" do not reflect any aspect of the emotional polarity of nudges, but the nudge's suggested direction. Thus, nudges with positive influence motivate one to adopt an ecological habit by presenting its advantages for the environment or by evoking positive emotions and nudges with negative influence invite one to abandon an ecological habit by showing the negative consequences of an ecological habit or evoking negative emotions.

We illustrate the schema of nudge's classification used in this experiment in Figure 3.1

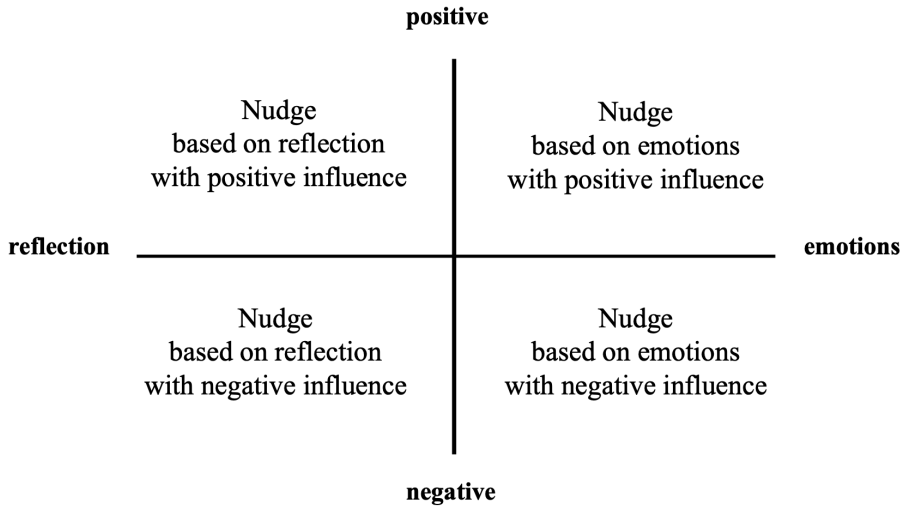


Figure 3.1: Classification of nudges used in the experiment with adults. The vertical axis indicates the nudge's type of influence - towards or against a certain ecological behavior. The horizontal axis indicates the nudge's base - reflection or emotions.

For this experiment, we proposed answering questions about participants' willingness to adopt certain ecological habits. For each of these habits, we created nudges with positive influence and nudges with negative influence, using either techniques based on reflection, or techniques based on emotions. Table 3.1 gives

an overview of the themes and techniques of nudges within the proposed framework. We present the nudges with positive and negative influence in Annex A.

The subject of environment was chosen because of its actuality and controversy. Indeed, we are told to make more efforts to slow down climate change, but most citizens do not know the hidden side effects of new habits. For example, we are encouraged to replace plastic bags with tote bags (shopping bags made of cotton). However, the production of tote bags is very water-intensive. There is no right answer on what choice is better, and we used this controversy to create nudges to initiate the reflection on their everyday ecological choices. We illustrate our approach in the following example of the use of tote bags:

*A **pregnant whale, whose stomach contained 22 kilograms of plastic waste, washed up on the beach** in the Mediterranean. The use of cotton bags reduces the amount of plastic in the oceans.*

The presented nudge is an example of a nudge with positive influence - we steer a user to use cotton bags by evoking negative emotions for using plastic bags. The strategy of nudge used several "units": pregnant whale, stomach full of plastic bags, its death on the beach.

On the other hand, to create a nudge with negative influence for the same theme, we used the strategy of deception (deceive, according to the framework of Caraban et al. (2019)) - use the deception for usual behavior:

*The production of cotton shopping bags is very **water-intensive**. To recoup its production cost, a cloth bag needs to be used **at least 327 times, unlike a plastic bag, which only needs to be used 7 times**.*

The nudge with negative influence is composed of the following "units" that create deception: production is water-intensive and demands many uses to be cost-effective.

Table 3.1 demonstrates nudges with positive and negative influences and the strategies of nudging.

Ecological habit	**Nudge with positive influence**	**Nudge with negative influence**
Use of tote bags vs. use of plastic bags	**Fear**: whales' description with a stomach full of plastic bags	**Deceive**: production of tote bags wastes more water
Self-made cleaning products	**Fear**: fish poisoned with plastic of bottles of cleaning products	**Fear**: no standards applied to home-made cleaning products
Purchase of electric car vs. Purchase of gas car	**Facilitate**: electric car is less expensive for maintenance	**Confront**: use of rare metals for electric cars' production
Travel on a train in France vs. Travel on a plane in France	**Social influence**: eco-conscious citizens take trains	**Deceive**: railways impacts biodiversity
Animal vs. Plant-based proteins	**Confront**: there are more animals for human consumption than the total number of humans	**Deceive**: soja production leads to deforestation & new diseases
Use of electric scooter	Joke about funny accident on a scooter	**Fear**: example of an accident
Green beans cultivated in France vs. Imported green beans	**Social influence**: responsible citizens prefer local products	**Fear**: use of pesticides to cultivate green beans in France

Table 3.1: 7 ecological habits and types of nudges used in the experiment with adults

The main idea behind the scenario was to create a framework that would measure the willingness to follow certain ecological habits before the nudging and how their level of willingness changes after nudges. For that, participants were proposed to answer baseline questions that were formed as follows:

On a scale from 1 to 5, how willing are you (to buy an electric car, make your cleaning products, etc.)?

Participants noted their willingness to adopt the ecological habits from 1 to 5 at this step. After baseline questions, we introduced nudges with positive and negative influences followed by the same questions (as presented above). The difference between scores given to questions with nudges and to baseline questions at the beginning of the experiment indicated if nudging induced differences in participants' willingness to adopt a certain ecological behavior.

The first version of the methodology was tested during the pilot sessions at LISN lab and Collège des Bernardins. In this experiment version, we measured participants' general ecological investment in terms of time, money, and ideas. For example:

How much more money are you willing to pay for environmentally friendly products?

Their baseline score of willingness to adopt ecological habits, on a scale from 1 to 5, was measured during an oral exchange with a conversational agent. To cover the real subject of the study, we asked them general questions about technologies, ecology, etc. We also interviewed them about five ecological habits and not seven: the use of tote bags, the self-made cleaning products, the purchase of an electric car, the travel on train in France, and the consumption of animal proteins. The information presented in the Table 3.1 stayed, nevertheless, the same.

During the step of nudges, we added "quiz" type questions to distract participants from the questions about their willingness to adopt ecological habits and to make the exchange more educational.

At the end of the exchange, an agent asked a participant whether they had learned anything new, and knowing this new information about ecological habits, whether they would spend more time and/or money on ecological problems.

However, we observed several issues with this version of the methodology.

Firstly, participants seemed irritated at the step introducing nudges. They said they had felt repeating themselves and were more concentrated on irritation provoked by the same questions rather than on information presented by nudges.

Secondly, we found it difficult for participants to estimate the time and money they were willing to devote to ecological problems without any situational anchor.

Thirdly, participants were divided into two groups within each conversational agent group: the one who received more nudges with positive influence and the other who received more nudges with negative influence. The order of type of influence

was alternated. However, this approach added more biases to the participants' responses since we could not distinguish between the current nudge's influence and the previous one's influence for analyzing the results.

Finally, the question about the disadvantages of buying green beans produced in France presented as a part of a quiz, drew participants' attention and induced discussion after the experiment.

Considering these observations, we applied the following modifications to produce the version of the methodology that was tested during two main recording sessions:

1. We changed the part of the experiment where we measured participants' willingness to adopt ecological habits to the written survey that participants filled out before the oral part of the experiment.

2. To estimate their general investment in terms of time and money in ecological problems, we transformed these questions into hypothetical situations. For example:

   You have 100 euros to do your grocery shopping for one week in a supermarket. You can also grocery shop at a local market, but it will cost you more. What will you choose?

   To measure how this parameter evolved over the conversation, we reintroduced similar hypothetical situations after the step of nudging to measure if any of the analyzed criteria influenced their level of engagement.

3. The question about the consumption of green beans produced in France was transformed into one of the questions with nudges. We also added a question about using an electric scooter to enrich the number of nudges based on emotions.

4. To limit biases influencing participants' responses, participants were divided into a group receiving nudges with only negative influence and a group receiving nudges with only positive influence.

3.1.2 Experimental procedure

We tested technical equipment at LISN lab in November 2021 and realized a pilot session at Collège des Bernardins in Paris, France in December 2021. We recorded two main sessions at Collège des Bernardins in April and June 2022.

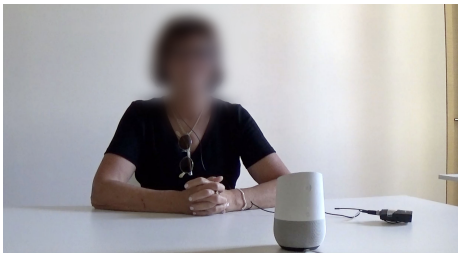
In the first place, we welcomed participants at the entrance of Collège des Bernardins. Our research team explained how the experiment would be held and submitted the consent notice to sign. During a pilot session in December 2021, participants went directly to recording rooms and orally answered questions measuring their willingness to adopt 7 ecological habits in exchange with a conversational agent. However, in the main April and June 2022 sessions, they filled out a written survey measuring their willingness to adopt 7 ecological habits before starting the recording.

In the second place, one of our research team members accompanied them to the recording room, corresponding to one of the three conversational agents. In every room, two team members are present to manage the technical part of the experiment. One was in charge of taking video and audio recordings and taking notes of participants' answers. The second one controlled the Wizard-of-Oz or played human agent.

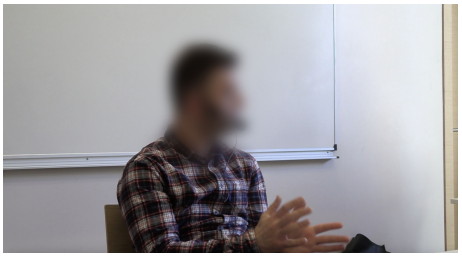
In the two main recording sessions, the agent established common ground with a participant through small talk, e.g., asking the participants about their day, if they were still willing to participate, etc. Afterward, the agent presented two hypothetical situations in which participants chose between the default and the more investing but eco-friendly options. The next step took the same questions as those from the written form of baseline questions preceded by nudges with positive or negative influences for each of the 7 habits. These questions were mixed with quiz-type questions. In the final step, the agent reproduced similar but slightly different hypothetical situations from the beginning of the exchange.

Figure 3.2 demonstrates participants in the three experimental conditions: with

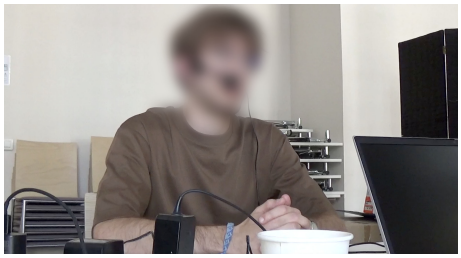
a smart-speaker, a robot, and a human.



(a) Participant with a smart-speaker



(b) Participant with a robot



(c) Participant with a human

Figure 3.2: Captures of adult participants in three experimental conditions: with a smart-speaker, a robot, and a human

In the third place, when the recording was done, experimenters thanked the participants and led them to the organizers' room, where they were offered a snack and filled out the OCEAN personality test.

We sum up the flow of the experiment in Figure 3.3.

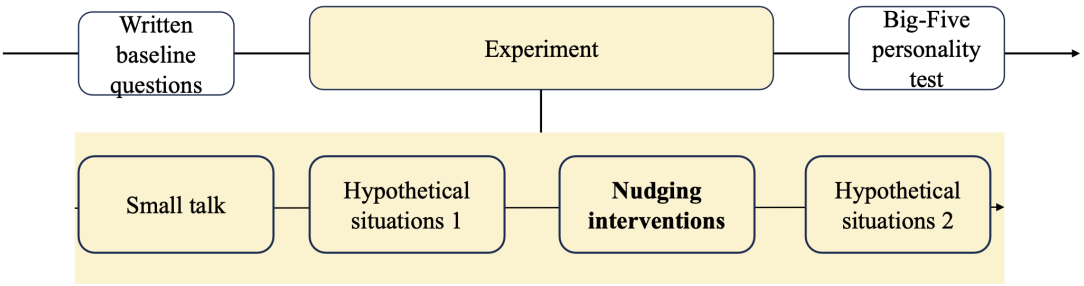


Figure 3.3: Flow of the experiment with adults.

One of the members of our research team played the human agent role. Three female and one male colleagues participated in this role during the two main recording sessions. They read the same script aloud as for the devices' conditions. However, only two participants communicated with the male human agent, and they did not

correspond to our criteria of annotation, they, therefore, were not considered for the further linguistic and paralinguistic analysis of the data. The exchanges with a robot and a smart-speaker were realized using the Wizard-of-Oz paradigm inspired by Mehenni et al. (2020). The voice provided by default settings of the Pepper robot was used for the smart-speaker and robot conditions. The default settings of the Pepper robot provide the synthesized voice with a mean pitch of 230 Hz, corresponding to the pitch of a teenage girl's or high-pitched adult female voice. We used the high-pitched female voice for the device condition to approach the human condition.

We used a unidirectional headset microphone (AKG45) to record audio data using Audacity at 44.1 kHz, 16 bits, and a Sony camera (HDR-CX240E) to record video data. We placed cameras near the conversational agent and focused them on the upper part of the body of the participants. This setup allows us to record the voices of the conversational agent and a subject.

3.1.3 Annotation

One of the focuses of this thesis is the relationship between the propensity to be nudged and emotional states. In particular, we are interested in the differences in emotional states of the "successfully nudged" participants and those who were not nudged. To categorize the "successfully nudged" participants, we concentrated on the subset of data satisfying the following criteria:

- Scores of willingness to adopt a certain ecological behavior were changed for at least two out of seven questions;
- The difference of one of the scores of willingness to adopt a certain ecological behavior is at least two points.

We explain more about these criteria in the following Chapter 4.

The participants who corresponded to both of these two groups were annotated. We also selected data where participants did not change their rates at all or changed their rates to two questions and, at most, for 1 point. In this manner, we can compare and describe the participants' propensity to be nudged.

Three master students from the sociology, literature, and philosophy departments were in charge of the annotation process. The annotators were selected based on their competencies of these domains to provide richer feedback on the data. The two female and one male annotators are French native speakers aged 22 years old. However, due to the lack of time and resources, the data from each participant were annotated by only two of the three labelers.

The recorded sessions were manually segmented and transcribed using ELAN software (Sloetjes and Wittenburg, 2008). The segmentation was realized in two steps. Firstly, the speaker turns of interlocutors were selected. One speaker turn is defined as a segment of speech of one interlocutor realized between two other segments of speech of another interlocutor and starts at the moment of active speaking. Secondly, if the speaker's turn of the participants exceeded 30 seconds, it was cut into several grammatically and semantically cohesive segments and separated by pauses. Pauses were included as a part of a segment when they occurred during the speech of a participant, but they were excluded if they occurred between the turns of the agent and the participant.

After being segmented, speaker turns were orthographically transcribed by one of the annotators. False starts and different types of affect bursts were also annotated. False starts were indicated in parentheses and transcribed as many times as they were repeated (e.g., *"euh je (s-) je sais pas"* Eng.: *"hmm I (d-) I don't know"*). However, there were not enough false starts to study whether they were indicative or not of the speaker being nudged. Affect bursts contained filled pauses (e.g., "euh"), laughs, sighs (if they were signs of emotional states, e.g., irritation), and any sounds indicating hesitation. They were indicated in square brackets. No punctuation mark

was used for transcription.

Annotators listened to the entire conversation between a participant and their interlocutor (human, smart-speaker, or robot) to take the conversational continuity into account and progressively labeled segments that corresponded to the participant's speech. The annotators were trained for the annotation using the video recording of an interaction. However, for the annotation of the datasets, they did not use the video to concentrate on the acoustic emotional expression.

The annotation was done at several affective levels. The choice for the annotation levels was inspired by the existing datasets, annotated on affects, such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019a), RECOLA (Ringeval et al., 2013), etc. which used the same dimensions.

Therefore, our dataset with adults was annotated on the following levels:

- **Valence** was annotated as the acoustically perceived polarity of speech using the labels *positive, negative, or neutral.*

- **Activation** was defined as the intensity of an expressed emotion. This parameter was analyzed on a scale from 1 to 5, where:

  - 1: corresponding to the neutral emotional state;

  - 2: corresponding to the lowest emotional level when it was not considered neutral or the annotator was not sure about the label;

  - 3: the emotion was expressed rather strongly and the annotator was sure about the label used;

  - 4 & 5: a participant was at the highest point when expressing the emotion.

- **Engagement** was considered as a level of a participant's investment in the dialogue. We analyzed engagement on a scale from -2 to 2, where negative values were associated with negative engagement, positive values with positive engagement, and 0 with disengagement. Positive engagement is defined as a

participant's interest and involvement in the communication with a demonstrated desire to cooperate. Negative engagement is considered a participant's interest in the dialog, but with a desire to oppose the agent.

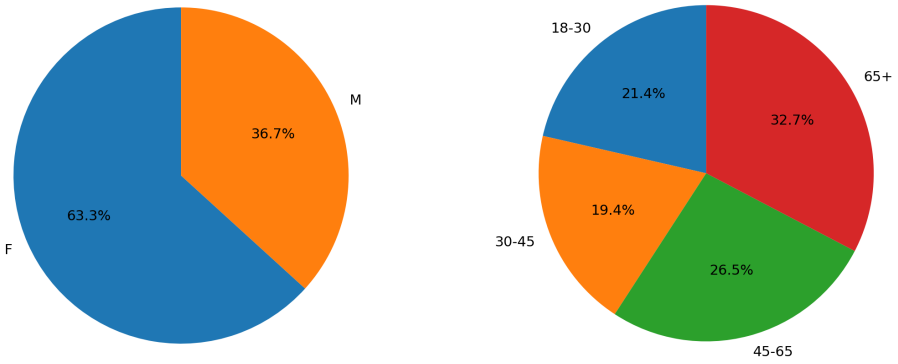
- We adapted the annotation scheme of Vidrascu and Devillers (2005) to define a list of 18 fine-grained **emotion** labels for annotation at a segment level. The fine labels were then merged into 7 macro-classes (fine-grained emotion labels are indicated in italic):
  - Anger: *irritation, aggressivity*;
  - Disgust: *irony, mockery, contempt*;
  - Fear: *embarrassment, anxiety (stress), doubt, reluctance*;
  - Sadness: *lack of interest*;
  - Joy: *interest, amusement, satisfaction, confidence, enthusiasm, relief*;
  - Neutral;
  - Surprise.

  The annotators could use two labels in cases where they doubted between two labels, and/or to describe complex emotions.

3.1.4 Corpus description

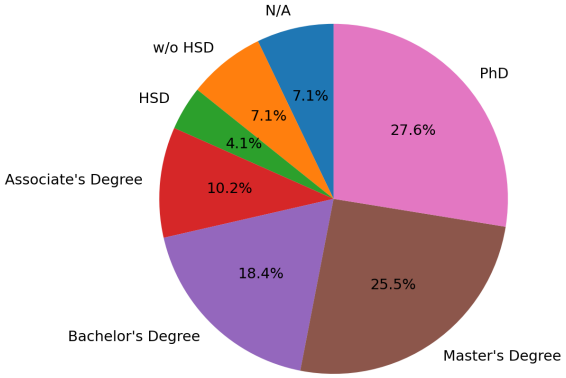
Recorded Data We recorded 98 participants during the two main sessions at Collège des Bernardins in April and June 2022 with a total duration of dialogs of more than 22 hours.

Figure 3.4 shows the diversity of participants in terms of gender, age, and educational level. 62 women and 36 men participated in our study. 58 people were aged 45 and older, and 40 were aged between 18 and 45 years old. Most of our participants possessed a higher education: 27 held Ph.D., 25 had Master's Degree, 18 - Bachelor's Degree, and 10 - Associate's Degree. In line with Hidalgo et al. (2021) we will use



(a) Gender.

(b) Age.

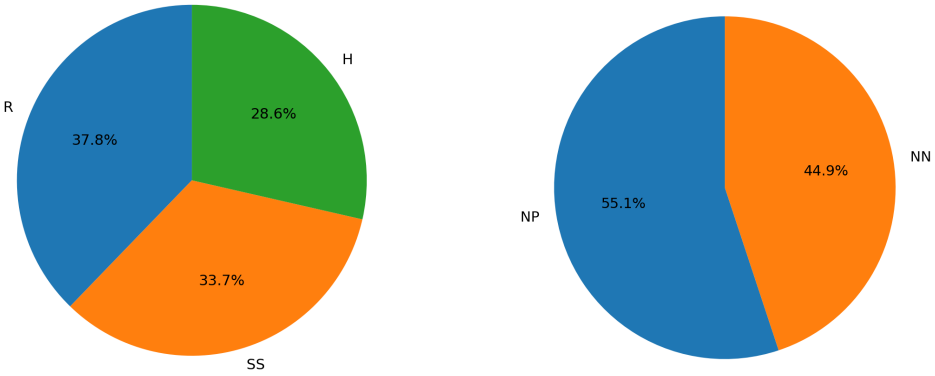


(c) Study level.

Figure 3.4: Distribution of participants in terms of gender, age, and study level. Indications in tables: "F" - female participants, "M" - male participants, "HSD" - high school degree, "w/o HSD" - without high school degree. Dataset with adults

this information to observe differences between these groups. Authors concluded that men were less judgemental than women and made the same choices regardless of their interlocutor (machine or human). They also found that respondents with higher educational levels were less judgemental than those with lower educational levels.

Figure 3.5 represents how recruited participants were distributed in groups of conversational agents and the type of influence that they received. As mentioned before, each group of conversational agent participants was divided into groups of nudges with positive influence and nudges with negative influence. Thus, 54 participants were assigned to a group with positive influence, and 44 to a group with negative influence. Similarly, 37 participants interacted with the robot, 33 with the smart-speaker, and 28 with the human. Table 3.2 represents participants'



(a) Type of agent. (b) Type of influence.

Figure 3.5: Distribution of participants in terms of the type of agent, and the type of influence. Indications in tables: "R" - robot, "SS" - smart-speaker, "H" - human, "NP" - nudge with positive influence, "NN" - nudge with negative influence. Dataset with adults

Type of agent	Positive Influence	Negative Influence
Human	16	12
Smart-speaker	19	14
Robot	19	18

Table 3.2: Distribution of participants per group of conversational agents and types of influence from data from adults

distribution per group of agents and per group of the type of influence. Some lack of balance in these distributions was due to technical problems during the experiment.

Annotated Data

62 participants corresponded to the selection criteria: 1) participants who changed their scores for at least 2 questions and by at least 2 points, 2) and participants who did not change at all their scores. However, we added 12 participants, who changed their scores for two questions, and the difference between scores was less than two points. This choice was made in anticipation of the model's development in case there would not be enough data to generalize. In that way, a total of 74 participants were annotated. After the selection, we annotated 22 participants who exchanged with the human agent, 23 with the smart-speaker agent, and 29 with the robot agent. 35 annotated participants were in a group with positive influence, and 39 were in a group with negative influence. More than 16 hours of exchanges between partic-

Type of agent	Average n° of Tokens	Average n° of Turns	Total n° of Turns	Average Duration	Total Duration
Human	21.16 tokens	51 turns	1121 turns	7.2 secs	8060 secs
Smart-speaker	14.3 tokens	37 turns	847 turns	5.7 secs	4806 secs
Robot	11.7 tokens	34 turns	988 turns	4.3 secs	4295 secs

Table 3.3: The average number of tokens per speaker turn, the average number of speaker turns, the total number of speaker turns, the average duration of a speaker turn, and the total duration of participants' active speech for three conversational agents. Dataset with adults

ipants and three conversational agents were segmented. Almost 5 hours of active participants' speech were transcribed and annotated at affective levels. Table 3.3 describes the annotated data details for each conversational agent's group.

We calculated the Cohen's Kappa coefficient to measure the level of inter-annotator reliability of two annotators (McHugh, 2012). The following results were obtained:

- Fine-grained labels of emotions: 0.66.
- Classes of emotions: 0.67.
- Polarity: 0.29.
- Activation: 0.24.
- Engagement: 0.25.

According to Cohen (1960), a score between 0.21 and 0.40 is considered fair, and a score between 0.61 and 0.80 is considered substantial.

Cohen's Kappa score is influenced by the number of labels used for the annotation. As a reminder, Cohen's Kappa formula is calculated as follows:

$$k = (p_0 - p_e)/(1 - p_e)$$

where p_0 is the empirical probability of agreement on the label assigned to any sample (the observed agreement ratio), and p_e is the expected agreement when both annotators assign labels randomly" (Pedregosa et al., 2011).

When less labels are used for the annotation (which is the case for polarity, activation, and engagement), errors are weighted more than when more labels are used. The low agreement score for these levels illustrates the complexity of emotions. Moreover, annotations were done using only audio recordings.

The distribution of emotional labels of the three annotators is the same. Thus, the order of labels' frequency is the following: "Joy", "Fear", "Sadness", "Anger", "Surprise", "Disgust", and "Neutral", with more than half of the segments annotated with the labels that correspond to the macro-class "Joy".

3.2 Experiment with children

3.2.1 Methodology

The methodology of the experiment with children is based on a dictator game. In this experimental paradigm, a participant receives money and needs to share it between themselves and another anonymous participant. The participant can keep or give all the money (Leder and Schütz, 2018). We adapted this experimental paradigm for children by replacing money with balls. The presented methodology was inspired by the previous experiments realized by Ali Mehenni (2023) during his PhD thesis.

We propose a framework where children play with one of the conversational agents in three games. During the first game, three bowls are in front of the children. Ten little balls are placed in the middle bowl. The conversational agent asks a child to divide ten balls into two empty bowls, one dedicated to keeping balls for themselves and another dedicated to offering balls to other children of the school. The quantity of balls in each bowl is considered the baseline score for a statistical

analysis. After a child distributes balls between the two bowls, a conversational agent nudges the child to change the number of balls in the bowl dedicated to keeping the balls of the child regarding the number of balls that the child took for themselves. If they took less than 5 balls, a conversational agent proposes to take more balls for the child. If they took more than 5 balls, a conversational agent proposes to give more to others. If they took 5 balls, the type of proposition (take more or give more to others) is randomly assigned:

If X $>$ 5: less

otherwise more

If X $=$ 5: randomly between "less" and "more",

where X is the number of balls that the child keeps for themselves.

As was discussed in Chapter 2, nudges use cognitive biases to motivate someone to make a desired choice. In the experiment with children, we tested nudges of the category "Social Influence" according to the classification of Caraban et al. (2019). Nudges were based on two cognitive biases: peer-effect and first person. The nudge of type "peer-effect" compared the child's choice with the common choice of the group of children to which this child belonged and proposed to change the child's choice to the group's choice:

You kept X balls. Usually, other children keep less/more balls for themselves.

The nudge of type "first person" compared the child's choice to the choice of their interlocutor (in our experiment, it is one of the conversational agents) and proposed changing the child's choice to the same choice of the conversational agent.

You kept X balls. If you ask me, I would keep less/more for me.

There were two versions of introducing nudges at this step. During the pilot session, we merged the two nudges into one proposition to change the number of balls.

Agent: *You took 6 balls for yourself. Other children took less. Moreover, if you ask me, I would take less. Do you want to change the number of balls in your*

bowl?

During the other three sessions, we divided nudges into two propositions:

Agent: *You took 6 balls for yourself. Other children took less. Do you want to change the number of balls in your bowl?* *waiting for the answer*

Child: answers the number of balls.

Agent: *If you ask me, I would take less balls for me. Do you want to change the number of balls in your bowl?*

The difference between these two variants of introducing nudges allowed us to compare whether the combination of two nudges is more effective than two nudges presented one by one.

During the second game, we investigated children's attitudes toward their interlocutor. For this game, a conversational agent asked a child if they still wanted to play, and if yes, an agent proposed to a child to choose who could roll the dice: the agent or a human (in groups of a robot and a smart-speaker) / a computer (in a group of a human). The outcome number of the dice corresponded to the number of plastic cords to make bracelets. We analyzed whether 1) the child preferred the interlocutor with whom they started the experiment or the new one; 2) the child preferred to speak to a machine rather than speak to a human.

During the third game, we explored the children's willingness to hide the outcome of the game, depending on the type of conversational agent. Thus, an agent proposed to a child to roll the dice themselves but to hide the outcome and only say it out loud.

At the end of the experiment, an agent wrapped up the child's participation by asking them whether they enjoyed themselves and what game they would like to play another time. After that, a conversational agent joked and asked a child if they were willing to tell a funny story or sing a song.

3.2.2 Experimental procedure

All recording sessions took place at the outdoor centers of four public schools in the city of Sceaux in the Paris region in France between May and June 2023. Children were aged between 6 and 10 years old.

To enroll children for participation, our research team presented the Nao robot to children a few weeks before the recording date. Nao proposed to the children to guess some songs and their artists and to sing something to it, and made some jokes. The parents of those who were interested in participation signed the consent notice. Once we had a list of participants, they were randomly assigned to the groups of conversational agents.

For the recording, one of our research team members accompanied participating children one by one to the experimental room and back to the class after the recording.

During the recording, a child played with one of the conversational agents, as was described previously.

Figure 3.6 demonstrates participants in the three experimental conditions: with a smart-speaker, a robot, and a human.

At the end of the recording day, we distributed balls and elastic cords to the children as promised during the experiment. All children received the same amount of balls and cords regardless of the outcomes of their participation.

We sum up the flow of the experiment in Figure 3.7.

Two female members of our research team played the role of the human agent. They read the same script out loud as the one for machine agents. Robot and smart-speaker conditions were realized using the Wizard-of-Oz paradigm.

The voice parameters for robot and smart-speaker conditions and settings for audio and video recordings were used the same as for the experiment described in Section 3.1.

(a) Participant with a smart-speaker

(b) Participant with a robot

(c) Participant with a human

Figure 3.6: Captures of children participants in three experimental conditions: with a smart-speaker, a robot, and a human

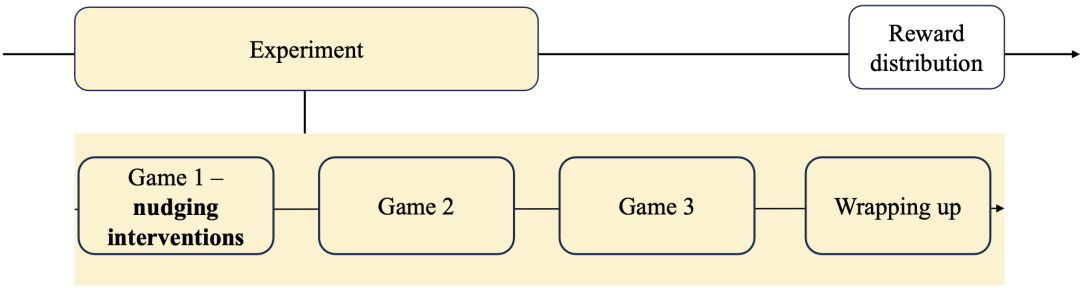


Figure 3.7: Flow of the experiment with children.

3.2.3 Annotation

Segmentation, transcription, and annotation were realized by one male and one female labelers, who previously worked on the corpus described in Section 3.1.

We used the same protocol for segmentation and transcription as in Section 3.1. However, data were analyzed on other levels. Compared to the data collected for the experiment with adults, children spoke less and were more stressed during the recording. Therefore, we hypothesized that the level of stress influenced their comprehension of the experiment and their propensity to be nudged. We also simplified the annotation scheme due to the low level of agreement on the previous corpus.

This corpus was annotated on the following levels:

- **Comprehension** noted if the child correctly answered questions and/or followed the instructions. Labels used:
  - 1: a child understood the instructions or questions questions (e.g.,
    Agent: *What other game would you like to play next time?*
    Child: *I want to play a board game with you.*);
  - 0: the labeler did not manage to identify the level of comprehension;
  - -1: a child did not understand the instructions or questions (e.g.,
    Agent: *What other game would you like to play next time?*
    Child: *Yes.*)
- We used a shortened list of **emotions** which better described children's behavior:
  - stress/embarrassment/fear;
  - impatience/anger;
  - amusement/interest;
  - neutral.

  The annotators could use two labels to describe complex emotional states or if they could not decide between two labels.
- **Engagement** indicated the level of the child's interest in the conversation:
  - 1: a child was interested in the conversation;
  - 0: a child showed no interest in the conversation;
  - -1: a child was interested in conversation but showed signs of irritation by the agent.

Type of agent	Number of participants	Duration
Human	12 (8 girls & 4 boys)	59 mins
Smart-speaker	17 (8 girls & 9 boys)	1 h 28 mins
Robot	15 (8 girls & 7 boys)	1 h 19 mins

Table 3.4: Number of participants and total duration of groups per conversational agent for the pilot session in schools. Dataset with children

- The scale of **Confidence / Hesitation** showed if the child was at ease during the conversation:
  - 1: a child is confident about themselves;
  - 0: the labeler did not manage to identify the level of comfort;
  - -1: a child hesitated a lot and/or did not show confidence in themselves.

3.2.4 Corpus description

Pilot session

24 girls and 20 boys aged from 6 to 10 years old participated in the pilot session. Table 3.4 shows the number of participants for each group and the total duration per group of conversational agents. A total of 226 minutes was recorded during this session.

Main sessions

We report the description of children's distribution in groups according to the following criteria:

- whether they were nudged to increase the number of balls (group "more") or to decrease it (group "less").
- what type of nudge they received in the first place: group "peer-effect" (e.g., *other chose X balls...*) or group "1st person" (e.g., *I would choose X balls...*)

Thus, 67 children out of 86 chose to keep 5 balls for themselves and give 5 balls to others. Due to randomization, 38 children were in the group "more" and 29 in

Type of agent	Number of participants	Duration (mins)
Human	12 (8 girls & 4 boys)	1 h 2 mins
Smart-speaker	36 (17 girls & 19 boys)	2 h 56 mins
Robot	38 (17 girls & 21 boys)	2 h 39 mins

Table 3.5: Number of children participants and total duration of groups per conversational agent for the main sessions in schools

the group "less". In total, 44 were in the group "more" and 42 children were in the group "less". 43 children were in the group "peer effect", including 19 children exchanging with the robot, 18 children communicating with the smart-speaker, and 6 children speaking to the human, the same distribution goes for the group "1st person".

During three main sessions, we recorded 86 children with a total duration of 397 minutes. The data from the four sessions were annotated with the following Cohen's Kappa agreement score:

- Emotions: 0.87
- Comprehension: 0.35
- Engagement: 0.3
- Confidence / Hesitation: 0.27

The agreement score for emotion level is higher than for the adults. However, the level of agreement at other levels is not sufficient using only audio.

The distribution of emotion labels among the annotators is almost the same. The frequency of the labels of the first annotator is the following: "interest", "neutral", "anger", and "stress". The second annotator's most used labels are: "interest", "neutral", "stress", and "anger".

3.3 Discussion

In this chapter, we presented the experimental setups of two data collections: with adults and with children. For the experiment with adults, we introduced the notions of "nudge based on reflection" (presents proven information and expects rational reflection) and "nudge based on emotions" (provokes emotional reaction). Within these two groups, we also presented "nudge with positive influence" (motivates *towards* a habit) and "nudge with negative influence" (motivates *against* a habit). We reviewed themes that were used for the experiment and classified them according to the type of influence (positive (towards) or negative (against)) and the origin of motivation (reflection or emotions). For the experiment with children, we studied the nudge of type "peer-effect" (compares the answer of a participant with "the most frequent answer" of the group) and the nudge of type "1st person" (compares the answer of a participant with the hypothetical answer of the experiment).

For both experiments, we explained the content of the experiment and experimental procedures, as well as our research questions. Both corpora were annotated on different affective levels, even though the agreement score of only the emotional level of annotation is substantial enough for further analysis. We collected and annotated 16 hours of exchanges with adults and 10 hours of dialogs with children.

Parts of the contributions of this chapter were published in Kalashnikova et al. (2022) and Kalashnikova et al. (2024).

We analyze the collected data in the following chapters.

Chapter 4

Nudges and emotions in spoken interactions

This chapter presents two axes of data analysis of the nudging spoken interactions: the effectiveness of nudges and participants' emotional states in spoken interactions. First, we introduce metrics that aim to estimate the effectiveness of nudges and present the statistical analysis. Secondly, we propose the correlation analysis between different criteria (such as the type of interlocutor, the type of nudge, the participant's propensity to be influenced, etc.) and the participants' emotional states.

We presented global research questions in Chapter 1, in this chapter, we detail these research questions. The similarity of methodologies for data acquisition between the experiment with adults and the experiment with children allows us to investigate the following global research questions:

- Do linguistic nudges influence someone's choice?
- Do different audiences (children and adults) have the same propensity to be nudged?
- How does the type of their interlocutor influence these choices?

- Is there any correlation between someone's propensity to be nudged and their emotional state?

- How does emotional state change regarding the type of conversational agent?

These research questions are the same for both experiments. At the same time, the differences in methodologies between these two experiments allow us to ask additional research questions regarding the type of audience. Thus, for the data acquired from adults, we also address the following research questions:

- Is it possible to influence someone's choice against mainstream ideas? We believe that in today's context of massive ecological engagement, nudges with negative influence would be harder for our participants to accept since these ideas go against the mainstream ideas that motivate people to make more efforts to slow down climate change.

- Is there any correlation between the propensity to be nudged and social demographic factors?

- Do nudges based on emotional criteria influence more than nudges based on reflection?

- Do nudges with positive influence have more impact on someone's opinions than nudges with negative influence?

- How has the nudging influenced participant's investment in terms of time and money?

Similarly, for the data acquired with children, we explore the following additional research questions:

- Which nudge is more effective - the nudge based on peer effect or the nudge based on first person?

- Does a child have a preference for the type or their interlocutor?
- If so, do they prefer to speak with a machine or with a human? Or do they prefer to speak to their usual interlocutor?
- When it is possible, do children seize the opportunity to hide?

4.1 Effectiveness of nudges

In this section, we study the effectiveness of nudges in spoken interactions based on multiple factors: the participant's interlocutor, the type of influence, the type of nudge, the audience, etc.

4.1.1 Metrics of effectiveness of nudges

Data from adults

We aimed to measure the effectiveness of nudges from quantitative and qualitative points of view. The quantitative measure focuses on how many ecological habits the nudges were effective in changing participants' scores of willingness to adopt an ecological habit. The qualitative measure aims to show how the participant was willing to change their score of adopting ecological habits.

The difference of one point out of four (the scale from 1 to 5) in scores given before and after the nudging intervention seems too small since the participant could simply forget their baseline score. On the contrary, the difference of three points seems too big since it can only be applied to extreme scores. In that manner, we propose to consider the difference of two points between the score before (baseline) and after the nudge as a threshold for a qualitative measure.

Considering the qualitative measure, we can now analyze the distribution of participants who changed their scores for at least two points for at least one question (Figure 4.1).

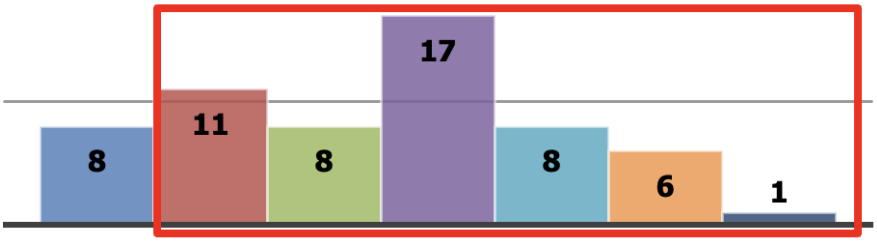


Figure 4.1: Distribution of number of scores changed (per participants who changed their scores for at least one question for at least two points.

We observe that 8 participants changed their scores for one question for at least two points, 11 participants changed their scores for 2 questions for at least two points, 8 participants changed their scores for 3 questions for at least two points, etc. Thus, participants who changed their scores for at least two questions and at least two points represent half of the recorded participants (51 of 98).

Taking this into account, we propose the following measures:

Quantitative measure. A participant is considered to be "nudged" from a quantitative point of view if they changed their scores for at least two answers to questions with nudges compared to scores given to baseline questions.

Qualitative measure. A participant changed their score by at least 2 points between answers to baseline questions and answers to questions with nudges, we considered them "nudged" from a qualitative point of view.

Regarding these criteria, we divided participants into three groups:

- **Nudged**: participants that changed their scores from qualitative and quantitative points of view (i.e., they changed their score for at least two answers AND one of the changed scores was for at least two points);

- **Moderately Nudged**: participants that changed their scores from either qualitative or quantitative points of view (i.e., they changed their scores for at least two questions OR they changed their scores for at least two points to one of the answers);

- **Not-Nudged**: they did not change their scores at all to answers to questions

with nudges.

Among 98 participants, 51 participants belonged to the group "nudged", 36 belonged to the group "moderately nudged" and 11 participants did not change their scores (group "not-nudged").

Data from children

We analyzed the results from the pilot session and the main sessions of the experiment separately since we did not use the same methodology.

Similarly to the work of Ali Mehenni (2023), we considered that a child was "nudged" if they had changed the number of balls after the nudge.

To estimate the influence of the nudge, we proposed the following formula:

$$X * (R_1 - R_O)$$

where $X = 1$ if the nudge suggested the direction "more" and X = -1 if the nudge suggested the direction "less", R_0 indicates the number of balls given by the child to the agent before the nudge, and R_1 stands for the number of balls given by the child after the nudge. In that manner, if the metric is negative, it indicates that the child did not respect the direction of the nudge and the degree to which they changed their scores, and vice versa.

Thus, we calculated these metrics twice: for the scores after the first nudge, and the scores after the second nudge.

We applied the metrics like percentage proposed by Ali Mehenni (2023) to evaluate the effectiveness of nudging from a quantitative point of view (*How many children were influenced?*) and the metrics of effectiveness of nudging described above to evaluate nudges from a qualitative point of view (*How many balls they were willing to keep after the nudges?*).

Methodology for statistical analysis

We used the Wilcoxon signed-rank test to measure if the difference in scores

Green beans	**Meat consumption**	**Tote bags**	**Cleaning products**	**Electric car**	**Electric scooter**	**Travel on train**
6.2e-06	0.32	**0.002**	0.32	**0.0008**	**3.73e-06**	0.62

Table 4.1: p-value per question regarding the type of nudge and the type of conversational agent. Data from adults

was significant between answers to baseline questions and answers to questions with nudges. This statistical hypothesis test analyzes two paired (dependent) data samples. For the data from adults, we measured if the score of willingness to adopt a certain ecological behavior given after questions with nudges significantly differed from the score given to baseline questions. For the data from children, we compared the number of balls given before and after nudging. The test was applied using Python and the module *stats* from the collection SciPy (Virtanen et al., 2020).

To analyze how each social-demographic category and each personal trait of character (measured with OCEAN personality test) influences the propensity to be nudged we applied Chi-squared and Pearson statistical hypothesis tests.

4.1.2 General effectiveness of nudges

Data from adults

We investigated how the theme of the questions influenced participants' propensity to be nudged regardless of the type of their conversational agent and the type of nudge. We hypothesized that the theme on which the nudge is applied is one of the parameters that influence the effectiveness of nudges. We suggested that nudges can be inefficient in questions where participants have already established a habit or the proposed changes demand too much effort.

In Table 4.1, we show p-values for differences between willingness scores given to baseline questions and questions with nudges for 98 participants. We observe that participants significantly changed their scores for four questions: the consumption

of green beans coming from abroad (p=6.2e-06), the use of tote bags (p=0.002), the purchase of an electric car (p=0.0008), and the use of an electric scooter (p=3.73e-06).

We conclude that for other questions (the partial replacement of meat consumption by vegetal proteins, self-made cleaning products, and travel on trains in France), the theme itself could influence participants' propensity to be nudged. When arguing their choices, participants explained that they had already had the habit of choosing the train over the plane when it was possible. Among the reasons they were not interested in making cleaning products, participants cited that it demanded too much time or they had already tried but were not satisfied with the result. Finally, as for the partial replacement of animal proteins, participants either followed one of the alimentary practices of abstaining from meat consumption or decreased the quantity of meat consumption. This question seemed to be one of the most discussed and established in participants' minds.

Data from children

Regardless of the type of nudge, 48% of children changed the number of balls after the first nudge, and 60% of children were nudged after the second nudge.

On average, children kept 5.24 balls before nudging, 4.88 balls after the first nudge (p=0.14), and 4.84 balls after the second nudge (p=0.15). However, the results are not statistically significant.

Among those who changed their scores after the first nudge, we found that 65% changed their scores also after the second nudge. Among those who changed the number of balls after the second nudge, 33% of children were not influenced after the first nudge. Almost 33% of all participants did not change their scores at all.

The comparison of the two audiences showed that nudges induced significant changes in part of the results of adults' data. We observed the tendency of children to decrease the number of balls after the nudging intervention.

Agent	Green beans	Meat consumption	Tote bags	Cleaning products	Electric car	Electric scooter	Travel on train
Human	**0.005**	0.35	**0.04**	0.25	**0.04**	**0.004**	0.16
Smart-speaker	0.34	0.7	**0.05**	0.68	0.19	**0.0005**	0.08
Robot	**0.0002**	0.52	0.07	0.27	**0.004**	0.07	0.4

Table 4.2: p-value per question per group of conversational agents regarding the type of nudge. Data from adults

4.1.3 Influence of agent

We aimed to answer how the type of conversational agent influenced participants' propensity to be nudged regardless of the type of nudge.

Data from adults

Table 4.2 presents p-values for the difference in scores between answers to baseline questions and questions with nudges per group of conversational agents for 98 participants.

Thus, when speaking to a human agent adult participants significantly changed their scores to questions about the consumption of green beans coming from abroad (p=0.005), the use of tote bags (p=0.04), the purchase of an electric car (p=0.04), and the use of an electric scooter (p=0.004).

Participants of a group discussing with a smart-speaker agent significantly changed their scores for the following questions: the use of tote bags (p=0.05) and the use of an electric scooter (p=0.0005).

Participants significantly changed their scores when communicating with a robot agent for questions about the consumption of green beans coming from abroad (p=0.0002) and the purchase of an electric car (p=0.004).

None of the conversational agents significantly impacted participants in questions of partial replacement of meat consumption, self-made cleaning products, and travel by train, as was predicted by the analysis of the theme's impact on participants' propensity to be nudged.

We consider that participants felt more free to discuss the differences in their opinions about ecological habits and presented information during the experiment when speaking to a human than to a device due to the more usual way of communication. We do not deny the possibility that participants could more easily reject proposed ideas when they were expressed by one of the devices, since humans trust more humans than machines, as it was shown by Hidalgo et al. (2021).

Data from children

After the first nudge, 50% of the children were influenced by the robot agent, 52% by the smart-speaker, and 25% by the human agent.

After the second nudge, 71% of the children were influenced by the robot agent, 55% by the smart-speaker, and 41% by the human agent.

In this manner, we conclude that the machine agents impacted more children during the game than the human agent.

Concerning the number of balls, the difference before and after nudging is not statistically significant:

- 1st nudge:
  - Human: t=10.5, p=0.53.
  - Smart-Speaker: t=170.0, p=0.15.
  - Robot: t=208.0, p=0.29.
- 2nd nudge:
  - Human: t=19.0, p=0.59.
  - Smart-Speaker: t=200.0, p=0.28.
  - Robot: t=245.0, p=0.17.

Comparing the results obtained with the data from adults and the data from children, we conclude that the embodiment of the interlocutor influenced both types of audiences.

Type of nudge	**Green beans**	**Meat consumption**	**Tote bags**	**Cleaning products**	**Electric car**	**Electric scooter**	**Travel on train**
Nudge with positive influence	0.14	0.2	**0.002**	0.35	0.07	**0.002**	0.23
Nudge with negative influence	**6.47e-06**	0.7	0.2	0.58	**0.0003**	**0.0007**	0.11

Table 4.3: p-value per question per group of types of nudges regarding the type of conversational agent. Data from adults

4.1.4 Influence of type of nudge

Data from adults

Participants significantly changed their scores of willingness when they were nudged by positive influence for two questions: the use of tote bags (p=0.002) and the use of an electric scooter (p=0.002). As for nudges with negative influence, significant changes occurred for questions of the consumption of green beans coming from abroad (p=6.47e-06), the future purchase of an electric car (p=0.0003), and the use of an electric scooter (p=0.0007).

Nudges with positive influence impacted participants on fewer questions since their baseline rates were already high, and this type of nudge only confirmed their ideas about the desired ecological behavior. Thus, the mean rate of willingness given by all participants to the question about the consumption of green beans coming from abroad was 4.68 points (out of 5). Nudges with negative influence presented new and unexpected information, that allowed participants to see their ecological behavior from a different point of view. We hypothesize that participants felt societal pressure and they gave high rates in a baseline survey to be judged as good citizens, and nudges with negative influence showed them that during the

experiment less popular ecological behavior was also accepted, allowing them to decrease their rates. For example, participants who were in a group of nudges with negative influence gave on average 4.66 points of willingness to the questions of the consumption of green beans coming from abroad. After the nudge, the average rate to the same question was decreased to 3.44 points. We presume that someone can indeed change their opinions against mainstream ideas since nudges with negative influence present ideas that go against mainstream ideas.

As indicated in Table 3.1 nudges with positive influence for questions on the use of tote bags, self-made cleaning products, travel by train, use of an electric scooter, and consumption of green beans coming from abroad and nudges with negative influence for questions on self-made cleaning products, use of an electric scooter, and consumption of green beans coming from abroad are nudges based on emotions. The nudges with negative influence on the questions of travel by train, and the use of tote bags, as well as nudges with two types of influences for questions of partial meat replacement by plant proteins and the purchase of an electric car, are nudges based on reflection. The p-values indicate that only one nudge based on reflection (nudge with negative influence for the question of the purchase of an electric car) had a significant impact on participants' rates. Most significant changes (for nudges with positive influence for questions of the tote bag use, and the use of an electric scooter; for nudges with negative influence for questions of the use of an electric scooter and the purchase of green beans) are observed for nudges based on emotions.

Data from children

The analysis of the proposed metric of nudging showed that, in general, children followed the direction suggested by the nudges. After the first nudge, the metric of nudging was the highest with the robot agent (0.55) and the smart-speaker agent (0.42). The children who communicated with the human agent almost did not change the number of balls (-0.08).

After the second nudge, the tendency stayed the same. Children changed the

Group	Human	Smart-Speaker	Robot	Total
Number of children	12	36	38	86
Nudged 1st nudge	25%	52%	50%	48%
Nudged 2nd nudge	41%	55%	71%	60%
Metric of nudging 1st nudge	-0.08	0.42	0.55	0.4
Metric of nudging 2nd nudge	0.42	0.5	0.79	0.62

Table 4.4: Results of nudging from data from children

number of balls after the nudge with the robot by 0.55 points, the smart-speaker by 0.5 points, and the human by 0.42 points.

We report the results in Table 4.4.

The reported results show how children respected the direction of the nudges (e.g., if the agent suggested to keep more balls, they increased the number of balls that they kept for themselves, and vice versa). However, this metric does not illustrate whether these differences are significant.

The results of Wilcoxon signed-rank test showed no significant differences in the number of balls before and after the two nudges (after the first nudge: t=290.5, p=0.7; after the second nudge: t=352.5, p=0.49) for the children who were suggested to keep more balls for themselves. But despite that, the results for the group "less" (the group who was suggested to keep less balls for themselves) are significant after both nudges (after the first nudge: t=180.5, p=0.02; after the second nudge: t=177.0, p=0.008).

When comparing what type of nudge was more effective, we found that more children changed the number of balls mostly after the nudge "first person" when it was presented in the first place (59% *vs* 41% after the nudge "peer-effect"). However, the difference in number of balls that children chose before and after the nudge is not statistically significant.

- 1st nudge:
  - 1st person: t=287.0, p=0.25.

  - Peer-effect: t=228.5, p=0.34.

- 2nd nudge:

  - 1st person: t=258.5, p=0.1.

  - Peer-effect: t=371.0, p=0.67.

These results conclude that the type of nudging (regarding the cognitive bias of its base) does not play a significant role in the children's decision-making.

Le Guel et al. (forthcoming) realized a joint statistical analysis of the results of our experiment and the experiment realized by Mehenni et al. (2020). The authors found that both types of nudges ("1st-person" or "peer-effect") had a significant effect on children to change the number of balls for the three conversational agents.

Comparison with fused nudges

As a reminder, for the pilot session of this experiment, we presented both nudges together, as we show in the following example:

You took X balls for yourself. Other children took less/more. Moreover, if you ask me, I would take less/more. Would you like to change the number of balls in your bowl?

The size of this subgroup does not allow us to make any conclusions, nor calculate the percentage of the children that changed their scores since the weight of every data sample becomes important. Any of the calculated p-values (comparison of the total data, and the type of agent) is statistically significant. However, we observe the significant p-values when comparing the number of nudges regarding the suggested direction (less/more). Using the metric of the effectiveness of the nudge, we found that children correctly followed the suggested direction of the nudge, and changed more bills when the nudge suggested taking less balls for themselves. We report the discussed results in Table 4.5.

This observation indicates that this axis of the research should be investigated in future research.

Metrics	Group "more"	Group "less"
Measure of efficiency	0.63	1.32
t	17.0	14.0
p	0.02	0.006

Table 4.5: Metrics of nudges' evaluation for groups "more" and "less". Data from children

Our analysis showed the tendency of children to be more impacted by the machine agent than by the human agent.

4.1.5 Correlation between propensity to be nudged and social-demographic categories

Data from adults

One of our research questions was to investigate whether there is a profile of a person who belongs to particular social-demographic groups that are susceptible to manipulation. To that purpose, we analyzed the correlation between social-demographic categories such as age, study level, and gender. Before the experiment, we hypothesized that people with a higher educational level are less susceptible to nudges because their critical thinking is more developed (Ren et al., 2020), women might change their rates more easily due to the developed habit of accommodation (Sabater, 2017), and elder people might trust more easily (Brashier and Schacter, 2020).

Chi-squared statistical hypothesis test for 98 participants did not show any statistically significant correlation between the propensity to be nudged from qualitative and quantitative points of view and neither age nor gender. However, a significant correlation (p=0.02) was observed between a group of participants who changed their rates from a quantitative point of view and their level of higher education. Table 4.6 reports the proportion of participants who changed their rates in groups of educational level. Nevertheless, the size of our sample is not significant enough to make conclusions. Thus, it contains 4 participants with only a high school degree.

w/o HSD	HSD	Associate's	Bachelor's	Master's	PhD
86% (± 26%)	25% (± 42%)	80% (± 25%)	83% (± 17%)	96% (± 7%)	70% (± 17%)

Table 4.6: Proportion and its confidence interval of 95% for the group of participants nudged from a quantitative point of view and their educational level from data from adults

Data from children

Similarly to the adults, age (p=0.1) and gender (p=0.8) did not impact the children's propensity to be nudged:

Contrary to our hypothesis of the correlation between social-demographic factors and the propensity to be nudged, we did not find any statistical evidence. Thus, we may conclude that in our experiments, the propensity to be nudged by both audiences did not depend on their social-demographic factors.

4.1.6 Correlation between propensity to be nudged and character traits

The correlation between the propensity to be nudged and particular character traits was only analyzed using adult data. 89 out of 98 participants completed the OCEAN personality test after the experiment. We analyzed their responses to OCEAN personality test in correlation to the number of answers that they changed during the experiment (quantitative measure) and the number of answers that were changed for at least two points (qualitative measure). The analysis was realized with Pearson correlation.

The only statistically significant correlation (t=0.18, p=0.04) in our experiment was observed between the trait of agreeableness and the number of answers changed for at least two points (qualitative measure). This result indicates that participants with a higher level of agreeableness were more susceptible to changing their answers to a greater degree.

Our finding is contrary to the study of Kawano et al. (2022), who found that the

propensity to be nudged is correlated with extraversion and conscientiousness. This difference raises the question if it is due to the cultural differences between Japanese and French participants, or the particularity of our audience, or whether the three characteristics could be correlated with the propensity to be nudged.

4.1.7 General investment in ecological problems

The participant's general level of investment in ecological problems was measured in terms of time and money. Therefore, these data were only acquired from adult participants.

As a quick reminder, before the step of nudging, we presented to participants two hypothetical situations where they could use the default option or more ecological option but which required more investment of time/money. The situations are briefed below:

Time: *You clean your closet and prepare a bag of clothes that you do not want to wear anymore. You can throw it in the trash can in your building and it will take you 5 minutes. Or, you can go to the recycle bin but it will take more time. What will you choose? If you choose to go to the recycle bin, how much more time are you willing to spend?*

Money: *You have 100 euros to do your grocery shopping for one week in a supermarket. You can also grocery shop at a local market, but it will cost you more. What will you choose? If you choose to go to a market, how much more money are you willing to pay, knowing that your default budget is 100 euros?*

After the step of nudging, we proposed other slightly different hypothetical situations, which are described below:

Time: *After a party with your friends, there is a considerable amount of glass bottles to throw out. You have seen that your neighbors sometimes leave bottles next to garbage cans because the glass bin is quite far from your home. Leaving bottles next to garbage cans will take 5 minutes. Going to the glass bin will take longer.*

Measure	**Before nudging (yes; $\overline{x}$; σ)**	**After nudging (yes; $\overline{x}$; σ)**	p	t
Time	yes=53; $\overline{x}$=22.7;σ=24.7	yes=56; $\overline{x}$=11.8; σ=10.4	1.08e-06	207.5
Money	yes=56; $\overline{x}$=25.3; σ=29.8	yes=52; $\overline{x}$=103.8; σ=106	1.3e-07	198

Table 4.7: Answers of adult participants to the questions if participants were willing to spend more time/money on hypothetical ecological choices. Indications: yes - number of participants who answered yes; $\overline{x}$ - the average of how much time and money participants were willing to spend; σ - standard deviation; p - p-value; and t - t-statistic

What would you do?

Money: *Since you cleaned your closet you want to buy new clothes. You find a coat which costs 80 euros but is not of great quality. You saw another coat, which you like as much, it costs more but is of a greater quality, and will last you longer. What will you choose? If you choose a more expensive coat, how much more money you you willing to pay for it?*

Since the default price for "money" questions differs in situations before and after nudging, we calculated the percentage of additional money compared to the default price and realized our analysis using it.

For the questions of how much time and money participants were willing to spend in hypothetical situations, we analyzed the answers of 62 participants. 36 participants answered in an inoperable manner, e.g., "I don't know", "As much time as it needs", etc. They were, therefore, excluded from the analysis. We analyzed the number of participants who answered "yes" to the questions if they were willing to spend more time/money on hypothetical ecological choices, the average answer, and the standard deviation on how much additional time/money they were willing to spend on these choices. As a reminder, we presented hypothetical situations at the beginning and the end of the experiment, which allowed us to compare the differences in answers before and after nudging. We compared these differences with Wilcoxon signed-rank test which returned p-value and t-statistic. We report the results of the analysis in Table 4.7. We observed that for the first hypothetical situation, 53 participants were willing to spend more time on ecological choice,

spending 22.7 minutes more on average. For the second hypothetical situation, more (56) participants were willing to spend more time on ecological choice, however, the average time was less (11.8 minutes) than for the first situation. The difference in time that participants were willing to spend more on ecological choice is statistically significant (p=1.08e-06) between the first and the second situations.

As for the money, we observed the opposite tendency. More participants (56) were willing to spend more time on ecological choice for the first hypothetical situation than for the second hypothetical situation (52). However, on average, participants were willing to spend 25.3% more money for the first situation and 103.8% more money for the second situation. As for time, the difference in answers for the first and the second situations is statistically significant (p=1.3e-07).

In summary, we noted the following tendencies after nudging:

- more participants were willing to spend time on ecological choices but with less investment than before nudging;
- fewer participants were willing to spend money on ecological choices but with greater investment than before nudging.

We hypothesized that the difference in the investment of money may be due to social status, i.e., the default price that participants are willing to pay for a coat. Thus, the median value for the first situation (additional money spent on grocery shopping) is 20% and the maximum value is 200%, whereas for the second situation (additional money spent on a coat) the median and the maximum values are 68.75% and 500% respectively.

We also analyzed the participants' willingness to spend more time/money regarding the type of conversational agent with whom they communicated and whether they were in the group of positive or negative influence. Table 4.8 reports these results.

In all groups of conversational agents, participants were willing to spend less ad-

Measure	Smart-speaker				Robot				Human			
	$\overline{x}$	σ	p	t	$\overline{x}$	σ	p	t	$\overline{x}$	σ	p	t
Time	17.6; 12.7	17.9 13.5	0.16	55.5	25.4 12.4	31.4 8.2	0.003	41.0	24.0 9.7	18.5 10.0	0.003	5.0
Money	31.3; 56.3	45.0; 52.9	0.04	42.0	19.7; 111.1	11.8; 100.7	5.5e-06	20.5	27.6; 148.0	28.8; 140.5	0.004	15.0

Table 4.8: Willingness to spend more time/money on ecological choices regarding the type of conversational agent from data from adults. The first value in results given on two rows indicates the result for a situation before nudging, and the second value represents the result for a situation after nudging

ditional time after nudging. Participants were mostly impacted in groups of robot and human agents: in these groups, the difference in time investment in ecological choices between answers before and after nudging is statistically significant (p=0.003). Moreover, participants in all groups were willing to spend more money after nudging than before it and in all groups this difference is statistically significant (smart-speaker: p=0.04; robot p=5.5e-06; human p=0.004).

However, we realized an additional t-test to compare the significance of answers between groups of different conversational agents. We found that participants of the group of smart-speaker were willing to spend significantly more money than participants of other groups in both situations. We report the additional results below.

The situation before nudging:

- Smart-speaker *vs* Robot: p=0.03; t=2.2;
- Smart-speaker *vs* Human: p=0.01; t=2.6;

The situation after nudging:

- Smart-speaker *vs* Robot: p=0.03; t=2.2;
- Smart-speaker *vs* Human: p=0.01; t=2.6;

Since these values stay the same after nudging, we hypothesize that it illustrates the particularity of the participants in this group or the influence of the agent.

Measure	NPI				NNI			
	$\overline{x}$	σ	p	t	$\overline{x}$	σ	p	t
Time	22.3; 10.9	28.7; 9.0	0.003	66.0	23.1; 12.7	19.7; 11.9	0.001	45.5
Money	29.8; 114.1	37.8; 108.7	2.47e-05	61.0	20.3; 92.2	15.9; 103.5	0.0005	44.0

Table 4.9: Willingness to spend more time/money on ecological choices regarding the type of conversational agent from data from adults. The first value in results given on two rows indicates the result for a situation before nudging, and the second value represents the result for a situation after nudging. *NPI* - nudge with positive influence; *NNI* - nudge with negative influence

In the same vein, the type of influence did not impact the participants' willingness to invest time/money in ecological choices. In both groups, we obtained the same tendency of participants to spend more money and less time after nudging. Even though, participants of the group of nudges with positive influence seem to spend more money on both hypothetical situations than participants of the group of nudges with negative influence, an additional comparative t-test showed that this difference is not statistically significant.

This result confirmed the result obtained for the pilot session of the experiment with adults (Kalashnikova et al., 2022), where we found that participants were willing to spend more money than time on ecological problems.

4.1.8 Preference for interlocutor in the corpus of children

As a reminder, for the second game of our experiment, we asked children to choose who would roll the dice: his usual interlocutor or a new one, the outcome of the dice indicated the number of plastic cords to make bracelets. If a child was in a group with a human, he had to choose between the human agent and a computer (synthesized voice announced the random number between 1 and 6); a participant from the "machine" group (with a smart-speaker or a robot) had to choose between the machine agent (the smart-speaker or robot) and a human who assisted the experiment (the person responsible for the recording).

Groups of comparison	t	p
Human *vs* Smart-Speaker	-5.31	1.35e-06
Human *vs* Robot	-3.48	0.0009
Robot *vs* Smart-Speaker	-2.07	0.04

Table 4.10: t-test results of comparison of the preference of interlocutor. Data from children

Since this part of the experiment was the same for all sessions, we analyzed all participants together. However, 9 participants did not answer the question in an operable manner, thus we withdrew their data from this analysis.

Group Human. 65% of the participants who communicated with the human agent preferred the computer to roll the dice.

Group Smart-Speaker. Some children from the group communicating with machine agents did not follow the instructions and chose to roll the dice themselves. Thus, 83% of the group who played with the smart-speaker preferred to stay with the smart-speaker, and 8.5% chose a human to roll the dice.

Group Robot. 72.5% of the children who participated with the robot preferred to play with the robot, and 25.5% preferred a human to roll the dice.

We tested whether these observations are statistically significant with t-test for independent samples of scores. We report these results in Table 4.10.

As we can observe, in all comparisons, children chose statistically more often a machine to roll the dice than a human. In the comparison between the groups of the smart-speaker and the robot, the result is still statistically significant with the preference for the smart-speaker. However, our analysis does not allow us to conclude whether the children of the group of smart-speaker preferred the agent itself, or whether they were impacted by the humans who assisted in the room with the smart-speaker. The condition of the smart-speaker was realized by two male members of our research group, whereas the two other conditions were assisted mainly by female members.

Groups of comparison	t	p
Human *vs* Smart-Speaker	-5.31	1.35e-06
Human *vs* Robot	3.1	0.003
Robot *vs* Smart-Speaker	-0.88	0.4

Table 4.11: t-test results of comparison of the propensity to hide regarding the type of interlocutor. Data from children.

4.1.9 Propensity to hide

For the third game of our experiment, we proposed to the children to roll the dice themselves, hide the dice, and just tell the agent the outcome, which corresponded to the number of plastic cords for bracelets that the child would get as a reward. Through this experiment, we aimed to analyze the children's attitudes toward the different conversational agents.

Since it was simpler to check whether the child had hidden for the human agent than for the machines, we hypothesized that children would hide less often the dice with the human agent. However, contrary to our hypothesis, we found that children hid more often with the human agent (87%), less with the smart-speaker (60%), and even less with the robot (51%).

We report the significance scores of comparison between groups of pairs of conversational agents in Table 4.11.

We propose several explanations for this outcome. On the one hand, we hypothesize that children could trust more machine agent than human agent. On the other hand, as the machine agent did not move, children could suppose that it could not check the outcome of the dice, and thus, there was no need to hide the dice.

4.2 Emotions in nudging interactions

In this section, we provide the analysis of emotional labels used by annotators and their evolution during the exchange, the correlation between groups regarding their interlocutor, the type of influence, etc.

As a reminder, annotators could use at least one label and at most two labels for each speaking turn to describe participants' emotional states. Since all labels had the same weight, we proposed to use annotations from both labelers to preserve the richness and complexity of participants' emotional states. In this way, we summed all labels of emotions of each speaking tour for each conversational step to calculate the proportion of emotional labels. We then used a t-test to calculate the statistical significance of different groups.

For the experiment with adults, we used 18 fine-grained labels for the annotation, and for the more detailed description of the participants' emotional states, we used these 18 labels in the analysis presented in this section. These 18 labels are: irritation, aggressivity (macro-class Anger); irony, mockery, contempt (macro-class Disgust); embarrassment, anxiety, doubt (macro-class Fear); lack of interest, reluctance (macro-class Sadness); interest, amusement, satisfaction, confidence, enthusiasm, relief (macro-class Joy); neutral; surprise.

The conversational steps for the experiment with adults consist of:

1. **S0**: establishing common ground with small talk;
2. **S1**: presenting hypothetical situations with the choice between the default and eco-friendly options;
3. **S2**: introducing nudges followed by questions of willingness to adopt ecological habits;
4. **S3**: reproducing similar but slightly different hypothetical situations from step "S1".

For the annotation of the data recorded with children, we used fewer labels to describe their emotional states. We annotated stress, anger, interest, and neutral state. The experiment was divided into 5 steps:

- **Hello**: greetings and small-talk;

- **Game 1**: the game with small balls;
- **Game 2**: the game when they chose the interlocutor who rolled the dice;
- **Game 3**: the game where they hid the outcome of the dice;
- **Bye**: closing small-talk (whether they appreciated the exchange, what they would like to do another time).

However, we did not report the results for the step "Bye" since it did not contain enough data to proceed to the statistical analysis.

We analyzed the described emotion labels for the whole conversation and separate conversational steps for both audiences regarding:

- the type of conversational agent;
- the propensity to be nudged;
- the type of nudge.

In this section, we report only the statistically significant results in tables.

4.2.1 Emotional state and conversational agent

Data from adults

Emotion	**Intro**		**S0**		**S1**		**S2**		**S3**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-1.55	0.13	-0.26	0.8	0.42	0.68	3.88	**0.0002**	3.81	**0.0002**
Amusement	2.2	**0.04**	4.71	**2.1e-05**	4.21	**9.6e-05**	2.87	**0.005**	2.48	**0.01**
Lack of interest	-1.42	0.17	-2.32	**0.02**	-1.84	0.07	-4.8	**9.1e-06**	-5.4	**1.13e-06**
Irritation	-1.03	0.3	-0.009	0.99	-1.79	0.07	-5.12	**1.84e-06**	-4.5	**2.82e-05**

Table 4.12: Significant test statistics comparing labels of emotions between a group addressing a human and a group addressing a smart-speaker. Data from adults.

Human vs Smart-Speaker.

When comparing the emotional labels of the group speaking to a human and the group speaking to a smart-speaker, we observe that the beginning is characterized by

significant differences in labels of amusement ("Intro": t=2.2, p=0.04; "S0": t=4.7, p=2.1e-05) and anxiety ("Intro": t=-1.95, p=0.05; "S0": t=-2.43, p=0.02). These results indicate that at the beginning of the conversation participants speaking to a smart-speaker felt more anxiety and participants speaking to a human were amused.

The statistical comparison between participants speaking to a human agent and participants speaking to a smart-speaker shows significant differences in the labels "interest" (t=3.88, p=0.0002; t=3.81, p=0.0002) and "lack of interest" (t=-4.8, p=9.1e-06; t=-5.4, p=1.13e-06) at the step of nudges ("S2") and hypothetical situations ("S3") indicating that when speaking to a smart-speaker participants are less interested in conversation and drop out at the second part of the exchange. At the same steps, speech addressed to a smart-speaker has a significantly higher score of the label of "irritation" than speech addressed to a human at steps "S2" and "S3" (t=-5.12, p=1.84e-06; t=-4.5, p=2.82e-05). Throughout all conversational steps, participants have a significantly higher level of amusement when speaking to a human (t=2.2, p=0.04; t=4.71, p=2.1e-05; t=4.21, p=9.6e-05; t=2.87, p=0.005; t=2.48, p=0.01). Table 4.12 reports results for these four emotions for each conversational step.

Globally for the whole conversation significant differences were observed for the following labels: interest (t=4, p=0.0001), reluctance (t=2.37, p=0.02), amusement (t=3.7, p=0.0004), and lack of interest (t=-5.07, p=3.36e-06). These results describe participants speaking to a human as having more interest and being more amused and at the same time with a high level of reluctance. Participants speaking to a smart-speaker are characterized by a lack of interest.

Human vs Robot.

The comparison between the group speaking to a robot and the group speaking to a human shows that at the beginning of the conversation, participants were significantly more interested when they spoke to a robot than to a human (t=-2.71, p=0.008; t=-2.69, p=0.008). However, this difference is no more significant at step

Emotion	**Intro**		**S0**		**S1**		**S2**		**S3**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-2.71	**0.008**	-2.69	**0.008**	-1.92	0.06	-0.28	0.77	1.17	0.24
Amusement	1.27	0.2	2.25	**0.03**	2.94	**0.004**	2.88	**0.005**	2.91	**0.005**
Hesitation	-1.76	0.08	-1.76	0.08	-3.03	**0.003**	-3.15	**0.002**	-2.85	**0.006**
Lack of interest	NA	NA	-0.5	0.6	-1.93	**0.05**	-3.32	**0.001**	-5.06	**3.15e-06**

Table 4.13: Significant test statistics comparing labels of emotions between a group addressing a human and a group addressing a robot. Data from adults.

"S1" till the end of the conversation. At the same time, the use of the label "lack of interest" became significant at step "S1" till the end of the conversation (t=-1.93, p=0.05; t=-3.32, p=0.001; t=-5.06, p=3.15e-06). We interpret these contradictory results by the fact that participants of both groups lost interest in the exchange after the first two steps of the conversation.

During the conversation, participants were significantly more amused with a human agent ("S0": t=2.25, p=0.03; "S1": t=2.94, p=0.004; "S2": t=2.88, p=0.005; "S3": t=2.91, p=0.005), and they hesitated more with the robot ("S1": t=-3.01, p=0.003; "S2": t=-3.32, p=0.001; "S3": t=-5.06, p=0.006). For more details see Table 4.13.

When analyzing the whole conversation, we observed statistically significant differences for the following emotional labels: reluctance (t=4.43, p=4e-05), amusement (t=3.07, p=0.003), hesitation (t=-3.3, p=0.002), irony (t=-2.17, p=0.04), lack of interest (t=-3.83, p=0.0003), relief (t=2.59, p=0.01). These results show that participants speaking to a human were annotated as expressing more reluctance, amusement, and relief. Participants exchanging with a robot were considered expressing more hesitation, irony, and lack of interest.

Emotion	**Intro**		**S0**		**S1**		**S2**		**S3**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	-0.69	0.5	-2.56	**0.01**	-2.34	**0.02**	-3.62	**0.0005**	-1.93	**0.05**
Hesitation	-1.75	0.08	-1.76	0.08	-3.03	**0.003**	-3.15	**0.003**	-2.85	**0.006**
Irritation	2.25	**0.03**	2.11	**0.04**	0.58	0.6	3.74	**0.0003**	2.62	**0.01**

Table 4.14: Significant test statistics comparing labels of emotions between a group addressing a smart-speaker and a group addressing a robot. Data from adults

Smart-Speaker vs Robot.

The significant differences observed at the beginning of the exchange between the group speaking to a robot and the group speaking to a smart-speaker are the labels "irritation" ("Intro": t=2.25, p=0.03, "S0": t=2.11, p=0.04), "amusement" ("S0": t=-3.04, p=0.003), "confidence" ("Intro": t=-2.07, p=0.04), indicating that when speaking to a robot participants were perceived as feeling amused and confident, whereas when speaking to a smart-speaker they were considered as irritated.

As for the other steps of the conversation, *t-statistic* indicates that participants were significantly more interested in speaking to a robot ("S0": t=-2.56, p=0.01; "S1": t=-2.34, p=0.02; "S2": t=-3.62, p=0.0005; "S3": t=-1.93, p=0.5) than to a smart-speaker, and also significantly hesitated more ("S1": t=-3.03, p=0.003; "S2": t=-3.15, p=0.003; "S3": t=-2.85, p=0.006). However, similar to the comparison with the group speaking to a human, participants were significantly more irritated when speaking to a smart-speaker almost throughout the entire conversation. Table 4.14 presents test statistics for this comparison.

The analysis of the whole conversation showed significant differences in the following labels: reluctance(t=2.7, p=0.008), hesitation (t=-3.3, p=0.002), aggressivity (t=2.71, p=0.008), irritation (t=4.11, p=8.41e-05). Thus, we conclude that when speaking to a robot participants showed more hesitation, whereas when speaking to a smart-speaker participants showed more reluctance, aggressivity, and irritation.

Data from children

The tendency that we observed in children's emotional states regarding their interlocutor was that the emotions of children who spoke to the human agent were different from the emotions of children who spoke to machines regardless of if it was a robot or a smart-speaker. Thus, we found significant differences between groups of the human and the smart-speaker, and of the human and the robot for the label "stress". These differences were statistically significant for the whole conversation but also for each step analyzed separately for both pairs of comparison, indicating

that children were more stressed to speak to the human agent than to the machine. Moreover, when comparing the labels of the group speaking to the robot and the group speaking to the smart-speaker we found a significant difference only for one conversational step which corresponded to the game of little balls (the step where we introduced nudges). Thus, children of the group robot were significantly more stressed and less interested at this step compared to the group of children speaking with the smart-speaker. Nonetheless, the differences were not significant for either the whole conversation or other conversational steps.

However, this result may be explained since children were told that they were going to speak to "robot" (machine), and their frustration with this incoherence was perceived as stress. This hypothesis was supported by the statistically significant higher level of the label "interest" for describing the emotional states of children interacting with the smart-speaker compared to the children communicating with the human.

We report the presented results in the Table 4.15.

Emotion	**Hello**		**Game 1**		**Game 2**		**Game 3**		**Global**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Human *vs* Smart-Speaker										
Stress	2.42	0.02	4.4	2.3e-05	2.6	0.01	3.05	0.003	4.29	3.66e-05
Interest	-2.12	0.04	-2.33	0.02	-0.99	0.32	-0.66	0.5	-2.46	0.02
Human *vs* Robot										
Stress	2.27	0.03	2.39	0.02	1.97	0.05	2.0	0.04	2.75	0.007
Smart-Speaker *vs* Robot										
Stress	-0.11	0.91	-2.19	0.03	-0.52	0.6	-1.2	0.23	-1.72	0.09
Interest	0.77	0.44	2.0	0.04	-0.42	0.67	0.3	0.76	0.9	0.37

Table 4.15: Significant test statistics comparing emotional labels between pairs of groups of conversational agents. Data from children.

4.2.2 Emotional state and propensity to be nudged

Data from adults

Table 4.16 reports comparative test statistics of most frequent labels between participants who were influenced by nudges (group "nudged") and participants who

did not change their rates on their willingness to adopt ecological habits (group "not-nudged").

t-statistic has positive values during all conversation steps for "interest", indicating that the group of "nudged" participants was more interested than the group of "not-nudged" participants. The difference in "interest" becomes significant at step "S2" (p=0.01) which corresponds to the step where we introduced nudges, and stays significant (p=0.003) for the next step of hypothetical situations with the default and eco-friendly choices.

Contrary to interest, *t-statistic* is negative for the label "confidence", showing that "nudged" participants were significantly less confident ("Intro": p=0.02; "S0": p=0.004, "S1": p=0.02). However, in the last two steps, the differences decrease and are not significant anymore.

For the label of "embarrassment" *t-statistic* is negative at the beginning of the exchange and positive at the end, but the differences are not significant. This observation indicates that the "not-nudged" group felt more embarrassed at the beginning, while the "nudged" group felt significantly more embarrassed ("S3": p=0.04) at the end of the conversation.

Among other statistically significant results, we report that for step "S0" which corresponds to the small talk between the participant and the agent "nudged" participants were more amused (t=3.32, p=0.001).

When regarding the results of statistical tests for the whole conversation, we observe significant differences for the following emotional states between group "nudged" and group "not-nudged": interest (t=2.9, p=0.005) and irony (t=-2.32, p=0.03). These results indicate that the group "nudged" was globally more interested, whereas the group "not-nudged" was globally more ironic.

Data from children

To continue the work realized by Ali Mehenni (2023) we considered that children were nudged if they changed the number of balls that they kept for themselves after

Emotion	**Intro**		**S0**		**S1**		**S2**		**S3**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Interest	0.22	0.83	1.22	0.23	0.42	0.68	2.6	**0.01**	3.06	**0.003**
Confidence	-2.43	**0.025**	-3.04	**0.004**	-2.37	**0.02**	-1.43	0.16	-0.44	0.66
Embarrassment	-0.35	0.73	-0.52	0.6	1.65	0.1	0.76	0.45	2.09	**0.04**

Table 4.16: Significant test statistics comparing labels of interest, confidence, and embarrassment between group "nudged" and group "not-nudged". Data from adults.

at least one nudge. In that manner, we grouped the children who changed their scores (once or twice) into the category "nudged" and those who did not change their scores at all (after the first or the second nudge) were grouped into the category "not-nudged".

We observed statistically significant differences between these two groups for the labels "neutral" and "anger". We report these results in Table 4.17.

Emotion	**Hello**		**Game 1**		**Game 2**		**Game 3**		**Global**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Anger	9.32	6.36e-18	12.1	1.47e-26	8.35	1.93e-14	7.97	6.1e-12	13.14	9.05e-30
Neutral	-2.9	0.006	-4.8	1.38e-05	-4.31	9.66e-05	-6.22	8.42e-08	-6.87	5.31e-09

Table 4.17: Significant test statistics comparing labels of anger and neutral between group "nudged" and group "not-nudged". Data from children.

We observed that children who changed their scores were annotated more often with the label "anger" for the whole conversation and also for each step. On the other hand, children who did not change their scores were perceived more as neutral compared to the group who changed their scores.

Another significant difference was observed for the label "stress" at the step of the third game: t=2.2, p=0.03, which indicates that the group "nudged" was perceived as more stressed than the group "not-nudged" at the end of the conversation.

However, no statistically significant differences were observed for the label "interest".

4.2.3 Emotional state and type of nudge

Data from adults

For the analysis of the correlation between the type of influence and the emotional state of the participants, we were particularly interested in the following points:

- whether there is any difference in emotional state of participants at the beginning of the exchange;
- how the emotional state changed at step "S2" which corresponded to the step of nudges;
- whether there are any consequences on the emotional state of the participant after the step of nudges.

Thus, we found that at the beginning of the exchange participants of the group of nudges with positive influence were annotated more frequently with the label "interest" ("Intro": t=2.76, p=0.007; "S0": t=1.97, p=0.05) and "confidence" ("Intro": t=2.06, p=0.04). On the other hand, participants in the group of nudges with negative influence were more frequently annotated with the label "aggressivity" ("S0": t=-2.08, p=0.04).

Most statistically significant changes were observed at step "S2" as expected. Participants of the group of nudges with negative influence were mostly described by the following labels: "lack of interest" ("S2": t=-2.27, p=0.02), "surprise" ("S2": t=-2.12, p=0.03), and "relief" ("S2": t=-2.01, p=0.05). Whereas the label "irony" ("S2": t=2.15, p=0.03) differed significantly for the participants of the group of nudges with positive influence.

After the step of nudges, two labels had significant differences: "amusement" ("S3": t=2.02, p=0.04) and "contempt" ("S3": t=2.01, p=0.05), characterizing participants of the group of nudges with positive influence as being more amused and contemptuous.

The end of the conversation is characterized by frequent use of the label "interest" (t=2.7, p=0.009) for the participants of the group of nudges with positive influence, and of the label "embarrassment" (t=-2.35, p=0.02) for the participants of the group of nudges with negative influence.

These results allow us to conclude that nudges with different types of influence impacted the emotional state of participants. Nudges with negative influence incited a lack of interest, but also surprise and relief.

Data from children

As for the result of the effectiveness of the nudge, the type of nudge presented in the first place (peer-effect or first person) did not impact the emotional state of participants. We found no significant differences in the labels that were used for these two groups.

Emotional state and direction of nudge.

Finally, we analyzed the difference in emotional labels between the group who was suggested to increase the number of balls that they kept for themselves (group "more") and the group who was suggested to decrease the number of balls that they kept for themselves (group "less").

Significant differences were mainly observed for labels "anger" and "neutral", as we report in Table 4.18. For the whole conversation and at each conversational step participants of the group "less" were annotated more often with the label "anger" and less with the label "neutral" than the group "more". We hypothesized that children changed their scores due to the expectations that they imagined the experiment's observers had of them.

Moreover, we found that for the whole conversation, the group "less" was stressed less than the group "more" (t=-2.17, p=0.03).

Emotion	**Hello**		**Game 1**		**Game 2**		**Game 3**		**Global**	
T-statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
Anger	-8.8	2.06e-16	-11.6	1.0e-24	-9.9	6.4e-20	-10.9	3.04e-21	-14.3	3.24e-34
Neutral	3.8	0.0003	3.8	0.0003	5.7	2.54e-07	6.6	6.43e-09	6.9	1.23e-09

Table 4.18: Significant test statistics comparing emotional labels between pairs of groups "more" and "less". Data from children.

4.3 Discussion

This chapter introduced the metrics to evaluate whether the nudges impacted our participants. For the experiment with adults, we proposed to consider that the participant was "nudged" from the quantitative point of view if they had changed their scores for at least 2 questions and from the qualitative point of view if they had changed for at least two points one of their scores. For the experiment with children, we proposed a metric that estimated the degree to which the child respected or did not the suggested direction of the nudges.

In the experiment with adults, we found that participants changed their scores on more questions with the human agent. Our results showed that nudges with negative influence had more impact on participants' scores. However, the baseline scores were already high, so the impact of the nudges with positive influence was limited. Nevertheless, the effectiveness of nudges with negative influence illustrated that participants changed their scores against mainstream ideas of ecological investment.

Even though we found that nudges based on emotions incited greater participant score changes than nudges based on reflection, nudges based on reflection were still effective. This finding confirms the results of the experiment by Kawano et al. (2022).

We did not find any correlations between the propensity to be nudged and social-demographic categories. However, we observed that participants with a higher level of agreeableness changed their scores on more questions. The analysis of how nudges influenced participants' willingness to invest time and money in ecological choices showed that participants were willing to spend less time but more money after

the nudge. However, we could not estimate the influence of social status on these answers.

Adults were annotated when speaking to the human as being amused, interested, and relieved. The group communicated with the robot, who was initially perceived as interested, and expressed a lack of interest shortly after the beginning, with hesitation and irony throughout the conversation. Finally, participants who exchanged with the smart-speaker were mostly described with the labels "anxiety", "lack of interest", and "irritation".

Those participants who were nudged according to our measures were annotated as less confident and expressing interest, and those who were not nudged were described as ironic.

Contrary to our hypothesis, children were not as sensible to nudges as adults during our experiment. We observed the children's tendency to be more impacted by the machine agent than by the human. This tendency confirmed the first findings described by Ali Mehenni (2023). However, a joint statistical analysis of both experiments with children showed that children were efficiently nudged by the three conversational agents (Le Guel et al., forthcoming).

In comparison with nudges presented separately, fused nudges impacted more. For other research questions about children's behavior with different conversational agents, we found that they had hidden more with the human and preferred to play with the machine, especially the smart-speaker.

On the one hand, the statistical analysis of the number of balls before and after nudges showed that children were willing to give more balls to others, indicating a high level of altruism. On the other hand, the analysis of emotional labels used to annotate emotional states showed that children who were suggested to give more balls to others expressed anger. We hypothesized that the high level of altruism was due to the expectations that children thought others might have of them.

The annotation of children's emotional states showed that they distinguished

between machine and human agents. Thus, children who communicated with the human were perceived as more stressed and less interested compared to two other conversational agents, and no differences were observed between the emotional states of children who communicated with the smart-speaker and those who communicated with the robot.

Nudged children, as well as children who were suggested to take less balls for themselves, had a higher degree of anger, whereas children who did not change their scores and those who were suggested to take more balls for themselves were perceived as neutral.

The machine condition of the experiment did not meet the expectations of children. They expected a robot from science-fiction movies with the capacity for spontaneous communication, whereas the robot Pepper did not move and could not answer questions that were not part of the experiment.

Parts of the contributions of this chapter were published in Kalashnikova et al. (2023b) and Kalashnikova et al. (2023a).

Chapter 5

Influence of interlocutor and propensity to be nudged on speech production

In this chapter, we present the results of the linguistic analysis of our data. We are interested in describing the paralinguistic and lexical differences in speech regarding 1) the type of speaker's interlocutor and 2) the speaker's propensity to be influenced. We analyzed pitch, intensity, speech rate, duration of a speaking turn, and frequency of disfluences for the paralinguistic analysis, and total number of words, unique words, lexical and non-lexical words, and pronouns per speaker, the ratio of these parameters and the total number of words per speaker, and the ratio of total number of words per speaker and total number of utterances of the same speaker for the lexical analysis. We first present the methodology of the paralinguistic analysis and its results comparing paralinguistic parameters between two audiences: adults and children. Secondly, we explain the methodology of the lexical analysis and its comparative results for both experiments.

We hypothesized that one factor influencing someone's propensity to be nudged is their level of communicative alignment with their interlocutor. The higher level of

communicative alignment might be interpreted as a sign of the participant's confidence in the interlocutor's arguments and, therefore, the participant's propensity to follow the direction of the nudge more easily. We have already shown in Chapter 4.1 that the type of conversational agent indeed impacted participants' propensity to be nudged. However, the correlation between the level of communicative alignment and the propensity to be nudged is still unanswered.

In this work, we analyzed the communicative alignment from a paralinguistic and lexical points of view. Being the most common and natural kind of communication, human-human interactions in our data were considered baseline interactions to describe the communicative behavior and the level of communicative alignment in nudging interactions. In that way, the comparison of if and how human-machine interactions (with the Pepper robot and Google Home smart-speaker) were different could indicate the differences in linguistic and paralinguistic alignment with other interlocutors.

Thus, in this chapter, we addressed the following research questions:

1. What are the differences in communicative behavior between speeches addressed to the human, the smart-speaker, and the robot?

2. What are the differences in communicative behavior between participants who were efficiently nudged and those who did not react to nudges?

A better understanding of communicative behavior regarding the nudging interlocutor and reaction to nudges will clarify how nudges influence human speech and how to prevent humans from being nudged.

5.1 Methodology

5.1.1 Paralinguistic characteristics in nudging spoken interactions

To describe paralinguistic characteristics of speech, we analyzed the following parameters: speech rate, frequency of disfluencies, duration of speaking turn, pitch, and intensity at the conversational step level. The conversational steps corresponded to the logical parts of the experiment.

In the study with the adults, we distinguished the following conversational steps:

1. **S0**: establishing common ground with small talk;
2. **S1**: presenting hypothetical situations with the choice between the default and eco-friendly options;
3. **S2**: introducing nudges followed by questions of willingness to adopt ecological habits;
4. **S3**: reproducing similar but slightly different hypothetical situations from step "S1".

In the experiment with children, the conversational steps corresponded to the different games that were proposed to them:

1. **Game 1**: to distribute the small balls between two bowls (the nudging game);
2. **Game 2**: to choose who roll the dice;
3. **Game 3**: to roll the dice and hide the outcome.

We excluded participant's speech before and after the described steps since some participants did not speak when they were not directly asked questions. The presence of data for some participants and the absence of data for others would, therefore, bias the results of paralinguistic analysis.

Speech rate, frequency of disfluences, and duration of speaking turns were first calculated at the level of speaking turns using the manual transcription of the recorded data. Then, we calculated the average value of these parameters for each conversational step described previously, as participants did not have the same number of speaking turns.

- **Speech rate**. To calculate the speech rate, we first tokenized (using a space as a separator) the transcription, excluding the tokens corresponding to disfluences, and then divided the number of tokens of the current speaking turn by the duration of the current speaking turn.

- **Frequency of disfluences** was calculated by dividing the number of disfluences of the current speaking turn by the duration of the current speaking turn.

- **Duration of speaking turn** was estimated in seconds.

For both experiments (with adults and with children), the pitch and intensity were extracted using the script created by Setsuko Shirai (Shirai) on Praat (Boersma and Weenink, 2024). The values of these two parameters were extracted every 10 ms for the whole audio file. The extraction of pitch values was limited between 145 and 275 Hz for female participants and 75 and 175 for male participants (Davies and Goldberg, 2006). After the extraction, the values of pitch and intensity were filtered according to the timesteps of speaking turns. Then, we excluded the sequences of pitch values equal to zero. These sequences might correspond to the pauses within a speaking turn and wrongly detected values. Since we used the scale of minimum and maximum values for pitch detection, we did not normalize the extracted data.

We apply linear interpolation between the previous and the first valid value (less than 500 Hz) for wrong-detected values above this ceiling.

5.1.2 Lexical characteristics in nudging spoken interactions

Similarly to the previous work covered in Section 2.2, we analyzed the same lexical characteristics to define the lexical differences between human-directed speech, smart-speaker-directed speech, and robot-directed speech. We applied two methods to calculate the ratio of unique words, pronouns, lexical (nouns, adjectives, verbs, adverbs), and non-lexical (all other words that are not lexical) words (Fischer, 2011): 1) using lemmatization (as it was suggested by Berman and Nir-Sagiv (2009)), and 2) without lemmatization. However, the first approach (using lemmatization) did not show any significant results, we therefore describe the results obtained with the second approach (without lemmatization).

For the analysis, we measured the following parameters:

- **Total number of words** per speaker (wc);
- **Total number of unique words** per speaker (uw);
- **Total number of lexical words** per speaker (lex);
- **Total number of non-lexical words** per speaker (nlex);
- **Total number of pronouns** per speaker (pro);
- **Verbosity I** - the ratio between the total number of unique words per speaker / total number of words of the speaker (uw-r1);
- **Verbosity II** - the ratio between the total number of unique words per speaker / total number of words in the corpus (uw-r2);
- **Ratio between the total number of lexical words per speaker / total number of words of the speaker** (lex-r);
- **Ratio between the total number of pronouns per speaker / total number of words of the speaker** (pro-r);

- **MLU** - the ratio between the total number of words per speaker / total number of utterances by the same speaker (mlu).

For this analysis, we used manual transcription realized during the corpus annotation. Both lemmatization and tokenization were realized using spacy (Honnibal and Montani, 2017).

Statistical analysis

The significance of the results between groups was tested using a *t-test* for two independent samples, which was applied with SciPy (Virtanen et al., 2020). The obtained p-values were Bonferroni-corrected with threshold alpha = 0.5 using Statsmodels (Seabold and Perktold, 2010).

Using this statistic analysis, the paralinguistic description of speech addressed to different conversational agents resulted in two publications: Kalashnikova et al. (2023a) and Kalashnikova et al. (2023b). The reviewers of our submissions suggested applying the Linear Mixed Model approach, which allows the combination of multiple parameters to analyze continuous variables (e.g., how pitch is predicted by sex and emotion). In this way, we can only use this approach separately for each paralinguistic parameter (e.g., how the type of interlocutor of the group predicts each paralinguistic parameter). However, the model takes the default option for comparing groups without giving the information of comparison between the three groups. For example, when comparing how the type of a conversational agent influenced the duration of a speaking turn, the model took the value of the group of smart-speaker as the reference and calculated the p-value for the comparison between the groups of smart-speaker and human, and between the groups of smart-speaker and robot, but not between the groups of human and robot. Thus, regardless of the statistical power of this approach, it does not seem completely appropriate for our task.

5.2 How does speech differ regarding the type of the interlocutor?

5.2.1 Paralinguistic characteristics

In this subsection, we describe the differences in pitch, intensity, speech rate, frequency of disfluencies, and duration of a speaking turn between the participants' groups when they speak to the human, the smart-speaker, or the robot for the whole conversation and the evolution of values throughout the conversation.

The results of the t-test and p-value of the data from adults for the whole conversation are reported in Table 5.1. The results of the t-test of the data from children for the whole conversation are reported in Table 5.2.

Pitch

Experiment with adults

Groups	**Pitch F**		**Pitch M**		**Intensity**		**Speech rate**		**Disfluency**		**Duration**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
H *vs* R	-1.7	0.3	-3.3	**0.005**	4.9	**8.2e-06**	1.1	0.8	0.44	1.0	2.9	**0.01**
H *vs* S-S	-1.2	0.7	0.7	1.0	3.5	**0.002**	1.7	0.3	2.9	**0.01**	1.9	0.16
S-S *vs* R	-0.6	1.0	-4.2	**0.0003**	2.6	**0.03**	-0.7	1.0	-3.01	**0.007**	1.4	0.48

Table 5.1: Test statistics comparing paralinguistic parameters between pairs of groups of conversational agents. Indications: *F* - female participants; *M* - male participants; *H* - group who interacted with the human agent; *R* - group who interacted with the robot agent; *S-S* - group who interacted with the smart-speaker agent. Data from adults

We analyzed the pitch values of female and male participants separately due to the natural differences in their pitch ranges.

Regarding pitch analysis for the whole conversation, we observed no significant differences between groups of female participants who spoke to either of the three conversational agents. However, the value of the t-test indicated that when speaking to the robot, the pitch values of male participants were significantly higher comparing to conditions with the human (t=-3.3, p=0.005) and the smart-speaker (t=-4.2,

p=0.0003).

However, when comparing the values for each step separately, we found significant differences only for step "S0" between human and robot (t=-2.9, p=0.04) and for step "S3" between robot and smart-speaker (t=-2.66, p=0.05) in the group of male participants and no significant differences in the group of female participants.

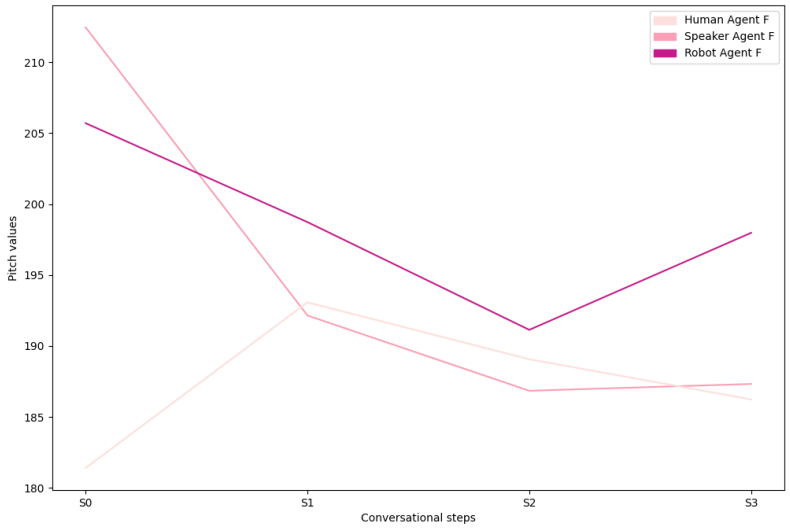


Figure 5.1: Mean pitch values of adult female participants extracted with Praat. The horizontal axis indicates conversational steps

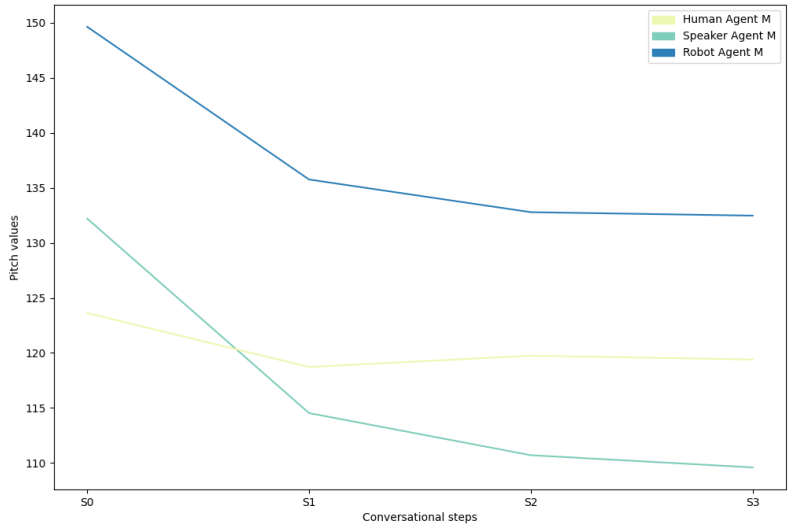


Figure 5.2: Mean pitch values of adult male participants extracted with Praat. The horizontal axis indicates conversational steps

As shown in Figures 5.1 and 5.2, female and male participants had a higher

average pitch at the beginning of the conversation when speaking with a machine. It confirmed previous observations of a higher-pitched voice in communication with a machine. We also observed that regardless of the different phonetic curves between both groups, they shared the same tendencies when speaking to the robot and the smart-speaker. The difference in phonetic curves between female and male participants consisted of a rise at the last step of conversation for female participants and a slight descent at the same step for male participants.

Takeaway

- Male adult participants had a higher mean pitch when speaking with the robot.
- Female and male adult participants had a higher pitch when speaking with the machines at the beginning of the conversation.

Experiment with children

Groups	**Pitch**		**Intensity**		**Speech rate**		**Disfluency**		**Duration**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*	*t*	*p*
H *vs* R	-0.5	0.6	-2.4	**0.02**	2.3	**0.03**	2.0	**0.05**	1.3	0.5
H *vs* S-S	-0.1	0.9	5.4	**1.6e-06**	2.5	**0.02**	3.05	**0.008**	1.1	0.9
S-S *vs* R	-0.5	0.6	-15.8	**1.7e-38**	-0.2	0.9	-1.3	0.5	0.4	1.0

Table 5.2: Test statistics comparing paralinguistic parameters between pairs of groups of conversational agents. Indications: *H* - group who interacted with the human agent; *R* - group who interacted with the robot agent; *S-S* - group who interacted with the smart-speaker agent. Data from children

Contrary to the experiment's procedure with adults, we did not divide participants into groups of their gender, since in their age the difference in pitch is not as important as it is for adults.

The t-test of the pitch values for the whole conversation did not show any significant difference in F0 regarding the type of interlocutor. Moreover, no differences were observed when comparing the values of each conversational step separately.

Figure 5.3 shows the evolution of pitch in human-directed speech (dark pur-

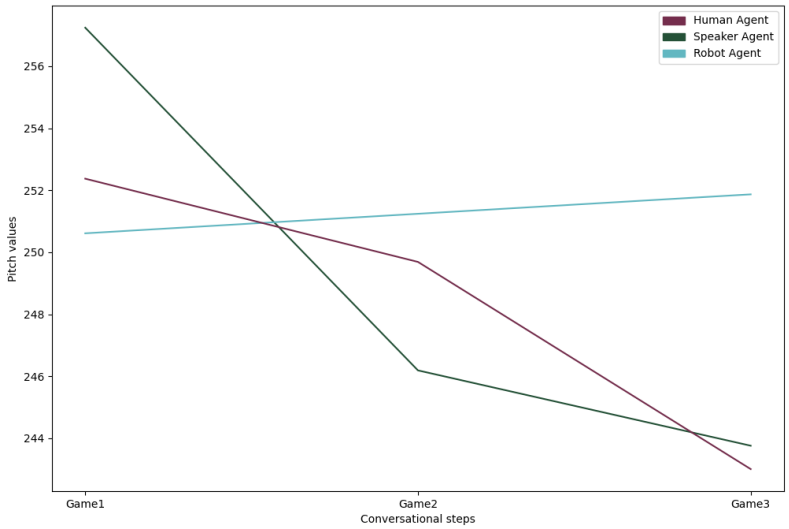


Figure 5.3: Mean pitch values per conversational step. Data from children. The horizontal axis indicates conversational steps

ple), smart-speaker-directed speech (dark green), and robot-directed speech (blue). The F0 values stayed at the same level in robot-directed speech, whereas they decreased in smart-speaker-directed speech and human-directed speech. Even though the pitch values are lower at the beginning of the conversation in robot-directed speech, they stayed higher than those in smart-speaker-directed speech and human-directed speech at steps "Game2" and "Game3".

This observation followed the same tendency observed in the experiment with adults: higher pitch at the beginning of the conversation, especially in machine-directed speech. If a high pitch from adults addressing the machine is associated with infant-directed speech, in the case of children, we hypothesize that the high pitch values at the beginning of the conversation and for the whole conversation in robot-directed speech may be associated with stress. However, this observation is only an observed tendency, and more data are needed to perform a more robust statistical analysis.

Takeaway

- The children's pitch was not significantly impacted by the embodiment

of the conversational agent.

- More pitch variations in children's speech were observed when talking to the human and the smart-speaker.

Intensity

Data from adults

The level of intensity for the whole conversation differed in all pairs of comparison:

- human-directed speech *vs* robot-directed speech: t=4.9, p=**8.2e-06**;
- human-directed speech *vs* smart-speaker-directed speech: t=3.5, p=**0.002**;
- smart-speaker-directed speech *vs* robot-directed speech: t=2.6, p=**0.03**.

We observed that the intensity was the highest in the human-directed speech and the lowest in the robot-directed speech.

The comparison of intensity values for each conversational step showed significant differences between human-directed speech and machine-directed speech for all conversational steps except "S0" and no significant differences between robot-directed speech and smart-speaker-directed speech for any conversational step. We report the results of the statistical analysis in Table 5.3.

Groups	**S1**		**S2**		**S3**	
Test statistics	t	p	t	p	t	p
H *vs* R	5.0	**3.5e-05**	5.6	**3.5e-06**	5.3	**1.2e-05**
H *vs* S-S	3.6	**0.003**	4.5	**0.0002**	4.1	**0.0006**

Table 5.3: Test statistics comparing intensity levels between pairs of groups of conversational agents. Indications: *H* - group who interacted with the human agent; *R* - group who interacted with the robot agent; *S-S* - group who interacted with the smart-speaker agent. Data from adults

We observed (Figure 5.4) that participants had the intensity at almost the same level when speaking with the three conversational agents at the beginning of the conversation. However, from step "S1" participants had a higher level of intensity

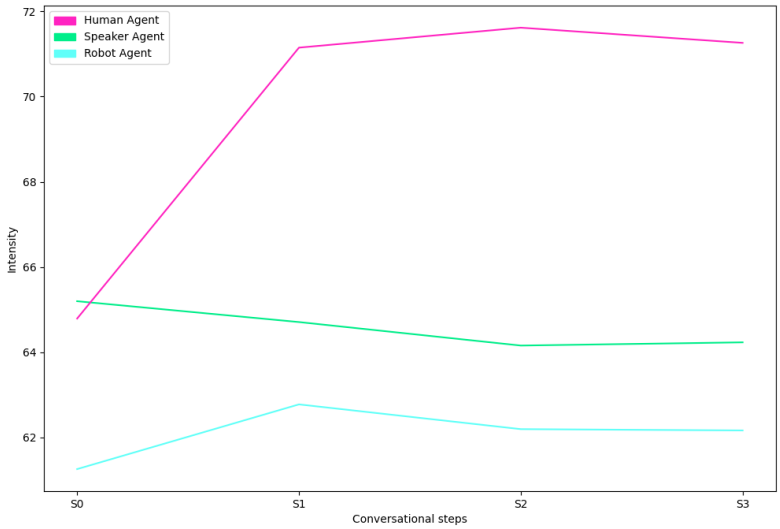


Figure 5.4: Mean intensity values. Data from adults. The horizontal axis indicates conversational steps

till the end of the conversation when speaking to the human. The level of the intensity was also increased at step "S1" in robot-directed speech but to a lesser extent.

> **Takeaway**
>
> - The human-directed speech of adults was characterized by the highest intensity and the robot-directed speech by the lowest.

Data from children

In general, the intensity in robot-directed speech (mean=70 dB) was significantly higher than in smart-speaker-directed speech (mean=52.6 dB) and human-directed speech (mean=64.9 dB). The intensity was also significantly higher in human-directed speech compared to smart-speaker-directed speech.

The intensity level stayed almost the same throughout the conversation in all three conditions. Contrary to the experiment with adults, the highest intensity was observed in robot-directed speech, and the lowest in smart-speaker-directed speech (Figure 5.5). Thus, the analysis of the intensity comparison between the three types of directed speeches showed significant differences for all steps between robot-

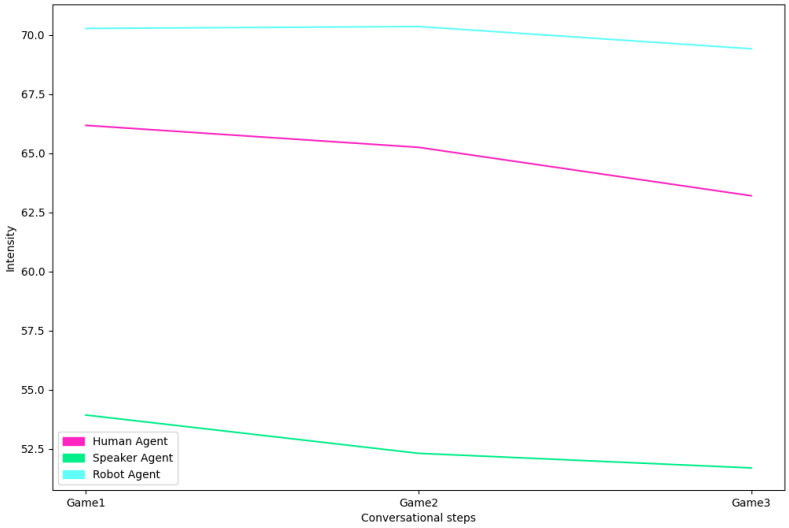


Figure 5.5: Mean intensity values per conversational step. Data from children. The horizontal axis indicates conversational steps

directed speech and smart-speaker-directed speech, and between human-directed speech and smart-speaker-directed speech. We report these results in Table 5.4.

Groups	**S1**		**S2**		**S3**	
Test statistics	*t*	*p*	*t*	*p*	*t*	*p*
H *vs* SS	3.4	**0.002**	3.4	**0.001**	2.5	**0.02**
S-S *vs* R	-11.3	**2.1e-17**	-9.6	**7.9e-14**	-7.6	**1.9e-11**

Table 5.4: Test statistics comparing intensity levels between pairs of groups of conversational agents. Indications: *H* - group who interacted with the human agent; *R* - group who interacted with the robot agent; *S-S* - group who interacted with the smart-speaker agent. Data from children

> **Takeaway**
>
> - The robot-directed speech of children was described with the highest intensity, and the smart-speaker-directed speech with the lowest.

Speech rate

Data from adults

We found no significant differences in speech rate for the whole conversation. Moreover, when comparing the evolution of speech rate during the conversation, the only significant difference was observed for step "S2" between human-directed speech

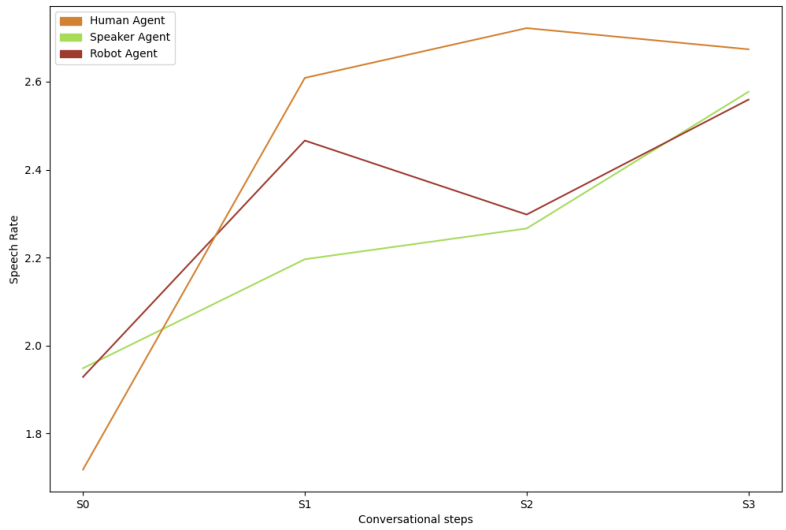


Figure 5.6: Mean speech rate. Data from adults. The horizontal axis indicates conversational steps

and robot-directed speech: t=3.7, p=0.002; and between human-directed speech and smart-speaker-directed speech: t=3.8, p=0.001. Since this step corresponds to the step where we nudged participants, we noted that when arguing their choices, participants slowed their speech rate when speaking to machines.

As shown in Figure 5.6, when addressing the human, participants spoke with a higher speech rate except for the beginning of the conversation than the groups who spoke to the robot and the smart-speaker. Whereas the speech rate in the smart-speaker-directed speech increased throughout the conversation, this parameter increased at step "S1", decreased at step "S2", and increased again at the end of the conversation in robot-directed speech.

> **Takeaway**
>
> - During nudging interventions, adults significantly slowed their speech rate when talking to the machines.

Data from children

The group of children interacting with the human agent spoke significantly faster than those speaking to machines for the whole conversation. Thus, statistically sig-

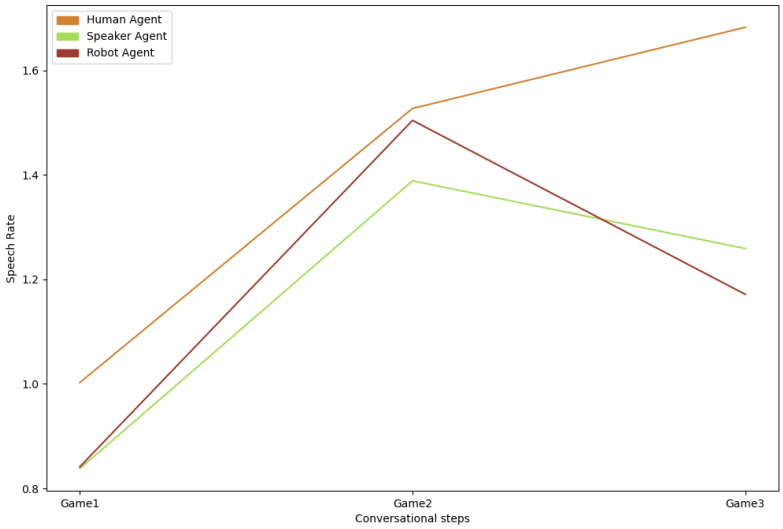


Figure 5.7: Mean speech rate per conversational step. Data from children. The horizontal axis indicates conversational steps

nificant differences were observed between human-directed speech and robot-directed speech (t=2.3, p=0.03), and between human-directed speech and smart-speaker-directed speech (t=2.5, p=0.02). However, even though the tendency to speak faster when talking to the human agent was observed throughout all conversational steps, it differed significantly only for the step "Game3" when analyzing the conversational steps separately:

- human-directed speech *vs* robot-directed speech: t=2.9, p=**0.006**;
- human-directed speech *vs* smart-speaker-directed speech: t=2.4, p=**0.02**.

Figure 5.7 shows that children spoke faster to the robot agent for the two first games but slower for the last game, compared to the speech rate of children who exchanged with the smart-speaker. Nevertheless, this tendency was not statistically significant, and we can conclude that globally, children's speech differed in terms of speech rate between human-directed speech and machine-directed speech.

Takeaway

- Children spoke significantly faster to the human than to the machines.

Frequency of disfluencies

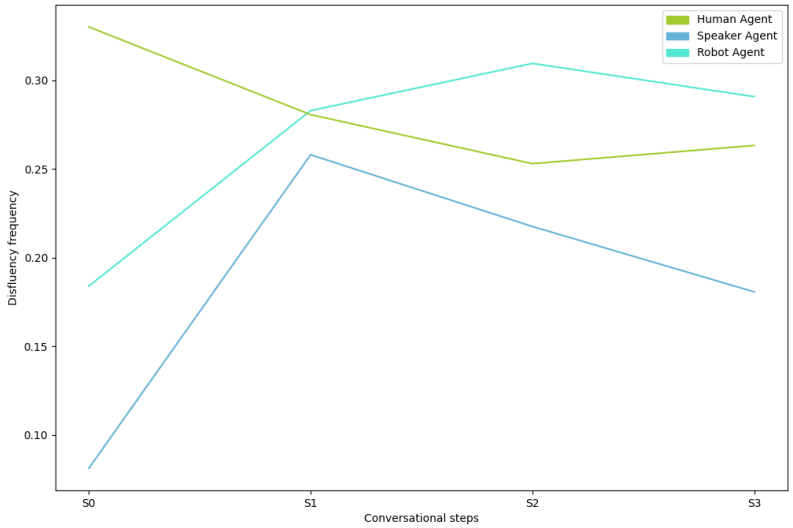


Figure 5.8: Mean frequency of disfluencies. Data from adults. The horizontal axis indicates conversational steps

Data from adults

In general, participants used significantly more disfluencies for the whole conversation when speaking to the human and the robot and less when speaking to the smart-speaker:

- human-directed speech *vs* smart-speaker-directed speech: t=2.9, p=**0.01**;
- smart-speaker-directed speech *vs* robot-directed speech: t=-3.01, p=**0.007**.

Nevertheless, throughout the conversational steps, the only statistically significant change in the frequency of disfluencies was observed at the step "S0" between human-directed speech and smart-speaker-directed speech: t=2.65, p=0.04.

As we can observe in Figure 5.8, participants used disfluencies the most in human-directed speech and the less in smart-speaker-directed speech. Moreover, the frequency of disfluencies stayed the least throughout the conversation in smart-speaker-directed speech. However, when introducing nudges ("S2"), the frequency of disfluencies is the highest in robot-directed speech. Taking this tendency into account, combined with observing a lower speech rate at the same step, we hypothesized that participants had difficulties arguing their choices when speaking to the robot.

Takeaway

- The smart-speaker-directed speech of adults was characterized by the lowest frequency of disfluencies.
- The robot-directed speech of adults during the step of nudging intervention had the highest use of disfluencies.

Data from children

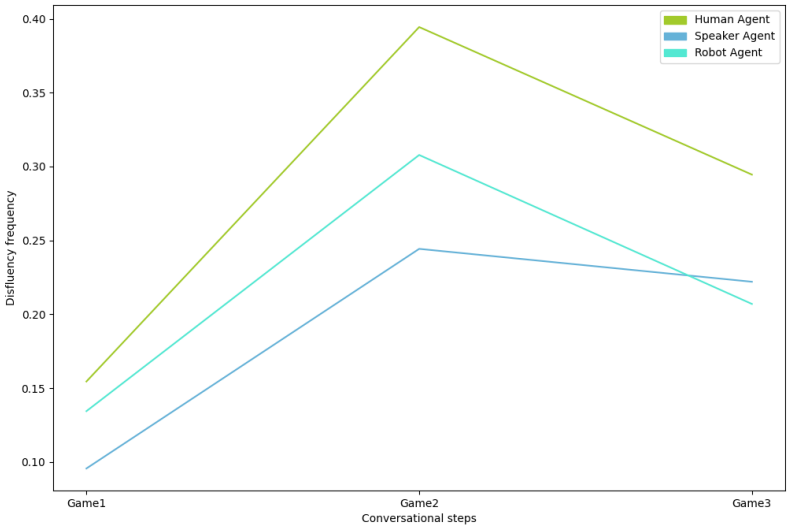


Figure 5.9: Mean frequency of disfluencies per conversational step. Data from children. The horizontal axis indicates conversational steps

Like the speech rate, children used significantly more disfluencies when speaking to the human agent than when speaking to one of the machine agents:

- human-directed speech *vs* robot-directed speech: t=2.0, p=**0.05**;
- human-directed speech *vs* smart-speaker-directed speech: t=3.05, p=**0.008**.

And if in adults' corpus, a higher frequency of disfluencies was correlated with the longer speaking turn, in children's corpus, the duration of a speaking turn did not differ between groups regarding the type of conversational agent.

However, at the level of conversational steps, the only significant difference was observed for the step "Game2" between human-directed speech and smart-speaker-directed speech: t=2.3, p=0.03.

When analyzing the evolution of the frequency of disfluencies during the conversation, we observed that participants of the three groups globally followed the same pattern. The least disfluencies were produced for the step "Game1" and the most for the step "Game2".

> **Takeaway**
>
> - The human-directed speech of children was characterized by the highest frequency of disfluencies.

Duration

Data from adults

The only significant difference in the duration of a speaking turn during the whole conversation was observed between human-directed speech and robot-directed speech: t=2.9, p=0.01, indicating that the speaking turn was significantly longer when the participant exchanged with the human than with the robot. The same tendency was observed when analyzing the evolution of values for the conversational steps. Thus, we reported significant differences between human-directed speech and robot-directed speech:

- "S2": t=3.3, p=**0.009**;
- "S3": t=3.01, p=**0.02**.

The average duration of a speaking turn followed the same pattern during the conversational steps for all groups, with the longest speaking turn during the step introducing nudges (Figure 5.10. Moreover, during the whole conversation, participants had the longest speaking turn when addressing the human and the shortest when addressing the robot.

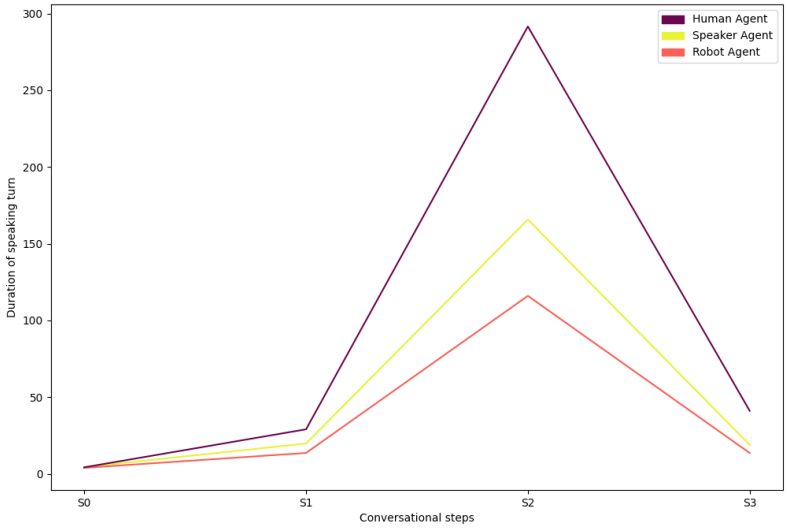


Figure 5.10: Mean duration of a speaking turn. Data from adults. The horizontal axis indicates conversational steps

Takeaway

- The speaking turn of adults was significantly longer in the human-directed speech.

Data from children

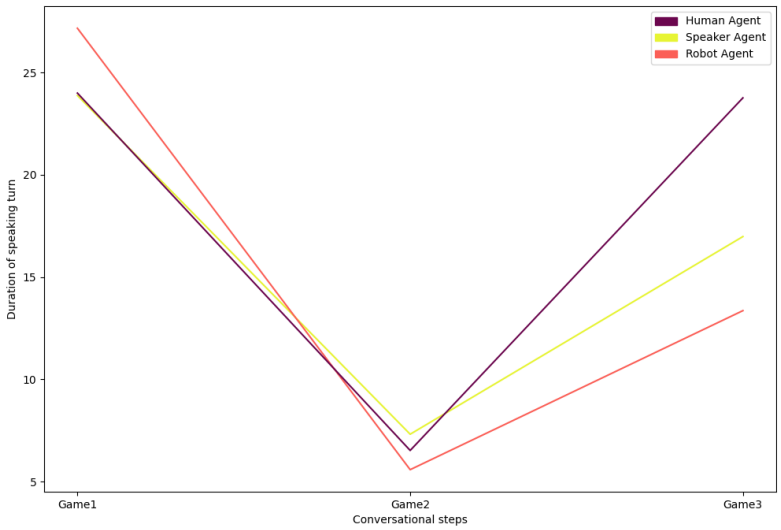


Figure 5.11: Mean duration of a speaking turn per conversational step. Data from children. The horizontal axis indicates conversational steps

Globally, children followed the same pattern of the duration of a speaking turn during the conversation regarding the type of their interlocutor (Figure 5.11). The

lowest speaking turn's duration was observed for the step "Game2" and the highest for the step "Game1". However, during the step "Game3", children spoke to the human agent for significantly longer than to the robot agent: t=2.7, p=0.03.

> **Takeaway**
>
> - The embodiment of a conversational agent did not influence the duration of a speaking turn of children.

5.2.2 Lexical characteristics

Data from adults

Table 5.5 presents the t-statistic value and p-value for the measured parameters in comparison of groups regarding their type of interlocutor for the whole conversation.

Measure	**wc**	**uw**	**lex**	**nlex**	**pro**	**mlu**	**uw-r1**	**lex-r**	**pro-r**	**uw-r2**
Human *vs* Smart-Speaker										
t	2.4	2.5	2.4	2.5	2.7	2.2	-3.5	-1.6	2.9	1.9
p	**0.02**	**0.02**	**0.02**	**0.02**	0.1	**0.03**	**0.0009**	0.1	**0.005**	0.08
Human *vs* Robot										
t	3.3	3.8	3.3	3.2	3.5	3.6	-5.2	-1.2	2.8	1.4
p	**0.003**	**0.0008**	**0.003**	**0.003**	**0.002**	**0.001**	**4.5e-06**	0.3	**0.007**	0.2
Robot *vs* Smart-Speaker										
t	-1.5	-1.6	-1.5	-1.4	-1.5	-1.3	1.3	-0.6	0.2	0.7
p	0.2	0.1	0.1	0.2	0.1	0.2	0.2	0.6	0.9	0.5

Table 5.5: Test statistics comparing lexical parameters between pairs of groups of conversational agents. Indications: *wc* - total number of words per speaker, *uw* - total number of unique words per speaker, *lex* - total number of lexical words per speaker, *nlex* - total number of non-lexical words per speaker, *pro* - total number of pronouns per speaker, *mlu* - ratio number of words per speaker / total number of utterances by the speaker, *uw-r1* - ratio unique-words / total number of words per speaker, *pro-r* - ratio pronouns / total number of words per speaker, *uw-r2* - ratio unique-words / total number of words in the corpus. Data from adults

The results of the lexical analysis repeated the same tendencies observed for the paralinguistic analysis: participants spoke more to the human than to the robot or the smart-speaker. Therefore, all measures are tightly correlated to the length of

the utterances/dialogue. Moreover, none of the measured parameters was significant between robot-directed speech and smart-speaker-directed speech. We argued that participants differentiated their lexical choices even more than their speech regarding the type of their interlocutor. In the following, we describe the lexical tendencies that differed between participants who exchanged with the human and participants who spoke to the robot and the smart-speaker.

Thus, when speaking with the human agent, participants used significantly more words in general, unique words, lexical words, non-lexical words, and pronouns, than when speaking to the smart-speaker or the robot. These observations confirm the results obtained by Fischer et al. (2011); Brennan (1991).

The significant differences were also observed for the following parameters: the mean length of utterance (ratio between the total number of words per speaker / total number of utterances by the same speaker), uw-r1 (ratio between the total number of unique words per speaker / total number of words of the speaker), and pro-r (ratio between the total number of pronouns per speaker / total number of words of the speaker).

A significantly higher measure of mean length of utterance indicated that participants used more complex utterances when speaking to the human agent. Even though the absolute value of unique words is significantly higher in human-directed speech, its proportional value is significantly lower in human-directed speech compared to robot-directed speech and smart-speaker-directed speech. Moreover, the proportional value of lexical words compared to the total number of words by the same person was not significantly different between the groups of different agents. These results indicated that participants spoke more when communicating with the human agent, but since the conversational theme was very specific, the participants of all groups used the same lexical choices.

The measure of verbosity in the approach of Fischer (2011) - the ratio of unique words compared to the total number of words in the corpus was not statistically

significant between groups.

Previous studies found that the differences decreased throughout the conversation (Amalberti et al., 1993). However, in our study, none of the measured parameters significantly differed among the three groups at the beginning of the conversation. We suggest that participants did not show any differences in their lexical choices regarding the type of interlocutor at the step of small-talk. Similarly, most parameters stayed statistically insignificant for the comparison between human-directed speech and smart-speaker-directed speech but became statistically significant between human-directed speech and robot-directed speech at the step "S1" (hypothetical situations). The differences increased significantly at step "S2" (nudging) between the group of human agent and the groups of machine agents till the end of the conversation. These results confirmed the tendencies observed during the paralinguistic analysis: participants distinguished the robot agent as the opposite of the human agent, placing the smart-speaker agent between them, however, closer to the robot agent.

We report these results in Table 5.6.

> **Takeaway**
>
> - Adult participants spoke significantly differently to the human than to the machines.
> - The human-directed speech of adults is characterized by more complex and longer utterances.

Experiment with children

We present the results of the statistical analysis of lexical measures for the whole conversation in Table 5.7.

Similarly to the results obtained for the experiment with adults, most statistically significant changes were observed between the group that exchanged with the human agent and the groups that spoke to two machine agents, and almost no signif-

Step	Measure	wc	uw	lex	nlex	pro	uw-r1	pro-r	mlu
Human *vs* Smart-Speaker									
S1	*t*	3.2	1.7	1.6	1.8	1.5	-1.8	0.8	1.0
	p	**0.004**	0.1	0.1	0.09	0.1	0.07	0.4	0.3
S2	*t*	2.5	2.6	2.5	2.5	2.7	-4.04	3.5	2.2
	p	**0.02**	**0.02**	**0.02**	**0.02**	**0.01**	**0.0002**	**0.001**	**0.03**
S3	*t*	2.3	2.5	2.2	2.3	2.3	-2.2	0.7	1.3
	p	**0.03**	**0.02**	**0.04**	**0.03**	**0.03**	**0.03**	0.5	0.2
Human *vs* Robot									
S1	*t*	3.2	2.6	2.2	2.2	1.9	-2.7	0.5	1.8
	p	**0.004**	**0.02**	**0.04**	**0.03**	0.07	**0.008**	0.6	0.08
S2	*t*	3.4	3.9	3.4	3.3	3.7	-5.7	3.1	3.4
	p	**0.003**	**0.0006**	**0.002**	**0.003**	**0.001**	**7.8e-07**	**0.003**	**0.0002**
S3	*t*	2.7	3.2	2.7	2.8	2.7	-3.4	0.2	2.5
	p	**0.01**	**0.004**	**0.01**	**0.01**	**0.01**	**0.002**	0.9	**0.01**

Table 5.6: Test statistics comparing lexical parameters between pairs of groups of conversational agents. Indications: *wc* - total number of words per speaker, *uw* - total number of unique words per speaker, *lex* - total number of lexical words per speaker, *nlex* - total number of non-lexical words per speaker, *pro* - total number of pronouns per speaker, *mlu* - ratio number of words per speaker / total number of utterances by the speaker, *uw-r1* - ratio unique-words / total number of words per speaker, *pro-r* - ratio pronouns / total number of words per speaker; "S0", "S1", "S2", "S3" - conversational steps. Data from adults

icant differences of lexical measures were found between robot-directed speech and smart-speaker-directed speech. We also noted that significant differences occurred only in the measures of total values. Thus, the particularities of human-directed speech were a significantly higher total number of words per conversation, a total number of unique words, lexical and non-lexical words, and pronouns relative to robot-directed speech and smart-speaker-directed speech. The only statistically significant difference between robot-directed speech and smart-speaker-directed speech was observed for the measure of the ratio of lexical words and the total number of words of the speaker. Thus, children who spoke to the smart-speaker used significantly higher pronouns than those who spoke to the robot in a proportional value.

The analysis of the evolution of the lexical measures throughout conversational steps did not show any significant differences at the step of nudging ("Game1").

Measure	wc	uw	lex	nlex	pro	uw-r1	lex-r	pro-r	mlu	uw-r2
Human *vs* Smart-Speaker										
t	2.9	3.0	3.0	2.6	2.8	-0.2	-0.2	1.9	1.7	-0.2
p	**0.007**	**0.004**	**0.005**	**0.01**	**0.007**	0.8	0.8	0.06	0.1	0.8
Human *vs* Robot										
t	3.7	4.03	4.3	3.2	3.7	-1.5	1.9	1.9	0.07	1.0
p	**0.007**	**0.0003**	**0.001**	**0.003**	**0.0008**	0.2	0.06	0.06	0.9	0.3
Robot *vs* Smart-Speaker										
t	-1.1	-0.9	-1.7	-0.7	-0.8	1.7	-2.5	0.03	1.4	-0.9
p	0.2	0.3	0.09	0.5	0.4	0.09	**0.01**	0.9	0.2	0.4

Table 5.7: Test statistics comparing lexical parameters between pairs of groups of conversational agents. Indications: *wc* - total number of words per speaker, *uw* - total number of unique words per speaker, *lex* - total number of lexical words per speaker, *nlex* - total number of non-lexical words per speaker, *pro* - total number of pronouns per speaker, *mlu* - ratio number of words per speaker / total number of utterances by the speaker, *uw-r1* - ratio unique-words / total number of words per speaker, *pro-r* - ratio pronouns / total number of words per speaker, *uw-r2* - ratio unique-words / total number of words in the corpus. Data from children

Takeaway

- The human-directed speech of children differed from machine-directed speech in absolute values.

To summarize:

Takeaway

- The robot-directed speech of adults was characterized by a higher mean pitch, a lower intensity during the conversation, and a slower speech rate and a higher frequency of disfluencies during nudging intervention.
- The smart-speaker-directed speech of adults was characterized by a lower frequency of disfluencies.
- The human-directed speech of adults was characterized by a higher intensity, a higher speech rate, and a longer speaking turn.
- The robot-directed speech of children was characterized by a higher in-

tensity.

- The smart-speaker-directed speech of children was characterized by a lower intensity.
- The human-directed speech of children was characterized by a higher frequency of disfluencies.

5.3 How does speech differ regarding the propensity to be influenced?

For this section, we analyzed if the same paralinguistic and lexical parameters changed regarding participants' propensity to be nudged. Similarly, we compared the significance of differences for the whole conversation and the conversational steps. As a reminder, we used two boolean measures to define the propensity to be nudged:

- qualitative - if a participant changed their willingness score for at least two points;
- quantitative - if a participant changed their willingness score for at least two answers.

Thus, we compared the paralinguistic parameters for the following groups: 1) those who changed their answers for at least two points *vs* those who did not (group "qualitative"); 2) those who changed their answers for at least two questions *vs* those who did not (group "quantitative").

Similarly, for analyzing data from adults, we compared twice the significance of paralinguistic parameters between children who changed their number of balls and those who did not after the first and second nudges.

5.3.1 Paralinguistic characteristics

Data from adults

The analysis showed that participants who changed their willingness scores for at least two points had significantly lower average pitch than those who did not change their willingness scores for the whole conversation: t=-1.93, p=0.05. However, no significant differences were observed for the group "quantitative" nor the conversational steps for both groups.

We found no significant differences in intensity or speech rate for either group for the whole conversation nor the separate analysis of conversational steps.

Disfluencies appeared significantly more frequently in participants' speech susceptible to nudges from a quantitative point of view: t=2.03, p=0.04. The comparison of the frequency of disfluencies throughout the conversational steps showed the same tendency with the significant differences for steps "S1" (t=2.73, p=0.02) and "S2" (t=2.21, p=0.05). Nevertheless, no significant differences were observed for the group "qualitative".

Participants with the propensity to be influenced from qualitative and quantitative points of view spoke significantly longer than those who were not susceptible to nudges:

- Group "qualitative": t=2.0, p=**0.04**;
- Group "quantitative": t=1.92, p=**0.05**.

When comparing the groups for each conversational step, we observed the significant differences only during step "S2", which corresponds to the step of nudging.

- Group "qualitative" ("S2"): t=2.34, p=**0.02**;
- Group "quantitative" ("S2"): t=2.4, p=**0.02**.

Data from children

Among all the paralinguistic parameters (pitch, intensity, speech rate, duration of a speaking turn, frequency of disfluencies) that we analyzed, only the difference in intensity was found to be statistically significant between children susceptible to influence and those who did not after the second nudge: t=2.69, p=0.008.

5.3.2 Lexical characteristics

Data from adults

Similar to the paralinguistic analysis, we investigated how lexical characteristics differed between participants susceptible to nudges from qualitative and quantitative points of view. We used the same methodology and lexical parameters to answer this question. The results of this analysis for the whole conversation are presented in Table 5.8.

Measure	**wc**	**uw**	**lex**	**nlex**	**pro**	**uw-r1**	**lex-r**	**pro-r**	**mlu**
Qualitative measure									
t	2.2	1.8	2.1	2.3	2.4	-2.2	-0.3	0.9	2.6
p	**0.03**	0.07	**0.04**	**0.02**	**0.02**	**0.03**	0.7	0.4	**0.009**
Quantitative measure									
t	2.4	1.4	2.1	2.5	1.9	-2.4	-1.7	-2.5	2.5
p	**0.02**	0.2	**0.04**	**0.01**	0.06	**0.03**	0.1	**0.03**	**0.02**

Table 5.8: Test statistics comparing lexical parameters between group "nudged" and group "not-nudged" for the whole conversation. Indications: *wc* - total number of words per speaker, *uw* - total number of unique words per speaker, *lex* - total number of lexical words per speaker, *nlex* - total number of non-lexical words per speaker, *pro* - total number of pronouns per speaker, *mlu* - ratio number of words per speaker / total number of utterances by the speaker, *uw-r1* - ratio unique-words / total number of words per speaker, *pro-r* - ratio pronouns / total number of words per speaker. Data from adults

We observed that participants susceptible to nudges from both qualitative and quantitative points of view used significantly more words in total, lexical and non-lexical words. Moreover, a significantly higher value of mean length of utterance indicated that their utterances were more complex.

The total number of unique words per speaker was not significantly different between groups of participants who changed their willingness score to adopt another

ecological behavior and those who did not. However, the ratio between the total number of unique words and the total number of words of the same speaker showed that "nudged" participants used significantly fewer unique words.

We also found differences between participants who corresponded to the qualitative measure of effectiveness of nudges and those who corresponded to the quantitative measure. Thus, the group that changed their scores according to qualitative criteria had a significantly higher total number of pronouns per speaker. However, the group that changed their willingness scores to more questions (quantitative measure) used significantly fewer pronouns than the group that did not change their scores. Nevertheless, due to the size of our corpus, the analysis of each conversational step was needed to establish if there were any differences between groups of two measures.

Concerning the evolution of the lexical parameters throughout conversational steps, we found that the statistically significant differences between the group susceptible to nudges and those not susceptible to nudges were observed only for the step "S2" corresponding to the step when we introduced nudges. It confirmed that the observations made during the analysis of the whole conversation were mainly expressed during the nudging step. We report the significant results for this step in Table 5.9.

We can conclude that participants susceptible to nudges spoke more with a higher frequency of pronouns and more complex utterances.

Data from children

When analyzing lexical measures of children who changed their number of balls after the first and the second nudges (group "nudged") and those who did not (group "not-nudged"), we did not find any significant differences.

However, due to the stress and unusual situation, children's speech was limited. The average total number of words per speaker was indeed rather low: 97 words per conversation in human-directed speech, 58 words per conversation in robot-directed

Measure	wc	lex	nlex	pro	uw-r1	mlu
Qualitative measure						
t	2.4	2.2	2.5	2.5	-2.0	2.4
p	**0.02**	**0.03**	**0.01**	**0.01**	**0.04**	**0.02**
Quantitative measure						
t	2.7	2.4	2.9	2.3	-2.7	3.1
p	**0.01**	**0.02**	**0.006**	**0.03**	**0.02**	**0.005**

Table 5.9: Test statistics comparing lexical parameters between group "nudged" and group "not-nudged" for the conversational step "S2". Indications: *wc* - total number of words per speaker, *uw* - total number of unique words per speaker, *lex* - total number of lexical words per speaker, *nlex* - total number of non-lexical words per speaker, *pro* - total number of pronouns per speaker, *mlu* - ratio number of words per speaker / total number of utterances by the speaker, *uw-r1* - ratio unique-words / total number of words per speaker, *pro-r* - ratio pronouns / total number of words per speaker. Data from adults

speech, and 66 words per conversation in smart-speaker-directed speech. Moreover, the value of the mean length of utterance was low compared to the results obtained for the adult's corpus and was not significant for any pair of comparisons, indicating the simplicity of utterances.

In that way, the statistical analysis is highly sensible to noise and extreme values. The presented results reveal only the tendencies in this particular study, and more data are needed to reproduce the analysis.

Takeaway

- The nudged adult participants spoke significantly longer with more complex utterances with more pronouns.
- The nudged children had a higher intensity.

5.4 Discussion

In this Chapter, we analyzed phonetic and lexical parameters to describe participants' speech according to 1) the type of their interlocutor and 2) their propensity to be influenced.

Regarding the type of conversational agent, the speech of participants who exchanged with the human agent was characterized by higher intensity, more frequent use of disfluencies, and a longer average speaking turn compared to one of the machine agents. The robot-directed speech was described by the lowest intensity level, lowest duration of a speaking turn, and less frequent use of disfluencies.

For the pitch values, we found no differences in pitch for female participants and significantly higher pitch for men when speaking to the robot compared to men speaking to the human and the smart-speaker.

In the experiment with children, no significant differences in pitch and duration were found between groups of different conversational agents. Similarly to the adults, children who spoke to the human had a higher speech rate and more frequent use of disfluencies compared to the other two groups. However, the highest intensity was observed for a group that exchanged with the robot.

Lexical analysis showed that adult participants mainly distinguished the human agent from two other agents, and the robot and the smart-speaker agents were seen similarly as machine agents. More complex and longer utterances (therefore, more unique words, pronouns, and lexical and non-lexical words in total value) characterized the lexical choice of participants who exchanged with the human. However, the proportional values of lexical measures indicate that their lexical choices were mostly impacted by the specific topic of the experiment and less by the type of interlocutor.

Similar to the experiment's results with adults, most changes in lexical measures were observed between children who exchanged with the human and two other groups. However, these changes were observed only in absolute values. Proportional values indicated that children's lexical choice was not affected by the type of interlocutor or their propensity to be nudged.

Thus, lexical and phonetic analyses of the experiment with adults allow us to conclude that the agent's embodiment plays an important role in human linguistic

behavior. If we can draw a scale of it, the human and the robot would be placed on polar edges and the smart-speaker somewhere between them, closer to the robot.

Regarding the propensity to be influenced, the peculiarities of speech of the group "nudged" were a lower pitch, more frequent use of disfluencies, and a higher duration of a speaking turn. These results indicate that participants susceptible to nudges intended to exchange ideas with their interlocutor and argue their opinions. This observation was confirmed during the lexical analysis. Thus, nudged adults spoke more and used more unique words and complex utterances. Moreover, these differences mainly occurred during the nudging step.

The speech of successfully nudged children differed only in intensity from that of not-nudged children. We hypothesize that the general environment (being filmed, communicating with an unusual interlocutor, being surrounded by unknown adults) could impact children and their speech more than nudges or the type of their interlocutor.

Chapter 6

Detection of nudges in spoken interactions

In previous chapters, we proposed metrics to consider the participant's propensity to be nudged. We also described participants' linguistic, paralinguistic, and emotional alignment regarding their propensity to be nudged. This chapter aims to model these linguistic, paralinguistic, and emotional characteristics to predict the outcome of the nudging spoken interactions regardless of the differences in speech related to the type of participants' interlocutor. We first explain the pre-processing techniques applied to audio and textual data and affective labels. Secondly, we present the experimental setting and how it differs from the baseline approach of the domain of context-aware emotion recognition in conversation. Thirdly, we demonstrate the results of the classification. Finally, we discuss the results and future research paths.

Our final research question that is addressed in this chapter is:

- How to predict the outcome of the nudging spoken interaction using linguistic, paralinguistic, and emotional cues?

As a reminder, we grouped participants into three classes, depending on their propensity to be nudged:

- "Not-Nudged" - participants who did not change their scores after nudging

intervention;

- "Moderately nudged" - participants who 1) changed their willingness score for at most one point and at most for two questions in the experiment with adults, or 2) changed their number of balls only after one out of two nudges in the experiment with children;

- "Nudged" - participants who 1) changed their willingness score for at least two points and at least for two questions, or 2) changed the number of balls after two nudges in the experiment with children.

Since the division of the groups was arbitrary, we hypothesize that the adult participants of the group "Moderately Nudged" possess characteristics closer to the group "Not-Nudged" and could be classified as such. In contrast, children participants of the group "Moderately Nudged" are characterized closer to the group "Nudged" and could be classified by our system as such.

To our knowledge, no system has been proposed to detect nudging interventions in spoken interactions. Therefore, our work is the first of this kind. However, we face several challenges:

- The small size (in terms of the number of participants) of our datasets raises the problem of limited generalization of the system predicting the outcome of the nudging spoken interactions.

- The classes in the datasets are not balanced. In the data from adults, 10 participants are considered "Not-Nudged", 13 - "Moderately Nudged", and 51 "Nudged". In the data from children, 26 participants belong to the group "Not-Nudged", 21 to the group "Moderately Nudged", and 31 to the group "Nudged".

- The proposed system should be capable of indicating whether the participants of the group "Moderately Nudged" share characteristics with one of the two

other groups and should be considered "Nudged" or "Not-Nudged".

> We propose to extract audio and text features, and emotion embeddings to train a model based on LSTM. We use a strategy based on ensemble methods to train multiple identical models to predict the outcome of the nudging interaction.

6.1 Pre-processing

In this Section, we describe the pre-processing techniques that were used to extract the audio and textual features from the recorded data. Figure 6.1 resumes the pre-processing.

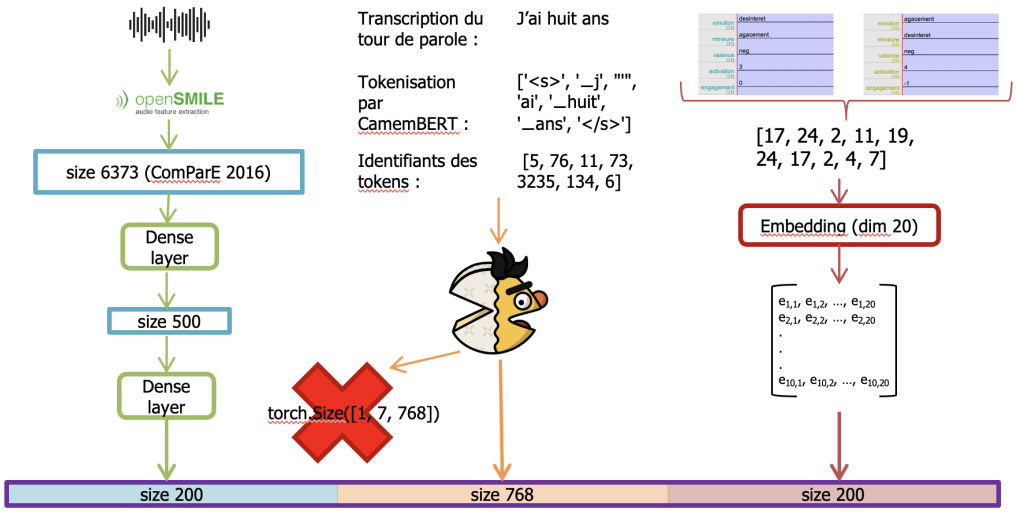


Figure 6.1: Illustration of the pre-processing.

Audio

In line with previous research (Majumder et al., 2019; Poria et al., 2017) in context-aware emotion recognition, we computed audio features using openSMILE software (Eyben et al., 2010) and specifically IS13-ComParE configuration file. This feature set consists of various functionals of low-level descriptor contours, such as MFCC, intensity, pitch, etc. and results in 6373 features.

We extracted these features from each speaker turn of the conversation. After extraction, each of the 6373 features was normalized independently. The maximum and the minimum values of each feature for the whole dataset were assigned to the values of 1 and -1. Linear scaling was then applied to other values of each feature.

Text

The textual features were extracted using CamemBERT (Martin et al., 2020) - BERT trained on French data. The use of a pre-trained language model such as BERT allows to produce the contextual representation of a speaking turn. CamemBERT computes a pooler output for a feature vector of dimension 768.

Emotion labels

For the affective representation of a speaking turn we used all levels of annotation: double emotion annotation, polarity, activation, and engagement. To keep the rich affective content of the speech, we kept the labels from both labelers, assigning the same weight to the labels for both annotators. In cases when the annotator used only one (major) emotion label and not two (major and minor), we repeated the label used twice. Each utterance is, therefore, presented by 10 labels: (2 emotions, polarity, activation, and engagement) x 2 annotators. The labels were passed to the Embedding layer of dimension 20 which creates a matrix of size 20 x 10, where each of ten labels is characterized by 20 embeddings.

Thus, affective labels are represented as embedding vectors which are concatenated together. As in the example below:

Original: *[euh] oui je pense que le train c'est [euh] je le trouve plus confortable et parfois plus vite pour des courtes oui des courtes distances [euh] nationales [rire]*

Translation: *[euh] yes I think that the train is [euh] I think it is more comfortable and sometimes faster for the short yes the short distances [euh] national [laugh]*

The list of concatenated labels ((major emotion, minor emotion, polarity, activation, engagement) x2) of the two annotators for this utterance: ['embarrassment', 'interest', 'pos', '2', '4', 'interest', 'embarrassment', 'pos', '2', '4'].

The vector of emotion embeddings given to the system: [0, 14, 10, 12, 11, 14, 0, 10, 12, 11].

6.2 Experimental setting

Baseline

The systems reviewed in Section 2.3, such as DialogRNN (Majumder et al., 2019), or CMN (Poria et al., 2017) do not apply to our data since they were explicitly designed for at least two speakers of the dialog. Our data from both experiments contain only the participant's speech. However, the state-of-the-art model Emocaps (Li et al., 2022a) is applicable for the modelization of the emotional states of one speaker. Thus, we used the model Emocaps as a baseline.

As a reminder, Emocaps extracts audio features with the IS13-ComParE configuration file of openSMILE and textual data with BERT. The emotional content from audio is extracted using a Transformer without a decoder (Emoformer block). Textual features are given to the Mapping Network as input and then merged with the sentence vector obtained from BERT. Finally, this merged vector is concatenated with the emotional representation of audio obtained after the Emoformer block and then given to a stack of dense layers. Its output is given to a bi-LSTM to produce contextual representations and classify emotions (Li et al., 2022a). The illustration of Emocaps network is presented in Figure 6.2.

In our case, we adapt this system to predict the outcome of the nudging spoken interactions.

Our approach

Our approach to predict the outcome of the nudging spoken interaction was inspired by the approach of context-aware emotion classification. The idea behind it is that there is a high probability of inter-utterance dependency. Thus, we argue that modeling the context can provide the necessary information for the representation

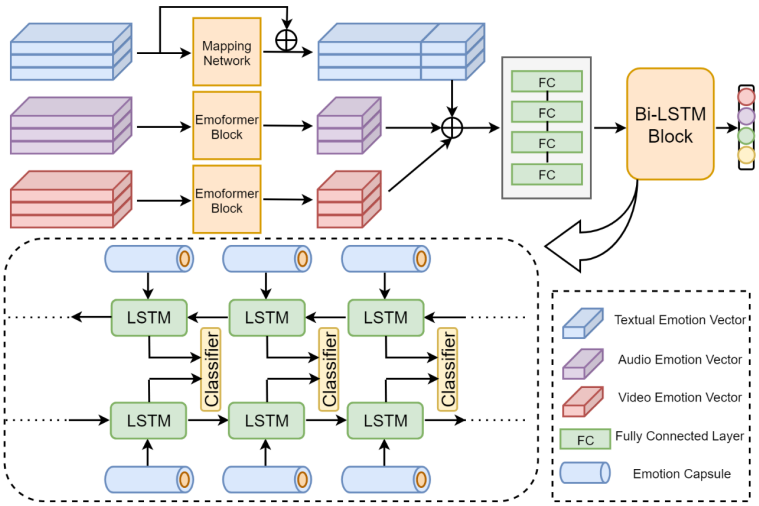


Figure 6.2: Illustration of the Emocaps model, taken from Li et al. (2022a)

of the conversation. We propose to use a LSTM-based system to capture the flow and model the context of the conversation. The main advantage of a LSTM neural network is to handle long-distance dependencies.

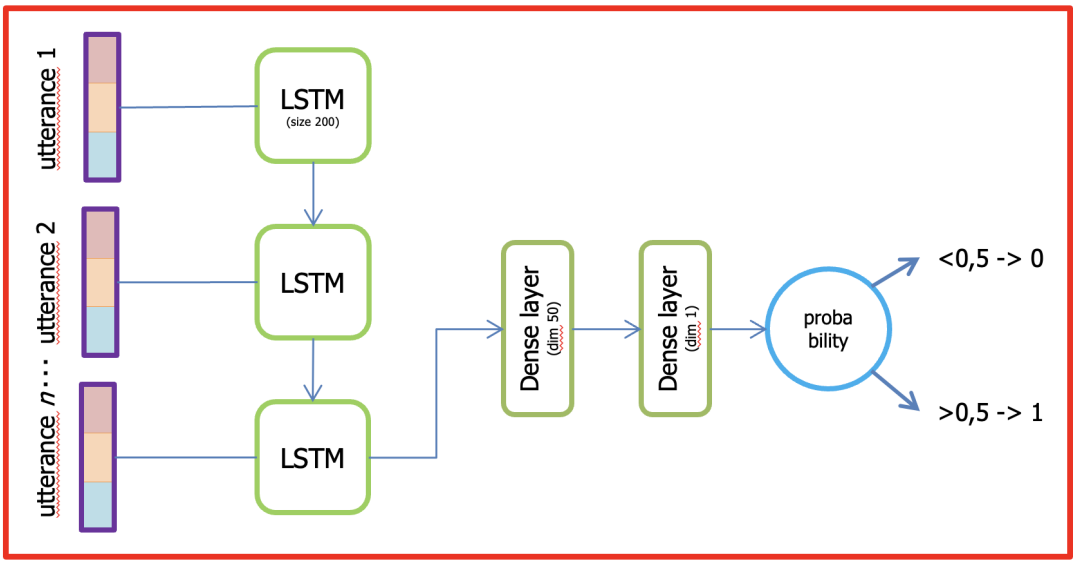


Figure 6.3: Architecture of the model predicting the outcome of the nudging conversation

The baseline version of our model is composed of the following steps (the schema representation of the model is presented in Figure 6.3):

1. Feature extraction (as explained in the previous Section 6.1).

- Audio: two successive linear projections are applied to reduce the dimension of the openSMILE features from 6373 to 200. We apply the Tanh activation function between the two layers.
- Text: we keep the output of CamemBERT of dimension 768.
- Emotion: the emotional representation of the utterance is realized with an embedding layer of dimension 20. Thus, the vector of emotional representation has a dimension of 200 (10 labels x the size of the embedding layer (20)).

2. Concatenation of all features (text, audio, and emotions).
3. Each speaking turn is given to a LSTM with a hidden size of 200.
4. The final representation of a LSTM is given to two successive linear projections of dimensions 50 and 1.
5. The output of the last linear projection is followed by the sigmoid activation function to obtain a probability of the model's prediction.

We apply dropout with a rate of 0.2 after the first linear projection of openSMILE features, on the concatenated vector of multi-modal features, and after the first linear projection of the model's output.

We run the model for 200 epochs and use a learning rate of 0.005.

Variants of the model

We also experimented with a bi-LSTM instead of LSTM. The bi-LSTM also had a hidden size of 200. Thus, the first linear projection of the output reduced the dimension from 400 to 50.

We tested both model's variants on different combinations of features:

- Acoustic features (A);
- Textual features (T);

- Emotional embeddings (E);
- Acoustic + Textual features (A + T);
- Acoustic features + Emotional embeddings (A + E);
- Textual features + Emotional embeddings (T + E);
- Acoustic + Textual features + Emotional embeddings (A + T+ E).

However, since Emocaps was not conceived to consider the emotional embeddings, this model was trained on the following combinations of features:

- Acoustic features (A);
- Textual features (T);
- Acoustic + Textual features (A + T).

Reduction to binary classification

Instead of performing classification over the three classes ("Not-Nudged", "Moderately Nudged", and "Nudged"), we reduced our problem to a series of binary classifications that will be explained in our inference procedure 6.2.1. In each of these binary classification tasks, we subsampled the most represented class in order to have an equal number of training samples from both classes.

We split our data into train and test sets using a random split with 80% of the data being used for the train set and 20% for the test set. Due to the size of our datasets, we did not use the validation set.

6.2.1 Inference procedure

Figure 6.4 summarizes the proposed approach.

We reduced our problem of predicting the outcome of nudging spoken interactions to the series of five binary classifications:

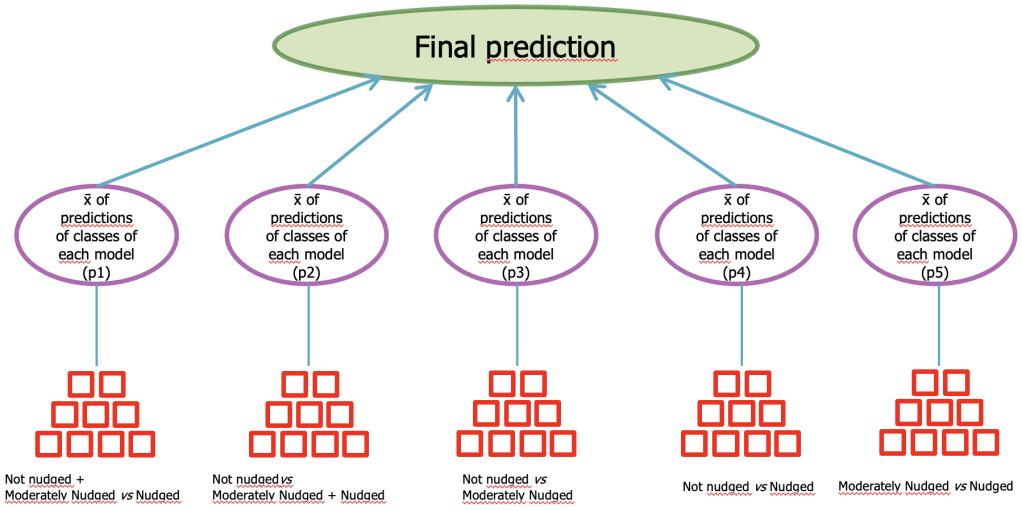


Figure 6.4: Illustration of the inference procedure.

- Group 1: "Not-Nudged" + "Moderately Nudged" *vs* "Nudged";
- Group 2: "Not-Nudged" *vs* "Moderately Nudged" + "Nudged";
- Group 3: "Not-Nudged" *vs* "Moderately Nudged";
- Group 4: "Not-Nudged" *vs* "Nudged";
- Group 5: "Moderately Nudged" *vs* "Nudged".

Due to the size of our dataset, this approach aimed to simplify the prediction task. Moreover, it allowed us to test whether the class "Moderately Nudged" was closer to one of the two other classes.

For each binary classification, we proposed to train several models on the same train set. We trained 9 models for each binary classification when we used multi-modal features and a group of 5 models for each binary classification trained on one modality. This approach is used to improve the predictive performance (Bühlmann, 2012).

Then, one model predicted 0 if the output of the model was lower than 0.5 and 1 otherwise. We assigned 0 to the class with the lowest level out of the two classes in the comparison of the propensity to be nudged and 1 to the class with the highest

level of propensity to be nudged out of the two classes. For each group of models, we got the prediction of all the models of the group and computed the mean.

To get the final probability prediction for each of the three classes, we proposed to compute the product of the likelihoods of the class in the four groups of models where this class appeared and to divide it by the sum of the three results. The class with the highest probability is then predicted. We were inspired by the traditional method of computing the probability of the prediction of an ensemble method. For example, to compute the final probability of the class "Nudged", we compute the product of its likelihoods predicted by groups 1, 2, 4, and 5 and divide the product by the sum of products of the three classes. To avoid a null result (in cases where a model predicted a probability of 0 for a class), we added (or substracted) 0.001 to all the results.

To illustrate this more formally, let l_1, l_2, l_3, l_4 and l_5 be the likelihoods computed by each group. The product of the likelihoods for each class can be expressed as follows:

$$
\begin{aligned}
s(\text{"Not-Nudged"}) &= (1 - l_1) \times (1 - l_2) \times (1 - l_3) \times (1 - l_4) \\
s(\text{"Moderately Nudged"}) &= (1 - l_1) \times l_2 \times l_3 \times (1 - l_5) \\
s(\text{"Nudged"}) &= l_1 \times l_2 \times l_4 \times l_5
\end{aligned}
$$

The likelihood of the class "Nudged" for example is then:

$$
p(\text{"Nudged"}) = \frac{s(\text{"Nudged"})}{s(\text{"Not-Nudged"}) + s(\text{"Moderately Nudged"}) + s(\text{"Nudged"})}
$$

Thus, as an example, to make a prediction on the whole set of features (text + audio + emotion), we trained 5 groups of 9 models of binary classification. The final outcome was then predicted based on these 45 models.

We present the obtained results in the following section.

6.3 Experiments

In this Section, we present the obtained results of the prediction of the nudging spoken interactions' outcome.

The performance of a group of models is evaluated as follows. For each class, we first compute its recall defined as:

$$\frac{TP}{TP + FN}$$

where TP and FN denote the true positives and false negatives for this class, respectively. The final metric is the Unweighted Average Recall (UAR) which computes the mean of the recall of each class. This is done to handle the imbalance between the three classes. If we denote our classes as 0, 1 and 2, we can express the UAR as:

$$\frac{recall_0 + recall_1 + recall_2}{3}$$

We ran each experimental setting three times and the reported results show the mean and standard deviation of the Unweighted Average Recall over these three runs.

Results

The classification results for each of the three classes are reported in Table 6.1 for the data from adults and in Table 6.2 for the data from children.

The baseline model Emocaps predicted worse than random the outcome of the nudging spoken interactions. For both datasets (with adults and with children), this model correctly predicted the class "Moderately Nudged", but it was achieved only by predicting all the test samples as "Moderately Nudged". Thus, we consider that this model is not adapted for our task.

For the data from adults, we observe that the best performance was achieved by a LSTM model trained on emotion embeddings with an recall of 0.73 and a

Model	Features	Not-Nudged	Moderately Nudged	Nudged	UAR
Emocaps	A	0.11 (±0.19)	0.44 (±0.2)	0.37 (±0.17)	0.31 (±0.1)
	T	0 (±0)	**1** (±0)	0 (±0)	0.33 (±0)
	A + T	0.11 (±0.19)	0.56 (±0.51)	0.44 (±0.12)	0.37 (±0.07)
LSTM	A	0.22 (±0.39)	0.44 (±0.39)	0.18 (±0.23)	0.28 (±0.08)
	T	0.22 (±0.19)	**1** (±0)	0.18 (±0.23)	0.46 (±0.02)
	E	0.44 (±0.2)	**1** (±0)	0.74 (±0.13)	**0.73** (±0.06)
	A + T	0.44 (±0.2)	0.44 (±0.2)	0.3 (±0.006)	0.39 (±0.006)
	A + E	**0.55** (±0.2)	0.89 (±0.19)	0.56 (±0)	0.67 (±0.17)
	T + E	0.45 (±0.39)	0.56 (±0.2)	0.56 (±0.2)	0.51 (±0.17)
	A + T + E	0.44 (±0.39)	0.56 (±0.19)	0.59 (±0.06)	0.53 (±0.13)
bi-LSTM	A	0.22 (±0.39)	0.56 (±0.2)	0.44 (±0.3)	0.4 (±0.16)
	T	0.22 (±0.39)	0.78 (±0.19)	0.04 (±0.06)	0.34 (±0.09)
	E	0.22 (±0.19)	0.56 (±0.2)	**0.78** (±0.1)	0.52 (±0.13)
	A + T	0.11 (±0.19)	0.33 (±0)	0.44 (±0.12)	0.29 (±0.09)
	A + E	0.44 (±0.2)	0.78 (±0.19)	0.54 (±0.2)	0.59 (±0.006)
	T + E	0.33 (±0.33)	0.56 (±0.2)	0.59 (±0.17)	0.49 (±0.13)
	A + T + E	0.44 (±0.2)	0.33 (±0.33)	0.7 (±0.06)	0.49 (±0.19)

Table 6.1: We report the accuracy for each class as well as the balanced accuracy. We indicate the means and standard deviations over three runs. Emocaps corresponds to the model proposed by Li et al. (2022a). UAR - Unweighted Average Recall. Data from adults

standard deviation of 0.06. For the bi-LSTM variant of our system, the one trained on acoustic features and emotion embeddings had the highest recall.

Regarding other prediction results for the data from adults, we observed that when groups of models were trained on emotion embeddings (with or without other features), their performance was better than without emotional representation. For comparison:

- A *vs* A + E:
  - LSTM: 0.28 (±0.08) *vs* 0.67 (±0.17)

  - bi-LSTM: 0.4 (±0.16) *vs* 0.59 (±0.006)

- T *vs* T + E:
  - LSTM: 0.46 (±0.02) *vs* 0.51 (±0.17)
  - bi-LSTM: 0.34 (±0.09) *vs* 0.49 (±0.13)

- A + T *vs* A + T + E:
  - LSTM: 0.39 (±0.006) *vs* 0.53 (±0.13)
  - bi-LSTM: 0.29 (±0.09) *vs* 0.49 (±0.19)

Model	**Features**	**Not-Nudged**	**Moderately Nudged**	**Nudged**	**UAR**
Emocaps	A	0.27 (±0.23)	0.17 (±0.29)	0.5 (±0.17)	0.31 (±0.2)
	T	0 (±0)	**1** (±0)	0 (±0)	0.33 (±0)
	A + T	0.33 (±0.12)	0 (±0)	0.56 (±0.1)	0.3 (±0.07)
LSTM	A	0.33 (±0.12)	0.17 (±0.14)	0.44 (±0.2)	0.32 (±0.13)
	T	0.2 (±0.2)	0 (±0)	0.55 (±0.25)	0.25 (±0.12)
	E	0.67 (±0.11)	0.08 (±0.12)	0.72 (±0.09)	0.49 (±0.01)
	A + T	0.47 (±0.12)	0 (±0)	0.45 (±0.25)	0.3 (±0.12)
	A + E	0.4 (±0)	0.25 (±0.25)	0.55 (±0.25)	0.4 (±0.15)
	T + E	0.6 (±0.2)	0.25 (±0.25)	0.5 (±0.29)	0.45 (±0.24)
	A + T + E	0.8 (±0.35)	0.08 (±0.14)	0.72 (±0.34)	**0.53** (±0.09)
bi-LSTM	A	0.27 (±0.3)	0.3 (±0.14)	**0.83** (±0.17)	0.48 (±0.1)
	T	0.47 (±0.3)	0 (±0)	0.55 (±0.25)	0.34 (±0.18)
	E	0.67 (±0.12)	0.17 (±0.14)	0.67 (±0.17)	0.5 (±0.06)
	A + T	0.67 (±0.3)	0.08 (±0.14)	0.39 (±0.19)	0.38 (±0.12)
	A + E	0.66 (±0.3)	0.33 (±0.14)	0.45 (±0.25)	0.48 (±0.05)
	T + E	**0.87** (±0.12)	0.08 (±0.14)	0.55 (±0.25)	0.5 (±0.13)
	A + T + E	0.6 (±0.2)	0 (±0)	0.61 (±0.19)	0.4 (±0.12)

Table 6.2: We report the accuracy for each class as well as the balanced accuracy. We indicate the means and standard deviations over three runs. Emocaps corresponds to the model proposed by Li et al. (2022a). UAR - Unweighted Average Recall. Data from children

For the data from children, the LSTM trained on the complete set of features (acoustic, textual, and emotional) had the highest Unweighted Average Recall of 0.53. For the bi-LSTM variant of the system, the best performance was achieved with the emotion embeddings (UAR = 0.5 (±0.06)) and with emotion embeddings and textual features (UAR = 0.5 (±0.13)).

Similar to the results obtained for the data from adults, adding emotion embeddings to the feature set increased the performance of the groups of models. Moreover, some models trained without emotional contextual representation performed no better than random prediction.

- A *vs* A + E:
  - LSTM: 0.32 (±0.13) *vs* 0.4 (±0.15)
  - bi-LSTM: 0.48 (±0.1) *vs* 0.48 (±0.05)
- T *vs* T + E:
  - LSTM: 0.25 (±0.12) *vs* 0.45 (±0.24)
  - bi-LSTM: 0.34 (±0.18) *vs* 0.5 (±0.13)
- A + T *vs* A + T + E:
  - LSTM: 0.3 (±0.12) *vs* 0.53 (±0.09)
  - bi-LSTM: 0.38 (±0.12) *vs* 0.4 (±0.12)

We note that the recall for the class "Moderately Nudged" was lower than for other classes and compared with the results obtained for the data from adults. We address this issue in the following.

We also observed that the proposed system performed better on the data from adults than on the data from children. Several factors could have influenced these results. First, the children spoke much less than adults, and the diversity of words was limited. Secondly, the configuration feature set IS-13 ComParE was introduced

for the feature extraction of adult voices, and the relevance of the proposed features is not established. Thirdly, as shown in the analysis of the correlation between emotional states and the propensity to be nudged, children seem to be more influenced by the novelty of the situation during the recording session than by the nudges or the type of conversational agent.

These results indicate that the emotional representation of the utterances plays a crucial role in predicting the outcome of the nudging spoken interactions. We can conclude that emotional alignment is a better prediction factor than linguistic and paralinguistic alignments in the automatic detection of nudges in spoken interactions.

Classification of the class "Moderately Nudged"

The separation between classes regarding participants' propensity to be nudged was arbitrary, so we hypothesized that the class of moderately nudged participants could share characteristics with one of the two other classes. To investigate this hypothesis, we analyzed what group of models performed better when combining this class with one of the two other classes for binary classification.

We report the comparative results in Table 6.3.

As predicted by our hypothesis, for the dataset with adults, most of the groups of models performed with higher recall when the samples of the class "Moderately Nudged" (1 in Table 6.3) were combined with the class "Nudged" (2 in Table 6.3). The best performance was achieved by LSTM trained on emotion embeddings: 0.8 ($\pm$0.05), and bi-LSTM trained on the set of all features: 0.8 ($\pm$0.04).

Contrary to adults, children from the group "Moderately Nudged" were closer to the class "Nudged" (2 in Table 6.3) than to the class "Not Nudged". The best prediction was obtained with bi-LSTM trained on the set of all features: 0.78 ($\pm$0.14).

Moreover, when combining the "Moderately Nudged" class with the class "Not-Nudged" for the dataset with adults, and with the class "Nudged" for the dataset with children, the groups of models with the highest Unweighted Average Recall

Model	Features	Data from adults		Data from children	
		0 + 1 *vs* **2**	**0** *vs* **1 + 2**	**0 + 1** *vs* **2**	**0** *vs* **1 + 2**
Emo caps	A	**0.52** (±0.08)	0.44 (±0.14)	0.51 (±0.07)	**0.63** (±0.09)
	T	0.5 (±0)	0.5 (±0)	0.5 (±0)	0.5 (±0)
	A + T	**0.57** (±0.05)	0.51 (±0.15)	**0.55** (±0.14)	0.48 (±0.09)
LSTM	A	**0.46** (±0.09)	0.29 (±0.07)	0.44 (±0.1)	**0.62** (±0.06)
	T	**0.47** (±0.08)	0.46 (±0.09)	0.35 (±0.15)	**0.47** (±0.12)
	E	**0.8** (±0.05)	0.56 (±0.09)	0.62 (±0.08)	**0.7** (±0.07)
	A + T	0.44 (±0.02)	**0.58** (±0.15)	0.48 (±0.11)	**0.53** (±0.06)
	A + E	**0.64** (±0.08)	0.52 (±0.06)	0.57 (±0.1)	**0.65** (±0.08)
	T + E	**0.67** (±0.04)	0.63 (±0.1)	0.58(±0.2)	**0.7** (±0.06)
	A + T + E	**0.7** (±0.15)	0.53 (±0.02)	0.69 (±0.1)	**0.78** (±0.14)
bi-LSTM	A	**0.44** (±0.17)	0.4 (±0.02)	0.6 (±0.09)	**0.63** (±0.13)
	T	0.44 (±0.08)	**0.5** (±0.02)	0.4 (±0.03)	**0.5** (±0.12)
	E	**0.68** (±0.09)	0.54 (±0.03)	0.67 (±0.12)	**0.72** (±0.02)
	A + T	0.44 (±0.08)	**0.58** (±0.03)	**0.57** (±0.03)	0.55 (±0.15)
	A + E	**0.78** (±0.07)	0.56 (±0.14)	0.53 (±0.12)	**0.77** (±0.13)
	T + E	**0.72** (±0.1)	0.54 (±0.14)	0.63 (±0.04)	**0.7** (±0.04)
	A + T + E	**0.8** (±0.04)	0.44 (±0.1)	0.55 (±0.04)	**0.75** (±0.08)

Table 6.3: Comparative results of the class "Moderately Nudged". The results are measured in balanced accuracy. Indications: 0 stands for the class "Not-Nudged", 1 - "Moderately Nudged", 2 - "Nudged". We report the means and standard deviations of Unweighted Average Recall over three runs.

performed better than the groups of models predicting three classes separately.

These results confirmed our hypothesis and go along with the analysis of the correlation between the participant's propensity to be nudged and their emotional states presented in Chapter 4.

Discussion about emotions

As we observed that the affective representation of the conversation is the most useful modality to predict whether a participant was nudged or not, we analyzed whether there were differences in the distribution of affective labels between the

Nudged	**Moderately Nudged**	**Not-Nudged**
Interest 24.2%	Interest 18.6%	Interest 19.6%
Confidence 14.6%	Confidence 15.5%	Confidence 15.7%
Embarrassment 14.2%	Embarrassment 13%	Embarrassment 15.8%
Stress 6.3%	Stress 7.9%	Lack of interest 7.3%
Lack of interest 6%	Lack of interest 7.8%	Stress 6%

Table 6.4: Distribution of emotional labels according to the adult participants' level of propensity to be nudged

Nudged	**Moderately Nudged**	**Not-Nudged**
Interest 37.5%	Interest 37.5%	Interest 37.4%
Stress 32.2%	Stress 33.6%	Stress 31.7%
Neutral 19.3%	Neutral 19.5%	Neutral 21.9%
Anger 11%	Anger 9.2%	Anger 9%

Table 6.5: Distribution of emotional labels according to the children participants' level of propensity to be nudged

different classes.

We report the five most frequent labels of emotions for the dataset with adults for each class in Table 6.4 and for the dataset with children in Table 6.5.

For the dataset with adults, we report the distribution of activation labels in Table 6.6, polarity in Table 6.7, and engagement in Table 6.8.

For the dataset with children, we report the distribution of comprehension in

Label/Class	**Nudged**	**Moderately Nudged**	**Not-Nudged**
1	2.1%	2.7%	1.6%
2	17.9%	31.4%	22.1%
3	47.2%	46.4%	47.2%
4	29.1%	17.8%	25.6%
5	3.8%	1.7%	3.4%

Table 6.6: Distribution of labels of activation according to the adult participants' level of propensity to be nudged

Label/Class	Nudged	Moderately Nudged	Not-Nudged
Positive	61.3%	56.1%	51.06%
Negative	35.6%	43.3%	48.1%
Neutral	2%	0.6%	0.8%

Table 6.7: Distribution of labels of polarity according to the adult participants' level of propensity to be nudged

Label/Class	Nudged	Moderately Nudged	Not-Nudged
-2	3.5%	3.3%	3.3%
-1	11%	8.9%	14.9%
0	17.6%	13.3%	13.5%
1	47.8%	74.4%	58.6%
2	20.1%	-%	9.7%

Table 6.8: Distribution of labels of engagement according to the adult participants' level of propensity to be nudged

Table 6.9, hesitation in Table 6.10, and engagement in Table 6.11.

As we can see, the distribution of emotional labels over the three classes for both datasets is similar.

However, we observe the difference in the distribution of other affective classes. For the data with adults, the main difference is observed in the level of engagement. This observation indicates that the model generalizes the prediction pattern based on these affective levels.

Label/Class	Nudged	Moderately Nudged	Not-Nudged
-1	9.2%	13.5%	7.6%
0	5.6%	8.8%	9.6%
1	85.2%	77.6%	87.8%

Table 6.9: Distribution of labels of comprehension according to the children participants' level of propensity to be nudged

Label/Class	Nudged	Moderately Nudged	Not-Nudged
-1	31.9%	33.7%	23.5%
0	18.2%	11.8%	23.4%
1	59.8%	54.6%	53.1%

Table 6.10: Distribution of labels of hesitation according to the children participants' level of propensity to be nudged

Label/Class	Nudged	Moderately Nudged	Not-Nudged
-1	6.6%	5.4%	8.1%
0	23.9%	27.2%	29.7%
1	69.4%	67.3%	62.1%

Table 6.11: Distribution of labels of engagement according to the children participants' level of propensity to be nudged

Discussion about multi-class classification

To highlight the effectiveness of our strategy, we compared it against the same architecture applied to multi-class classification. For both datasets with adults and children, we chose the best-performing feature set (as presented in Table 6.1 and 6.2). For the dataset with adults, our approach obtained an recall of 0.73, and for the dataset with children an recall of 0.53. Thus, we trained a single LSTM-based model on emotion embeddings of data with adults, and a single LSTM-based model on textual and audio features and emotion embeddings of data with children.

We present the obtained results in Table 6.12.

Model	Not-Nudged	Moderately Nudged	Nudged	UAR
Model 1	0.1 (±0.17)	0.42 (±0.21)	0.89 (±0.21)	0.47 (±0.06)
Model 2	0.53 (±0.12)	0.17 (±0.14)	0.39 (±0.1)	0.36 (±0.11)

Table 6.12: We report the accuracy for each class as well as the balanced accuracy. We indicate the means and standard deviations over three runs of a single model. UAR - Unweighted Average Recall. Model 1 - LSTM-based model trained on emotion embeddings for the dataset with adults; Model 2 - LSTM-based model trained on textual and audio features, and emotion embeddings.

We see that our approach significantly outperforms a simple multi-class classifier. These results show that reducing the task to a binary classification problem allowed a model to generalize the prediction better.

6.4 Argumentation

Validation set

Due to the size of our datasets, we did not validate the proposed hyper-parameters on the validation set. Indeed, the class "Not-Nudged" of the dataset with adults contains only 7 examples in the train set and 3 examples in the test set. If we add the validation set, the train set will contain even less examples, making the training process even more complicated.

Moreover, with a small number of examples, the validation set is not representative of the data. We illustrate this point in Figure 6.5. The green loss curve in 6.5b shows that hyperparameters are not adapted for the data. However, the green loss curve for the test set demonstrates a good performance of the model on the test data.

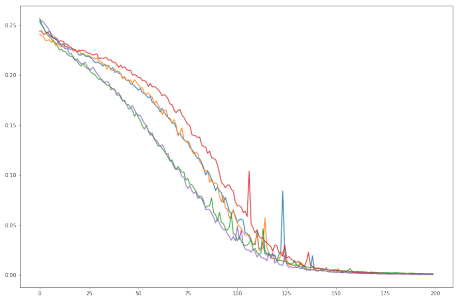
Considering these observations and the particularities of our datasets, we argue that prioritizing more examples in the train set results in a model's better performance.

Cross-validation

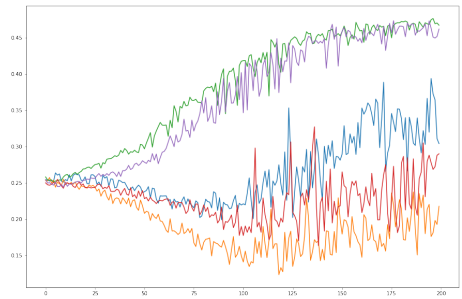
The same reasons can be applied to cross-validation. Models cannot generalize well since there are only a few examples in datasets. Therefore, the models are not stable and the results of the cross-validation highly depend on the split of data.

We tested the following parameters for the cross-validation:

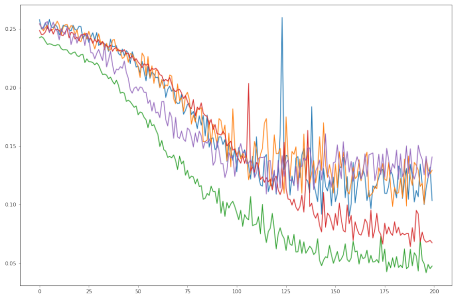
- Embedding of affective labels: 10, 20, 50
- Size of the vector of audio features: 100, 200, 500
- Size of the vector of textual features: 50, 100, 200, 400



(a) Loss curve for the train set.



(b) Loss curve for the validation set.



(c) Loss curve for the test set.

Figure 6.5: Loss curves for the train, validation and test sets.

For the dataset with adults, the best performance for the cross-validation was obtained for the combination of 50 affective embeddings, the vector of audio features of size 200, and the vector of textual features of size 50. However, models trained with these parameters obtained a worse performance compared to our approach (Table 6.13).

Model	**Not-Nudged**	**Moderately Nudged**	**Nudged**	**UAR**
Best configuration according to the cross-validation	0.22 (±0.39)	0.67 (±0.58)	0.63 (±0.26)	0.5 (±0.21)
Our approach	0.44 (±0.39)	0.56 (±0.19)	0.59 (±0.06)	0.53 (±0.13)

Table 6.13: Comparison of performances of the best configuration according to the cross-validation and our approach. The models were trained on the three modalities.

Moreover, the comparison of the performance of each analyzed parameter does not show significant differences, indicating that the particularities of the data play

a more important role in the performance of the models than its parameters.

Textual modality

In this thesis, we proposed to use the pooler output of the CamemBERT. It is the representation of CLS token. This output contains contextualized information of the whole utterance. Another common method is to use the mean of vectors of each token. However, this method is not suitable for our data and it is preferable to use the pooler output (Table 6.14).

Model	**Not-Nudged**	**Moderately Nudged**	**Nudged**	**UAR**
Matrix representation (A + T + E)	0.44 (±0.5)	0.55 (±0.38)	0.48 (±0.23)	0.49 (±0.27)
Our approach (A + T + E)	0.44 (±0.39)	0.56 (±0.19)	0.59 (±0.06)	0.53 (±0.13)
Matrix representation (T)	0 (±0)	0.44 (±0.2)	0.67 (±0.11)	0.37 (±0.04)
Our approach (T)	0.22 (±0.19)	1 (±0)	0.18 (±0.23)	0.46 (±0.02)

Table 6.14: Comparison of performances of models trained on the mean of vectors of each token (matrix representation) and the pooler output (our approach).

6.5 Discussion

In this chapter, we propose a system that automatically predicts the outcome of the nudging spoken interactions based on acoustic and textual features and emotion embeddings.

The main challenge for this task was the imbalance of classes in the datasets. We handled this issue by subsampling the most represented class and by reducing the prediction problem to the binary classification task. Moreover, we proposed to train multiple groups of binary classification models to handle the problem of datasets' small size.

We found that affective representation of the conversation was crucial for the model's performance for both datasets. Thus, the best performances were obtained by the groups of models that were based on emotion embeddings (with or without linguistic and paralinguistic features). The emotion embeddings are based on the labels of annotation which was realized perceptively without any visual cues. Therefore, we hypothesize that the linguistic and paralinguistic differences are correlated more with the agent's embodiment than with the propensity to be nudged. Moreover, we argue that the affective representation of the conversation contains relevant information for predicting the outcome of nudging spoken interactions.

We also found that the adult participants who were moderately nudged shared characteristics that were close to the characteristics of participants who did not nudge. Inversely to adults, children who were moderately nudged shared more characteristics with children belong to the group "Nudged". This result validated our approach of analyzing data in Chapter 4.

The distribution of affective labels showed that classes of participants regarding their propensity to be nudged were similar at emotional level, but differed mainly at other levels of annotation.

Chapter 7

Conclusion

7.1 Conclusion

In this thesis, we aimed to propose an approach to automatically detecting nudges in spoken human-human and human-machine interactions. To this aim, we proposed a methodology for data acquisition for two audiences: adults and children. We validated the methodology by the statistical analysis of the effectiveness of nudges. The recorded data were annotated on multiple affective levels. We analyzed linguistic and paralinguistic characteristics and emotional cues. Finally, we used the acquired data, annotations, and realized analyses to develop a system of automatic prediction of the outcome of nudging spoken interactions.

Data acquisition. We proposed a methodology for data collection that aimed to measure the effectiveness of nudges in spoken human-human and human-machine interactions of two audiences - adults and children. We also proposed an annotation guide to label recorded data at different affective levels. We recorded 98 adult participants and annotated 5 hours of participants' active speech. The dataset with children contains more than 6 hours from 86 participants.

Effectiveness of nudges in spoken interactions. For each dataset, we proposed metrics to estimate whether the participant was nudged. We used these

metrics to classify participants according to their propensity to be nudged. We analyzed whether nudges made statistically significant changes in participants' scores in general, regarding the type of their interlocutor and the type of nudge.

The statistical analysis showed that adults were especially impacted by the questions where they had not established their habits yet or the changes did not require much effort. All types of conversational agents impacted their choices, but with the human agent, they significantly changed their scores for more questions. The statistical analysis also showed that adults were more influenced by the nudges based on emotions and steering them against certain ecological behaviors. We can conclude that affect had a great impact on participants' decision-making.

Nudging interaction also impacted participants' interest in environmental issues, measured by their willingness to spend time and money on ecological problems. Regardless of the conversational agent or the type of nudge, participants were willing to spend more additional money but less time after the nudging intervention.

The analysis of the correlation of character traits with the propensity to be nudged showed that participants with a strong trait of agreeableness changed their scores by more points.

However, children were not significantly influenced by the nudging intervention. We suggest that the unusual environment during the recording had a more important effect on children's behavior than the nudge. Even though the type of their interlocutor did not impact their answers to nudges, the attitude of the children towards the different agents was not similar. Thus, we established that they chose significantly more often to interact with a machine than with a human when having the choice. Moreover, children lied significantly more often to the human agent than to one of the machine agents.

Emotional alignment. To better understand the emotional component of nudging spoken interactions, we proposed to analyze the correlation between participants' emotional states and 1) the type of their interlocutor, 2) the type of nudge,

and 3) their propensity to be nudged.

For both datasets, we found that participants speaking to the human agent were more amused and interested than those speaking to machines. Moreover, children, more than adults, expressed similar emotions towards the robot and the smart-speaker agents.

Nudged participants showed more interest (in a dataset with adults) and more anger (in a dataset with children) than not-nudged participants. Moreover, children who were suggested to keep less balls for themselves were also characterized by anger and less stress. Adults, in their turn, were perceived as surprised, relieved, and with a lack of interest after the nudge with negative influence.

Lexical and paralinguistic alignment. We investigated the differences in the participants' speech with paralinguistic and lexical measures. These cues showed that participants of both audiences mainly distinguished between the human agent and the machine agents. Thus, adults and children who spoke to the machine agents had a lower speech rate and less frequent use of disfluencies. With such a specific topic as ecology, the lexical choices of adult participants were restricted. Nevertheless, they produced more complex and longer utterances when addressing the human agent.

Automatic prediction of the outcome of nudging spoken interactions. To automatically predict the outcome of the nudging spoken interactions, we proposed a context-aware neural network with a LSTM that is based on textual and acoustic features and emotion embeddings. The proposed model handles the issue of class imbalance in small datasets by subsampling datasets, reducing the prediction problem to binary classification, and training multiple models of binary classification. We compared the performance of the approach on different combinations of features and datasets of adults and children. The best performance was achieved when predicting two classes: "Not-Nudged" and "Nudged".

For the dataset with adults, the bi-LSTM model trained on textual and acoustic

features and emotion embeddings and the LSTM trained on emotion embeddings were the best-performing models with an accuracy of 0.8. For the dataset with children, the best prediction was achieved by the LSTM model trained on the textual and acoustic features and emotion embeddings with an accuracy of 0.78.

The results of the proposed system indicate that the detection of nudges in spoken human-human and human-computer interactions can be realized through the analysis of linguistic, paralinguistic and emotional alignment.

7.2 Future research directions

Our research covers an interdisciplinary topic and opens a discussion in the domains of sociology, linguistics, ethics, etc.

One of the future perspectives would be the investigation of the role of engagement in nudging spoken interactions. As presented in this thesis, the prediction of the outcome of nudging spoken interactions based on the cues of participant's engagement could be analyzed. In the same vein, the correlation between hesitation expressed by affect bursts and false starts and the propensity to be nudged opens another path for studying nudges in spoken interactions.

More data is needed for a more robust statistical analysis and a more stable system of nudges detection. Moreover, to develop less arbitrary metrics of the effectiveness of nudges, the data acquisition methodology should be changed to record participants without nudging intervention during spoken interactions.

The proposed system can be combined with continuous emotion recognition in conversation and applied to non-annotated data. Thus, the system could make its prediction and propose some alarm once it has a sufficiently high level of certainty before the end of the conversation.

We tested our methodology on two specific topics (one by audience). Thus, linguistic nudges during spoken interactions might be applied in other domains to

compare results. We can imagine interactions where different conversational agents steer participants to buy something that they do not need, as we have seen that linguistic nudges are common in marketing. Moreover, the results in different domains could indicate a more general effectiveness of linguistic nudges in spoken interactions.

We raised some ethical questions in Chapter 2, which are still unanswered. More general conclusions about the effectiveness of nudges in speech could fill that gap. First, certain standards should be developed for the choice architects. As suggested in the literature review, end-users should be aware of the cognitive biases and all the available options for their choices. Similar to messages indicating that the conversation is being recorded, we can imagine a similar message that prevents users that their decision-making is being impacted. Secondly, similar to nudging interventions in the form of little notes that steer us not to forget our bag on the train, it is possible to create similar notes to raise the consciousness about cognitive biases that influence our decisions. For example, on e-commerce websites, a note in bright colors could explain the loss aversion bias or just highlight "Attention! You are being influenced by a hidden manipulation!" before you see the appealing message that

50 people are currently looking at this item and there are only 10 left with 50% off!

Bibliography

Eyal Aharoni and Alan Fridlund. 2007. Social reactions toward people vs. computers: How mere lables shape interactions. *Computers in Human Behavior*, 23:2175–2189.

Hugues Ali Mehenni. 2023. *'Nudges' dans l'interaction homme-machine : analyse et modélisation d'un agent capable de nudges personnalisés*. Theses, Université Paris-Saclay.

René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies*, 38(4):547–566.

Jodi N. Beggs. 2016. Private-sector nudging: The good, the bad, and the uncertain. In *Nudge Theory in Action: Behavioral Design in Policy and Markets*, pages 125–158, Cham. Springer International Publishing.

Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The digital landscape of nudging: A systematic literature review of empirical research on digital nudges. In *CHI '22: CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Ruth Berman and B. Nir-Sagiv. 2009. Clause-packaging in narratives: A crosslinguistic developmental study. *Crosslinguistic Approaches to the Psychology of Language*, pages 149–162.

Anjali Bhavan, Pankaj Chauhan, Hitkul, and Rajiv Ratn Shah. 2019. Bagged support vector machines for emotion recognition from speech. *Knowledge-Based Systems*, 184:104886.

Paul Boersma and David Weenink. 2024. Praat: doing phonetics by computer [computer program]. http://www.praat.org/. Retrieved: 2024-01-27.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5.

Francesca Bonin, Ronald Böck, and Nick Campbell. 2012. How do we react to context? annotation of individual and group engagement in a video corpus. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 899–903.

Francesca Bonin, Celine de looze, Sucheta Ghosh, Emer Gilmartin, Carl Vogel, Anna Polychroniou, Hugues Salamin, Alessandro Vinciarelli, and Nick Campbell. 2013. Investigating fine temporal dynamics of prosodic and lexical accommodation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Heather Bortfeld and Susan Brennan. 1997. Use and acquisition of idiomatic expressions in referring by native and non-native speakers. *Discourse Processes*, 23:119–147.

Holly Branigan, Martin Pickering, and Alexandra Cleland. 2000. Syntactic coordination in dialogue. *Cognition*, 75:B13–25.

Holly Branigan, Martin Pickering, Jamie Pearson, and Janet McLean. 2010. Linguistic alignment between people and computers. *Journal of Pragmatics*, 42:2355–2368.

Holly Branigan, Martin Pickering, Jamie Pearson, Janet Mclean, and Clifford Nass. 2003. Syntactic alignment between computers and people: The role of belief about mental states. *Cognitive Science - COGSCI*.

Nadia Brashier and Daniel Schacter. 2020. Aging in an era of fake news. *Current Directions in Psychological Science*, 29:096372142091587.

Susan Brennan. 1991. Conversation with and through computers. *User Modeling and User-Adapted Interaction*, 1:67–86.

Peter Bühlmann. 2012. Bagging, boosting and ensemble methods. In *Handbook of Computational Statistics: Concepts and Methods*, pages 985–1022, Berlin, Heidelberg. Springer Berlin Heidelberg.

Denis Burnham, Sebastian Joeffry, and L. Rice. 2010. Computer- and human-directed speech before and after correction. *Thirtheenth Australasian International Conference on Speech Science and Technology*, pages 13–17.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower Provost, Samuel Kim, Jeannette Chang, Sungbok Lee, and Shrikanth Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Ana Caraban, Evangelos Karapanos, Daniel Gonçalves, and Pedro Campos. 2019. 23 ways to nudge: A review of technology-mediated nudging in human-computer interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–15, New York, NY, USA. Association for Computing Machinery.

Martine Caris, H.A. Labuschagne, Mireille Dekker, M.H.H. Kramer, Michiel Agtmael, and Christina Vandenbroucke-Grauls. 2017. Nudging to improve hand hygiene. *Journal of Hospital Infection*, 98.

Marcela Charfuelan, Marc Schröder, and Ingmar Steiner. 2010. Prosody and voice quality of vocal social signals: The case of dominance in scenario meetings. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010*.

Lucy Chipchase, Megan Davidson, Felicity Blackstock, Ros Bye, Peter Clothier, Nerida Klupp, Wendy Nickson, Debbie Turner, and Mark Williams. 2017. Conceptualising and measuring student disengagement in higher education: A synthesis of the literature. *International Journal of Higher Education*, 6(2):31–42.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Michelle Cohn, Bruno Ferenc Segedin, and Georgia Zellou. 2022. Acoustic-phonetic properties of siri- and human-directed speech. *Journal of Phonetics*, 90.

Michelle Cohn and Georgia Zellou. 2021. Prosodic differences in human- and alexa-directed speech, but similar local intelligibility adjustments. *Frontiers in Communication*, 6:675704.

Niko Colnerič and Janez Demšar. 2020. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 11(3):433–446.

Brendan Conway-Smith and Robert L. West. 2022. System-1 and System-2 realized within the Common Model of Cognition. In *AAAI 2022 Fall Symposium*, Arlington, Virginia, United States. Association for the Advancement of Artificial Intelligence.

Benjamin Cowan and Holly Branigan. 2015. Does voice anthropomorphism affect lexical alignment in speech-based human-computer dialogue? In *Proc. Interspeech 2015*, pages 155–159.

Benjamin Cowan, Holly Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human-computer dialogue. *International Journal of Human-Computer Studies*.

Hengchen Dai, Silvia Saccardo, Maria Han, Lily Roh, Naveen Raja, Sitaram Vangala, Hardikkumar Modi, Shital Pandya, and Daniel Croymans. 2021. Behavioral nudges increase covid-19 vaccinations: Two randomized controlled trials. *SSRN Electronic Journal*.

Shelagh Davies and Joshua M. Goldberg. 2006. Clinical aspects of transgender speech feminization and masculinization. In *International Journal of Transgenderism*, volume 9, pages 167–196. Taylor & Francis.

Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. Covarep — a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964.

Agnes Delaborde, Marie Tahon, Claude Barras, and Laurence Devillers. 2009. A wizard-of-oz game for collecting emotional audio data in a children-robot interaction. In *Proceedings of the International Workshop on Affective-Aware Virtual Agents and Social Robots*, AFFINE '09, New York, NY, USA. Association for Computing Machinery.

Bella DePaulo and Lerita M. Coleman. 1986. Talking to children, foreigners, and retarded adults. *Journal of personality and social psychology*, 51 5:945–59.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

P. Dolan, M. Hallsworth, D. Halpern, D. King, R. Metcalfe, and I. Vlaev. 2012. Influencing behaviour: The mindspace way. *Journal of Economic Psychology*, 33(1):264–277.

Francesca Ervas, Artur Gunia, Giuseppe Lorini, Georgi Stojanov, and Bipin Indurkhya. 2022. Fostering safe behaviors via metaphor-based nudging technologies. In *Software Engineering and Formal Methods. SEFM 2021 Collocated Workshops*, pages 53–63, Cham. Springer International Publishing.

Harriet Rosanne Etheredge. 2021. Assessing global organ donation policies: Opt-in vs opt-out. *Risk management and healthcare policy*, 14:1985–1998.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Anna Fenko, Teun Keizer, and Ad Pruyn. 2017. Do social proof and scarcity work in the online context? In *ICORIA 2017*.

Charles A. Ferguson. 1982. Simplified registers and linguistic theory. In Loraine K. Obler and Lisa Menn, editors, *Exceptional Language and Linguistics*, pages 49–66. Academic Press, New York.

Kerstin Fischer. 2011. How people talk with robots: Designing dialog to reduce user uncertainty. *AI Magazine*, 32:31–38.

Kerstin Fischer, Kilian Foth, Katharina J. Rohlfing, and Britta Wrede. 2011. Mindful tutors: Linguistic choice and action demonstration in speech to infants and a simulated robot. *Interaction Studies*, 12:134–161.

B. J. Fogg and Clifford Nass. 1997. Silicon sycophants: the effects of computers that flatter. *Int. J. Hum. Comput. Stud.*, 46:551–561.

Simon Garrod and Anthony Anderson. 1987. Saying what you mean in dialogue: A study in conceptual and semantic coordination. *Cognition*, 27:181–218.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation Learning for Speech Emotion Recognition. In *Proc. Interspeech 2016*, pages 3603–3607.

Marcel Gohsen, Johannes Kiesel, Mariam Korashi, Jan Ehlers, and Benno Stein. 2023. Guiding oral conversations: How to nudge users towards asking questions? In *CHIIR '23: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*, pages 34–42.

Li Gong. 2008. How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior*, 24(4):1494–1509. Including the Special Issue: Integration of Human Factors in Networked Computing.

Pelle Hansen and Andreas Jespersen. 2013. Nudge and the manipulation of choice: A framework for the responsible use of the nudge approach to behaviour change in public policy. *European Journal of Risk Regulation*, 1.

Martie G. Haselton, Daniel Nettle, and Damian R. Murray. 2015. *The Evolution of Cognitive Bias*, chapter 41. John Wiley & Sons, Ltd.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, New Orleans, Louisiana. Association for Computational Linguistics.

Ralph Hertwig and Till Grüne-Yanoff. 2017. Nudging and boosting: Steering or empowering good decisions. *Perspectives on Psychological Science*, 12:973 – 986.

Cesar Hidalgo, Diana Orghian, Jordi Canals, Filipa de Almeida, and Natalia Martin. 2021. *How Humans Judge Machines*. The MIT Press.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Dias Issa, M. Fatih Demirci, and Adnan Yazici. 2020. Speech emotion recognition with deep convolutional neural networks. *Biomedical Signal Processing and Control*, 59:101894.

Richard David Jacques. 1996. *The nature of engagement and its role in hypermedia evaluation and design.* Ph.D. thesis, South Bank University.

Wendy Ju and Leila Takayama. 2009. Approachability: How people interpret automatic door movement as gesture. *International Journal of Design*, 3.

Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux, New York: New York.

Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers. 2023a. The effect of human-likeliness in french robot-directed speech: A study of speech rate and fluency. In *Text, Speech, and Dialogue: 26th International Conference, TSD 2023, Pilsen, Czech Republic, September 4–6, 2023, Proceedings*, page 249–257, Berlin, Heidelberg. Springer-Verlag.

Natalia Kalashnikova, Mathilde Hutin, Ioana Vasilescu, and Laurence Devillers. 2023b. Pitch in french human-machine and human-human interactions. do we speak to robots looking like humans as we speak to humans? In *25th ACM International Conference on Multimodal Interaction, ICMI 2023, Paris, France, October 9–13, 2023, Proceedings.*

Natalia Kalashnikova, Serge Pajak, Fabrice Le Guel, Ioana Vasilescu, Gemma Serrano, and Laurence Devillers. 2022. Corpus design for studying linguistic nudges

in human-computer spoken interactions. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4079–4087, Marseille, France. European Language Resources Association.

Natalia Kalashnikova, Ioana Vasilescu, and Laurence Devillers. 2024. Linguistic nudges and verbal interaction with robots, smart-speakers, and humans. In *Proceedings of the Fourteenth Language Resources and Evaluation Conference*, Turin, Italy. European Language Resources Association.

Seiya Kawano, Muteki Arioka, Akishige Yuguchi, Kenta Yamamoto, Koji Inoue, Tatsuya Kawahara, Satoshi Nakamura, and Koichiro Yoshino. 2022. Multimodal Persuasive Dialogue Corpus using Teleoperated Android. In *Proc. Interspeech 2022*, pages 2308–2312.

Kristin Kostick-Quenet and Sara Gerke. 2022. Ai in the hands of imperfect users. *npj Digital Medicine*, 5.

Sören Krach, Frank Hegel, Britta Wrede, Gerhard Sagerer, Ferdinand Binkofski, and Tilo Kircher. 2008. Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS ONE*, 3(7):1494–1509. Including the Special Issue: Integration of Human Factors in Networked Computing.

Bernhard Kratzwald, Suzana Ilić, Mathias Kraus, Stefan Feuerriegel, and Helmut Prendinger. 2018. Deep learning for affective computing: Text-based emotion recognition in decision support. *Decision Support Systems*, 115:24–35.

Sarah Kriz, Gregory Anderson, Magdalena Bugajska, and J. Gregory Trafton. 2009. Robot-directed speech as a means of exploring conceptualizations of robots. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pages 271–272, New York, NY, USA. Association for Computing Machinery.

Sarah Kriz, Gregory Anderson, and J. Gregory Trafton. 2010. Robot-directed speech: Using language to assess first-time users' conceptualizations of a robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 267–274.

Jacek Kudera, Katharina Zahner-Ritter, Jakob Engel, Nathalie Elsässer, Philipp Hutmacher, and Carolin Worstbrock. 2023. Speech Enhancement Patterns in Human-Robot Interaction: A Cross-Linguistic Perspective. In *Proc. INTERSPEECH 2023*, pages 4783–4787.

Leonhard K. Lades and Liam Delaney. 2020. Nudge for good. *Behavioural Public Policy*, 6:75 – 94.

Fabrice Le Guel, Théo Marquis, and Serge Pajak. forthcoming. Bad nudge, kids and voice assistants: A social preferences lab-in-the-field experiment.

Johannes Leder and Astrid Schütz. 2018. Dictator game. In *Encyclopedia of Personality and Individual Differences*, pages 1–4, Cham. Springer International Publishing.

Soëlie Lerch, Patrice Bellot, Emmanuel Bruno, and Elisabeth Murisasco. 2024. Emolis app et dataset pour suggérer des dessins animés proches émotionnellement. In *COnférence en Recherche d'Information et Applications*.

Jonathan Levav and Gavan J. Fitzsimons. 2006. When questions change behavior. *Psychological Science*, 17:207 – 213.

Zaijing Li, Fengxiao Tang, Ming Zhao, and Yusen Zhu. 2022a. EmoCaps: Emotion capsule based model for conversational emotion recognition. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1610–1618, Dublin, Ireland. Association for Computational Linguistics.

Ziming Li, Yan Zhou, Weibo Zhang, Yaxin Liu, Chuanpeng Yang, Zheng Lian, and Songlin Hu. 2022b. AMOA: Global acoustic feature enhanced modal-order-aware network for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7136–7146, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yi-Lin Lin and Gang Wei. 2005. Speech emotion recognition based on hmm and svm. In *2005 International Conference on Machine Learning and Cybernetics*, volume 8, pages 4898–4901 Vol. 8.

Zijie Lin, Bin Liang, Yunfei Long, Yixue Dang, Min Yang, Min Zhang, and Ruifeng Xu. 2022. Modeling intra- and inter-modal relations: Hierarchical graph contrastive learning for multimodal sentiment analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7124–7135, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors.

Rebecca Lunsford, Sharon Oviatt, and Alexander Arthur. 2006. Toward open-microphone engagement for multiparty interactions. In *Proceedings of the 8th International Conference on Multimodal Interfaces, ICMI 2006*, pages 273–280.

Yu Luo, Andrew Li, Dilip Soman, and Jiaying Zhao. 2022. A meta-analytic cognitive framework of nudge and sludge. *Royal Society Open Science*, 10.

Manon Macary, Marie Tahon, Yannick Estève, and Anthony Rousseau. 2021. On the use of self-supervised pre-trained acoustic and linguistic features for continuous speech emotion recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 373–380.

Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. 2021. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. In *IEEE Transactions on Affective Computing*, volume 14, pages 2276–2289.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: an attentive rnn for emotion detection in conversations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Catherine Mayo, Vincent Aubanel, and Martin Cooke. 2012. Effect of prosodic changes on speech intelligibility. In *Proc. Interspeech 2012*, pages 1708–1711.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

H. Ali Mehenni, S. Kobylyanskaya, I. Vasilescu, and L. Devillers. 2020. Nudges with conversational agents and social robots: A first experiment with children at a primary school. *Conversational Dialogue Systems for the Next Decade*, 704:257–270.

Christian Meske and Ireti Amojo. 2020. Ethical guidelines for the construction of digital nudges. In *Hawaii International Conference on System Sciences*.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Jane Mulderrig. 2018. Multimodal strategies of emotional governance: a critical analysis of 'nudge' tactics in health policy. *Critical Discourse Studies*, 15(1):39–67.

Md. Nasir, Wei Xia, Bo Xiao, Brian Baucom, Shrikanth S. Narayanan, and Panayiotis G. Georgiou. 2015. Still together?: the role of acoustic features in predicting marital outcome. In *Proc. Interspeech 2015*, pages 2499–2503.

Clifford Nass. 2004. Nass, c.: Etiquette equality: exhibitions and expectations of computer politeness. communications of the acm 47(4), 35-37. *Communications of the ACM*, 47.

Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT Press.

Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56:81–103.

Kristine Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence Teleoperators & Virtual Environments*, 12:481–494.

Heather L. O'Brien, Ido Roll, Andrea Kampen, and Nilou Davoudi. 2022. Rethinking (dis)engagement in human-computer interaction. *Comput. Hum. Behav.*, 128(C).

Catharine Oertel, Stefan Scherer, and Nick Campbell. 2011. On the use of multimodal cues for the prediction of degrees of involvement in spontaneous con-

versation. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 1541–1544.

Maxim OldenBeek, Till Winkler, Julie Buhl-Wiggers, and Daniel Hardt. 2019. Nudging in blended learning: Evaluation of email-based progress feedback in a flipped classroom information systems course. In *European Conference on Information Systems (ECIS)*.

Enriko Panai and Laurence Devillers. 2023. How ai-augmented nudges may impact eu consumer in a moral cituation? In *Governance of Artificial Intelligence in the European Union - What Place for Consumer Protection?*, pages 366–384. BRUYLANT.

Amit Kumar Pandey, Rodolphe Gelin, Rachid Alami, Renaud Viry, Axel Buendia, Roland Meertens, Mohamed Chetouani, Laurence Devillers, Marie Tahon, David Filliat, Yves Grenier, Mounira Maazaoui, Abderrahmane Kheddar, Frédéric Lerasle, and Laurent Fitte-Duval. 2014. Romeo2 project: Humanoid robot assistant and companion for everyday life: I. situation assessment for social intelligence. *CEUR Workshop Proceedings*, 1315.

Jennifer Pardo. 2006. On phonetic convergence during conversational interaction. In *The Journal of the Acoustical Society of America*, volume 119, pages 2382–2393.

Jamie Pearson, Jiang Hu, Holly P. Branigan, Martin J. Pickering, and Clifford I. Nass. 2006. Adaptive language behavior in hci: How expectations and beliefs about a system affect users' word choice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, page 1177–1180, New York, NY, USA. Association for Computing Machinery.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cour-

napeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Arthur Pellet-Rostaing, Roxane Bertrand, Auriane Boudin, Stéphane Rauzy, and Philippe Blache. 2023. A multimodal approach for modeling engagement in conversation. *Frontiers in Computer Science*, 5.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Jenefer Philp and Susan Duchesne. 2016. Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, 36:50 – 72.

Martin Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *The Behavioral and brain sciences*, 27:169–90; discussion 190.

Soujanya Poria, Erik Cambria, and Alexander Gelbukh. 2015. Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis. In *EMNLP*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Jean A. Pratt, Karina Hauser, Zsolt Ugray, and Olga Patterson. 2007. Looking at human–computer interface design: Effects of ethnicity in computer agents. *Interacting with Computers*, 19(4):512–523.

Eran Raveh, Ingmar Steiner, Ingo Siegert, Iona Gessinger, and Bernd Möbius. 2019. Comparing phonetic changes in computer-directed and human-directed speech. In *Elektronische Sprachsignalverarbeitung 2019*, pages 42–49.

Riccardo Rebonato. 2013. A critical assessment of libertarian paternalism. *SSRN Electronic Journal*.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Center for the Study of Language and Information; Cambridge University Press.

Xuezhu Ren, Yan Tong, Peng Peng, and Tengfei Wang. 2020. Critical thinking predicts academic performance beyond general cognitive ability: Evidence from adults and children. *Intelligence*, 82:101487.

Fabien Ringeval, Shahin Amiriparian, Florian Eyben, Klaus Scherer, and Björn Schuller. 2014. Emotion recognition in the wild. In *ICMI '14: Proceedings of the 16th International Conference on Multimodal Interaction*, pages 473–480.

Fabien Ringeval, Andreas Sonderegger, Jürgen S. Sauer, and Denis Lalanne. 2013. Introducing the recola multimodal corpus of remote collaborative and affective interactions. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8.

Viktor Rozgić, Sankaranarayanan Ananthakrishnan, Shirin Saleem, Rohit Kumar, and Rohit Prasad. 2012. Ensemble of svm trees for multimodal emotion recognition. In *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1–4.

Carmen Sabater. 2017. Linguistic accommodation in online communication: The role of language and gender. *Revista Signos*, 50:265–286.

Alessandra Maria Sabelli, Takayuki Kanda, and Norihiro Hagita. 2011. A conversational robot in an elderly care center: an ethnographic study. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, page 37–44, New York, NY, USA. Association for Computing Machinery.

Adam Sacarny, Michael L. Barnett, Jackson Le, Frank Tetkoski, David Yokum, and Shantanu Agrawal. 2018. Effect of peer comparison letters for high-volume primary care prescribers of quetiapine in older and disabled adults: A randomized clinical trial. *JAMA Psychiatry*, 75(10):1003–1011. Publisher Copyright: © 2018 American Medical Association. All rights reserved.

Yashar Saghai. 2013. Salvaging the concept of nudge: Table 1. *Journal of Medical Ethics*, 39(8):487–493.

S. Sasaki, T. Saito, and F. Ohtake. 2022. Nudges for covid-19 voluntary vaccination: How to explain peer information? *Social science & medicine*, 292.

Beth E Scott, Val Curtis, Tamer Samah Rabie, and Nana Garbrah-Aidoo. 2007. Health in our hands, but not in our heads: understanding hygiene motivation in ghana. *Health policy and planning*, 22 4:225–33.

Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

Setsuko Shirai. Praat scripted ressources. http://phonetics.linguistics.ucla.edu/facilities/acoustic/praat.html. Accessed: 2024-02-22.

Ingo Siegert and Julia Krüger. 2021. "speech melody and speech content didn't fit together"—differences in speech behavior for device directed and human directed interactions. In *Advances in Data Science: Methodologies and Applications*, pages 65–95. Springer International Publishing.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category - elan and iso dcr. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Catherine Snow. 1977. Mothers"speech research: From input to interaction. *Cambridge University Press*.

Zhongkai Sun, Prathusha Kameswara Sarma, William A. Sethares, and Yingyu Liang. 2019. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *AAAI Conference on Artificial Intelligence*.

Cass Sunstein. 2013. Impersonal default rules vs. active choices vs. personalized default rules: A triptych. *SSRN Electronic Journal*.

Cass R. Sunstein. 2020. Sludge audits. *Behavioural Public Policy*, page 1–20.

Marie Tahon and Laurence Devillers. 2010. Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Speech Prosody*, Chicago, United States.

Timm Teubner and Antje Graul. 2020. Only one room left! how scarcity cues affect booking intentions on hospitality platforms. *Electron. Commer. Rec. Appl.*, 39(C).

Richard H. Thaler. 2018. Nudge, not sludge. *Science*, 361.

Richard H. Thaler and Cass R. Sunstein. 2008. *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.

Vicki Trowler. 2010. *Student engagement literature review*. The Higher Education Academy.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences.

Anvarjon Tursunov, Mustaqeem Khan, and Soonil Kwon. 2020. Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features. *Sensors*, 20:5212.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Laurence Vidrascu and Laurence Devillers. 2005. Detection of real-life emotions in call centers. In *9th European Conference on Speech Communication and Technology (INTERSPEECH)*, pages 1841–1844.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P. Xing. 2016. Select-additive learning: Improving cross-individual generalization in multimodal sentiment analysis. *CoRR*, abs/1609.05244.

Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words can shift: dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Markus Weinmann, Christoph Schneider, and Jan vom Brocke. 2016. Digital nudging. *Business & Information Systems Engineering*, 58:433 – 436.

Yang Wu, Yanyan Zhao, Hao Yang, Song Chen, Bing Qin, Xiaohuan Cao, and Wenting Zhao. 2022. Sentiment word aware multimodal refinement for multimodal sentiment analysis with ASR errors. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1397–1406, Dublin, Ireland. Association for Computational Linguistics.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018a. Memory fusion network for multi-view sequential learning. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference*

and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018b. Multi-attention recurrent network for human communication comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Haidong Zhang and Yekun Chai. 2021. COIN: Conversational interactive networks for emotion recognition in conversation. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 12–18, Mexico City, Mexico. Association for Computational Linguistics.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press.

Appendix A

Dialogue Script for the Experiment with Adults

This annex presents the nudging strategies with positive and negative influences in the experiment with adults. Some of the seven ecological habits have scientifically proven advantages and disadvantages for the environment. Others use emotional triggers to provoke positive or negative emotions among participants. Within each group, communicating with different conversational agents, participants were divided into two groups: those who received only nudges presenting advantages or positive emotions and those who received only disadvantages or negative emotions.

Nudge with positive influence	Nudge with negative influence
Being a responsible citizen means buying green beans cultivated in France because by doing so, you support local farmers and create social bonds with your neighbors. Your answers indicate that you are already one of the most responsible citizens. On a scale from 1 to 5, how willing would you be to buy green beans cultivated in France?	The French soil is unsuitable for cultivating green beans. Farmers must use many pesticides, which can enter your body by skin, eyes, and respiration. This can cause cognitive dysfunctions, respiratory illnesses, and other health problems. On a scale from 1 to 5, how willing would you be to buy green beans cultivated in France?
A 2018 publication showed that in a year, there are more animals for human consumption than the total number of humans and wildlife. On a scale from 1 to 5, how willing would you be to replace some of your meat consumption with plant-based proteins, such as soy?	Deforestation caused by soy production leads to the destruction of the habitats of many wild species and increases the likelihood of the appearance of new diseases, such as Covid-19. On a scale from 1 to 5, how willing would you be to replace meat with plant proteins such as soy?
The electric car is a good solution to living without fossil fuels. Moreover, the maintenance cost is lower by at least 25%. On a scale from 1 to 5, how willing would you be to buy an electric car?	Electric cars' production is as polluting as gas cars' production. Moreover, we need rare metals to produce electric car batteries, which are hard to recycle. On a scale from 1 to 5, how willing would you be to buy an electric car?
Like electric cars, electric scooters are becoming increasingly popular. For example, a Belgian crossed the border into France for the first time on an electric scooter earlier this year. ***LAUGH*** On a scale from 1 to 5, how willing would you be to use an electric scooter?	Like electric cars, electric scooters are becoming increasingly popular. But have you noticed how dangerous they are? Last summer, for example, a young woman died after being hit by an electric scooter. On a scale from 1 to 5, how willing would you be to use an electric scooter?
Electric cars don't have much range. For a long journey in France, such as Paris-Marseille, ecologically responsible people prefer the train to the plane, as the latter emits 81 times more greenhouse gases. For the same price, on a scale between 1 and 5, how willing would you be to travel on a train in France?	Electric cars don't have much range. It is possible to fall back on trains or planes. The construction of railroads impacts the landscape's biodiversity, limiting access to territories and the reproduction of numerous animal and plant species. For the same price, on a scale from 1 to 5, how willing would you be to travel on a train in France?
The fish on our plates contain microplastics from plastic bottles, particularly cleaning products. Making your own cleaning products helps reduce the amount of plastic ingested by fish, making our food healthier. On a scale from 1 to 5, how willing would you be to consider making your own cleaning products?	Self-made cleaning products don't follow strict standards for harmful product content and can make the water uninhabitable for fish. On a scale from 1 to 5, how willing would you be to consider making your own cleaning products?

Table A.1: Nudges with positive and negative influences for the dataset with adults

Nudge with positive influence	**Nudge with negative influence**
A pregnant whale, whose stomach contained 22 kilograms of plastic waste, washed up on the beach in the Mediterranean. The use of cotton bags reduces the amount of plastic in the oceans. On a scale from 1 to 5, how willing would you be to use cotton shopping bags?	The production of cotton shopping bags is very water-intensive. To recoup its production cost, a cloth bag needs to be used at least 327 times, unlike a plastic bag, which only needs to be used 7 times. On a scale from 1 to 5, how willing would you be to use cotton shopping bags?

Table A.2: Continuation of Table A.1

Appendix B

French extended summary - Synthèse

Les techniques qui influencent indirectement la prise de décision des humains, connues sous le nom de "nudge" (Thaler and Sunstein, 2008), sont peu étudiées dans les interactions parlées. Les nudges linguistiques sont des techniques de manipulation douce fondées sur les biais cognitifs et utilisent les moyens linguistiques pour encourager les changements dans la prise de décision des humains sans aucune restrictions ou sanctions pour leur choix. Addressées directement au destinataire (par exemple, sous forme de lettre ou de note), ces techniques ont prouvé leur efficacité dans plusieurs domaines. Néanmoins, avec la présence de plus en plus répandu des agents conversationnels au quotidien, plusieurs questions se posent sur l'impact du type de l'interlocuteur (comme un robot ou une enceinte connectée) et la réaction de différents types de public (par exemple, les enfants et les adultes) aux nudges. En tenant compte de cette connaissance préalable, nous étudions plusieurs descripteurs linguistiques (complexité des énoncés, proportion de mots uniques, prénoms, mots lexicaux, etc.) et paralinguistiques (fréquence fondamentale, intensité, débit de parole, durée de l'énoncé, fréquence de disfluences) et posons la question de la pertinence d'un modèle qui prédit si quelqu'un a été verbalement manipulé.

Les recherches dans ce domaine en sont encore à leurs débuts, nous proposons donc d'abord une méthodologie innovative pour la collection de données dans le

but d'estimer la propension des participants à être nudgés (influencés). Nous avons testé deux types de publics : les enfants et les adultes. Le protocole compare les interactions contenant une intervention qui influence le choix (nudge ou incitation) avec trois agents conversationnels (robot Pepper, enceinte Google Home, humain). Dans l'expérience avec les adultes, nous avons comparé les scores des participants quant à leur volonté d'adopter des habitudes écologiques après le nudge avec leurs scores de base afin de mesurer l'influence des nudges. Dans l'expérience avec les enfants, nous avons comparé le nombre de billes qu'ils étaient prêts à garder pour eux après le nudge avec le nombre de billes qu'ils voulaient garder avant le nudge pendant le jeu. En utilisant cette méthodologie, nous avons enregistré 22 heures d'échanges entre des adultes et trois agents conversationnels (le robot Pepper, le haut-parleur Google Home et un humain) et 10 heures d'échanges entre des enfants et les mêmes agents conversationnels.

Dans un premier temps, ces données ont été transcrites manuellement et segmentées en tours de parole, puis annotées à différents niveaux affectifs (émotions, valence, engagement, activation). Deuxièmement, pour mesurer la capacité des différents agents conversationnels à donner des nudges de manière efficace, nous avons analysé la prise de décision des participants en fonction de l'interlocuteur et du type de nudge. De plus, nous avons étudié la corrélation entre les états émotionnels des participants et leurs réponses aux nudges et aux agents conversationnels. Troisièmement, pour mieux comprendre comment l'incarnation d'un agent conversationnel peut influencer la propension d'un participant à recevoir des encouragements, nous avons proposé une comparaison de certains éléments paralinguistiques, lexicaux et discursifs pertinents des participants selon le type d'agent conversationnel. Ces analyses ont montré que les deux types de publics font une distinction entre l'humain et les machines aux niveaux de l'expression des émotions ainsi que des patrons linguistiques au sens large. Enfin, nous avons utilisé différentes combinaisons d'annotations émotionnelles, de transcriptions et de données audio provenant des ex-

périences enregistrées pour construire un modèle d'apprentissage profond basé sur des caractéristiques acoustiques (extraites avec openSMILE (Eyben et al., 2010)), textuelles (extraites avec camemBERT (Martin et al., 2020)) et des états émotionnels afin de prédire si le participant a été nudgé. Pour répondre à nos objectifs de recherche et aux défis qui se sont posés pour cette tâche, nous avons proposé de réduire le problème de classification à une série de classifications binaires. Ensuite, de sous-échantillonner la classe majoritaire du corpus des adultes pour limiter les biais lors de la classification binaire. Et, enfin, d'appliquer la stratégie d'ensemble pour entraîner plusieurs modèles identiques rassemblés pour chaque classification binaire. Nous avons démontré que l'analyse des états émotionnels joue un rôle crucial pour prédire si la personne a été nudgée ou pas. Les principaux résultats soulignent que nos participants ont été nudgés quel que soit leur groupe d'âge, avec un effet plus important sur les adultes.