[Page de garde]

HEC MONTRÉAL

Vers l'adoption des interfaces utilisateurs conversationnelles: est-ce qu'une voix chaleureuse peut améliorer l'expérience utilisateur en cas d'échec?

par

Félix Le Pailleur

Sciences de la gestion
(Option Expérience utilisateur en contexte d'affaires)

Mémoire présenté en vue de l'obtention
du grade de maîtrise ès sciences en gestion
(M. Sc.)

Septembre 2020

Résumé

Ce mémoire par articles s'intéresse à l'évaluation de l'expérience utilisateur lors d'interactions avec des Interfaces Utilisateurs Conversationnelles (IUC). Ce sont des produits ayant la caractéristique de ne disposer d'aucune interface physique et dont l'intelligence artificielle permet la reconnaissance du langage naturel. Parmi eux, les plus populaires sont Alexa d'Amazon, l'Assistant Google, Siri d'Apple ainsi que Cortana de Microsoft. Bien que le nombre de détenteur d'enceintes intelligentes soit estimé à 83.1 millions (He, 2018), la recherche démontre que les utilisateurs font régulièrement face à de la frustration liée aux erreurs et contraintes technologiques uniques liées à ces produits.

Ce mémoire propose tout d'abord, une approche multiméthodes qui, par la triangulation des mesures psychométriques et physiologiques, vise à développer une compréhension plus riche de ce que les utilisateurs vivent lors d'une interaction avec un IUC. Nous avons réalisé, en laboratoire, une étude exploratoire dans laquelle les participants devaient interagir avec une application de nouvelles quotidiennes, via Alexa. Lors de cette expérience, nous avons volontairement conçu des tâches impossibles à effectuer, pour but de faire varier le niveau de frustration ainsi que l'état émotionnel des participants. Nos résultats suggèrent que les mesures psychométriques et physiologiques capturent une partie distincte de la variance émotionnelle des utilisateurs lors de l'interaction avec l'IUC.

Ensuite, dans le cadre d'une seconde recherche exploratoire, nous évaluons l'impact de la chaleur émotionnelle dans la voix de l'IUC, soit une dimension sociale présente dans les relations entre les humains, lorsque le participant reçoit une réponse non-pertinente à sa requête. Nos résultats suggèrent que l'utilisation de la chaleur émotionnelle permet d'augmenter l'intensité émotionnelle, la valence émotionnelle, ainsi que la diminuer la frustration perçue de l'utilisateur.

Ce mémoire contribue à la recherche de deux façons. Tout d'abord, il mobilise une nouvelle méthodologie qui, par la triangulation de mesures psychométriques et physiologiques, permet d'éviter plusieurs biais cognitifs, notamment la désirabilité sociale ainsi que le biais de la méthode unique. Ensuite, ce mémoire démontre que la chaleur émotionnelle dans les réponses données par les IUC, peut réduire les effets secondaires, entre autres liés à la frustration, lorsqu'il n'est pas en mesure de donner une réponse adéquate à la requête de l'utilisateur.

Mots clés: Interface Utilisateur Conversationnelle, chaleur émotionnelle, valence émotionnelle, intensité émotionnelle, frustration, expérience utilisateur

Table des matières

Liste des tableaux et des figures

Liste des figures

Premier article

Deuxième article

Liste des abréviations

CUI: « Conversational User Interface »

HCI: « Human-Computer Interaction »

FACS: « Facial Action Coding System »

AFA: « Automated facial analysis »

EDA: « Electrodermal Activity »

NLP: « Natural Language Processing »

SCR: « Skin Conductance Response »

WoZ: « Wizard of Oz »

UX: « User Experience »

UX: Expérience utilisateur

IA: Intelligence Artificielle

IUC: Interface Utilisateur Conversationnelle

CER: Comité d'Éthique en Recherche

Avant-propos

L'autorisation de rédiger ce mémoire par articles a été obtenue auprès de la direction du programme de la Maîtrise ès sciences en gestion en option Expérience utilisateur dans un contexte d'affaires. Ainsi, la rédaction de ce mémoire s'est faite sous la forme de 2 articles. L'accord de tous les coauteurs de ces articles a été obtenu afin de les inclure dans ce mémoire.

En juillet 2019, le comité d'éthique en recherche (CER) de HEC Montréal a approuvé ce projet de recherche.

Le premier article propose une nouvelle méthode qui permet de mesurer l'expérience vécue par les utilisateurs d'assistants vocaux intelligents selon la valence et l'activation émotionnelles, par la triangulation de mesures psychométriques et psychophysiologiques. Il a été présenté lors de la conférence virtuelle HCI au Danemark en juillet 2020 et sera publié dans les actes de la conférence.

Le second article destiné à être publié à la conférence SIGHCI, s'intéresse à l'importance de la chaleur émotionnelle dans les réponses données par un assistant vocal intelligent. Pour ce faire, nous avons mesuré l'impact de la chaleur émotionnelle sur la valence, l'activation ainsi que sur le niveau de frustration, à l'aide de la combinaison de données psychométriques et psychophysiologiques.

Remerciements

J'aimerais tout d'abord remercier mes codirecteurs de maîtrise, Sylvain Sénécal et Pierre-Majorique Léger qui, par leur mentorat exceptionnel, m'ont inculqué une rigueur ainsi qu'une nouvelle façon d'organiser mes idées, autant au niveau de la création de mon expérience que lors de l'écriture de mes articles. Mes deux années passées au Tech3lab ont été les plus formatrices de mon cheminement scolaire et c'est sans aucun doute, grâce à cette opportunité que je vais pouvoir me lancer sur le marché du travail avec une confiance ainsi qu'un bagage d'expérience hors pair.

Je voudrais aussi remercier l'ensemble des formidables membres de l'équipe du Tech3lab, pour leur constante disponibilité ainsi que leur précieuse aide pour l'ensemble de la réalisation de mon expérience. Aussi, je tiens à remercier les assistants et assistantes de recherche qui permettent aux activités quotidiennes de prendre place au laboratoire et sans eux, je n'aurais pu monter un projet d'une telle envergure. Je tiens aussi à remercier Shang Lin pour son aide lors de l'analyse statistique de mes données.

Ensuite, je tiens à remercier le Conseil de Recherches en Sciences Naturelles et en Génie du Canada (CRSNG), la Chaire de recherche industrielle CRSNG-Prompt en expérience utilisateur (UX) ainsi que la société Radio-Canada. Leur soutien financier m'a permis de réaliser mes deux années à la maîtrise dans un environnement optimal en me donnant la chance de concentrer mon énergie sur la réalisation de ce mémoire.

De plus, je tiens à faire quelques mentions spéciales. Tout d'abord, à Bo Huang, avec qui j'ai eu l'honneur de collaborer sur mon projet et de bénéficier de ses précieux conseils et de sa rétroaction sur mon écriture; tu as été pour moi un mentor, mais surtout un ami avec qui j'ai pu vivre les hauts et les bas découlant de cette aventure. Finalement, à Félix Giroux, m'ayant initié au monde de l'expérience utilisateur, pour son précieux support moral et académique

Finalement, je ne pourrais passer sous silence la contribution de ma magnifique copine Isabelle Bourque qui été une source de motivation tout au long de ce projet ainsi que les deux êtres humains qui me sont les plus chers, mes modèles personnels et professionnels, soit mes

parents Christine Montgrain et François Tremblay, qui par leur constant support m'ont permis d'entreprendre et de mener à terme cette importante étape de vie.

Introduction

Justification du contexte de l'étude

Les plus récentes avancées technologiques dans les domaines de l'intelligence artificielle et de l'apprentissage automatique permettent l'essor d'un nouveau mouvement de produits électroniques sans interfaces utilisant des IUC, soit les assistants vocaux intelligents. En effet, il est de plus en plus courant pour des consommateurs d'utiliser des commandes vocales dans leur quotidien afin d'interagir avec des IUC, via entre autres, d'enceintes Bluetooth et de téléphones intelligents, les plus populaires étant Alexa, Google Assistant, Cortana et Siri (Hoy, 2018).

Les IUC se différencient grandement des premiers systèmes utilisant la reconnaissance vocale, autrefois limités par une capacité maximale de fonctions intégrées, puisqu'ils font désormais partie de l'Internet des objets. Ainsi, étant désormais connectés à l'Internet, ces appareils enregistrent désormais la voix des utilisateurs, laquelle est envoyée vers un serveur informatisé externe faisant l'analyse et l'interprétation de la commande, afin que le IUC offre une réponse adéquate, et ce, en quelques secondes (Hoy, 2018).

Aux États-Unis seulement, on estime que le nombre d'utilisateurs de tels produits dépassera 123 millions d'ici 2021, ce qui représente une augmentation de 44 % depuis 2017 (Petrock, 2019). De plus, une récente étude de marché montre qu'Amazon a vendu pour 75 millions de dollars en haut-parleurs intelligents dans le monde en 2018, soit une croissance de 600 % par rapport à l'année précédente (Tung, 2018). La popularité de ces produits sans interface visuelle a suscité, dans les dernières années, un grand intérêt pour la recherche en ce qui a trait à l'évaluation de l'expérience utilisateur, laquelle, constitue la perception qu'une personne peut avoir lors de l'utilisation ou de l'anticipation d'un produit, système ou service (ISO, 2008).

En plus de leurs capacités de base (ex. : envoyer des messages textes, ajouter un rappel, répondre à des questions d'ordre général), il est possible d'élargir l'éventail de fonctionnalités des IUC par l'installation de « skills », l'équivalent des applications téléchargeables pour les téléphones intelligents. Ainsi, des tierces parties telles que des entreprises ou des particuliers peuvent désormais, grâce à des outils mis à leur disposition (ex: Amazon Developer (Amazon Inc, Seattle, WA)), atteindre leurs consommateurs en leur proposant une alternative entièrement vocale à leurs services traditionnels via par exemple, des jeux, chaînes de radio ou podcasts.

Toutefois, malgré les avancées technologiques et l'important gain en popularité de ces produits, les utilisateurs sont régulièrement confrontés à de nouveaux défis engendrés par l'absence d'interface visuelle, lesquels sont de potentielles sources de frustration (Myers et al. 2018). Par exemple, les trois principales difficultés rapportées sont : 1) Intention non familière (ex: lorsque le participant effectue une requête dont la réponse n'est pas prise en charge par l'appareil, 2) Erreurs liées à l'analyse du langage naturel (ex: lorsque l'appareil comprend la mauvaise requête et 3) Échec du retour d'information (ex: le retour d'information n'est pas clair ou ne répond pas à l'objectif de la requête (Myers et al. 2018).

Plusieurs études récentes se sont donc penchées sur l'interaction d'un utilisateur avec des IUC et sur les facteurs de satisfaction et d'insatisfaction afin d'en améliorer l'expérience (Lopatovska & Oropeza 2018; Lopatovska & Williams 2018, Jiang 2015; Purington et al. 2017; Kiseleva et al. 2016; Myers et al. 2018).

Ainsi, à des fins d'évaluation de l'expérience perçue par les utilisateurs, ces études ont principalement fait l'utilisation de mesures psychométriques de façon rétrospective, c.-à-d., post-tâche. Notamment, elles ont fait l'utilisation d'échelles de mesure, de journaux de bord ou d'entrevues (Jiang 2015; Easwara et al. 2014; Lopatovska and Williams, 2018; Lau et al. 2018; Sciuto et al. 2018; Lopatovska & Oropeza 2018; Porcheron and al. 2018).

Malgré le fait que ces méthodes soient reconnues et largement utilisées afin d'obtenir une grande quantité d'informations sur l'expérience utilisateur, elles peuvent manquer de précision et d'objectivité. En effet, elles peuvent faire face à certains biais cognitifs,

comme la désirabilité sociale. Par exemple, un chercheur qui obtient de la rétroaction d'un participant reflétant ce qui est socialement acceptable selon le contexte, plutôt que le réel fond de ses pensées (Piedmont 2014).

Aujourd'hui, des outils permettent de mesurer l'expérience physiologique vécue par les utilisateurs grâce à des signaux tels que l'activité électrodermale (EDA), le rythme cardiaque, l'oculométrie, la pupillométrie et l'analyse automatisée des expressions faciales. Ces outils sont rétroactifs, donc qu'il permet d'observer l'état émotionnel des participants précisément dans le temps, permettant ainsi d'éviter le biais lié à la rétrospection (De Guinea et al. 2014). De plus, ces outils sont extrêmement précis, non-obstructifs et capturent une dimension émotionnelle distincte, permettant ainsi de compléter les mesures subjectives traditionnellement rapportées (Riedl & Léger 2016).

De plus, une récente étude de Lourties et al. (2018) suggère que l'état émotionnel que les participants vivent n'est pas exactement le même qu'ils rapportent via les données psychométriques. Ainsi, nous nous intéressons à comprendre ce que vivent réellement les utilisateurs au moment de l'interaction avec un IUC. Pour ce faire, nous proposons une nouvelle approche méthodologique, ainsi que des recommandations aux chercheurs et designers de l'industrie leurs permettant de mieux comprendre et améliorer l'expérience utilisateur.

Objectifs et questions de recherche

Ce mémoire s'intéresse à l'évaluation de l'expérience utilisateur lors d'interactions avec des IUC. Ainsi, dans un premier temps, nous explorons les résultats provenant de la triangulation de la valence émotionnelle et d'activation émotionnelle mesurées de façons psychométriques et psychophysiologiques. Dans un deuxième temps, nous nous intéressons à l'impact de la chaleur émotionnelle perçue dans les réponses des IUC sur ces mêmes construits, mais aussi sur la frustration vécue par les utilisateurs. De fait, cela nous amène à poser les questions de recherche suivantes :

QR1 : *De quelles façons la valence émotionnelle et l'intensité émotionnelle permettent-elles de mesurer l'expérience utilisateur au moment d'interagir avec un IUC?*

***QR2** : Quel est l'effet de la chaleur émotionnelle dans les réponses de l'IUC sur la valence émotionnelle, l'intensité émotionnelle et la frustration, lorsqu'elle est non pertinente à la requête d'un utilisateur?*

Pour répondre à ces questions, nous avons mené deux expériences en laboratoire. Dans un premier temps, dans une étude portant sur la recherche d'informations liées aux fonctionnalités qu'offre le « skill » de Radio-Canada via Alexa (Amazon Inc, Seattle, WA), 11 participants ont réalisé l'étude en laboratoire ainsi que 17 dans une pièce silencieuse chez le partenaire, pour un total de 28 participants. Dans un second temps, 17 participants ont interagi avec un prototype de IUC. Ces deux expériences ont été approuvées par le comité d'éthique (CER) de HEC Montréal. Ensuite, pour mesurer la valence émotionnelle, nous avons utilisé le logiciel Face Reader (Noldus, Wageningen, Pays-Bas), qui analyse en temps réel des micro-expressions faciales du visage. Ensuite l'activation émotionnelle a été enregistrée à l'aide du logiciel Acknowledge (Biopac, Goleta, USA), en capturant la réponse électrodermale par la sudation, grâce à des capteurs placés sur la paume de la main non dominante des participants. L'activation émotionnelle génère de la sudation dermale, laquelle reflète l'intensité de l'état émotionnel de façon précise et en temps réel.

Contributions et implications

D'un point de vue théorique, la recherche actuelle s'ajoute à celle sur l'interaction humain-machine, en mettant en évidence l'effet de la chaleur émotionnelle dans une interaction uniquement vocale avec un IUC, en utilisant à la fois des mesures psychométriques et psychophysiologiques. Plus précisément, ce mémoire met en lumière l'impact de la chaleur émotionnelle sur la valence émotionnelle, l'intensité émotionnelle et la frustration

perçues dans la réponse donnée par le IUC, et ce, lorsque l'utilisateur n'est pas en mesure d'avoir une réponse pertinente à sa requête.

Ce mémoire propose une nouvelle méthodologie qui permet de capturer l'expérience réellement vécue par les utilisateurs et ainsi, enrichir les résultats provenant des méthodes traditionnelles auto-rapportées comme les questionnaires. Même si ces méthodes permettent de recueillir de l'information pertinente sur l'expérience, elles peuvent manquer d'objectivité et de précision. L'ajout des données physiologiques lors de l'évaluation de l'expérience d'interaction avec un IUC permet ainsi d'apporter une précision accrue et d'explorer une nouvelle dimension de l'expérience, soit les réactions automatiques et inconscientes des utilisateurs au stimuli.

Sur le plan managérial, les entreprises pourraient tirer profit de tels résultats, puisque certaines utilisent déjà le « sans interface visuelle » comme un avantage concurrentiel, ou pour atteindre un plus grand nombre de leur consommateurs, adeptes d'enceintes intelligentes. Par exemple, les centres d'appel utilisent déjà la technologie du traitement du langage pour permettre à leurs clients de naviguer dans les différentes options de leurs menus, alors que certaines banques permettent désormais d'effectuer quelques tâches simples comme consulter son solde ou faire un virement (Business Insider, 2016). Ces dites entreprises pourraient donc bénéficier de l'usage de la chaleur émotionnelle dans les réponses programmées, de sorte que l'expérience de l'utilisateur sera moins frustrante, même si les consommateurs sont confrontés à des problèmes techniques. Aussi, Amazon Developer (Amazon Inc, Seattle, WA) offre désormais la possibilité aux concepteurs d'utiliser des outils pour créer leurs propres "compétences", permettant ainsi aux utilisateurs d'assistants vocaux intelligents de consommer leur contenu via ce nouveau canal vocal Ainsi, les entreprises peuvent bénéficier de cette approche méthodologique en capturant différentes dimensions de l'expérience en utilisant des données psychophysiologiques

Structure du mémoire

La structure de ce mémoire est composée de deux articles de nature exploratoire, dont le premier a été publié dans les actes de la conférence HCI 2020 au Danemark. Ce premier article est une étude dans laquelle nous avons exploré la variance ainsi que la convergence d'outils physiologiques avec les mesures psychométriques lors d'interactions avec un assistant vocal intelligent. Le second article, se basant sur les apprentissages méthodologiques du premier, cherche à comprendre l'effet de la chaleur émotionnelle dans la réponse donnée par l'IUC sur la valence émotionnelle, l'intensité émotionnelle ainsi que la frustration perçue.

Résumé du premier article

Dans cette première phase exploratoire, nous proposons une approche multiméthodes pour évaluer l'expérience de l'utilisateur avec un assistant vocal intelligent par la triangulation de mesures provenant de la reconnaissance faciale et de l'activité électrodermale. Cette approche vise à développer une compréhension plus riche de ce que les utilisateurs vivent pendant l'interaction avec un assistant vocal intelligent. Les résultats suggèrent que la valence émotionnelle est mieux captée avec des mesures psychométriques, alors que l'activation est mieux détectée avec des mesures psychophysiologiques. Ainsi, cette étude démontre que chaque méthode capture une dimension émotionnelle distincte.

Résumé du deuxième article

Lors d'une seconde étude exploratoire, nous nous sommes intéressés à la chaleur émotionnelle dans la rétroaction donnée par le IUC, soit d'offrir une rétroaction empathique et chaleureuse, afin de déterminer son impact sur l'expérience utilisateur. Puisque l'utilisation de ces appareils se fait sous forme de discussions et que la littérature démontre clairement que certains utilisateurs entretiennent une forme de relation avec leurs appareils, nous investiguons l'impact qu'a une réponse chaleureuse de l'assistant vocal, lorsqu'il n'est pas en mesure de donner l'information demandée. Les résultats de cette étude suggèrent que la chaleur émotionnelle a un impact positif sur l'expérience

perçue par l'utilisateur, car les participants ont rapporté moins d'émotions négatives et moins de frustration, suivant l'interaction sans succès. De plus, les résultats suggèrent que la chaleur émotionnelle dans les réponses données par le IUC augmente l'intensité émotionnelle vécue par les participants.

Contributions à la recherche

Le tableau ci-dessous présente mon apport à la réalisation des études menées ainsi qu'à l'écriture à chacune des étapes. Ma contribution à chaque étape est présentée sous forme de pourcentage.

Tableau 1 – Contributions menant à la rédaction des articles

Étapes du processus	Contributions personnelles
Définition des attentes du partenaire	• Traduire les besoins d'affaires du partenaire sous forme de question de recherche - **100%**
Revue de littérature	• Rédaction d'une revue de littérature comprenant les principales recherches et construits – **100%**

Design expérimental	• Compléter la demande d'approbation au CER et y faire les modifications – **50%**
	  • Le formulaire a été complété avec l'aide de Bo Huang, candidat au Doctorat ainsi que partenaire de projet.
	• Élaboration et rédaction du protocole d'expérience – **100%**
	• Préparation de la salle d'expérience – **75%**
	  • Une aide des assistantes de recherche du Tech3lab fut fournie pour installer les outils neurophysiologiques

Recrutement des participants	• Élaborer et rédiger le message de recrutement – **100%**
Collectes de données	• Accueil des participants, rédactions des instructions, pose des outils – **75%**

	• Soutien et appui des assistant(e)s de recherche
Analyse des résultats	• Extraction des données – **100%** • Nettoyage des données – **100%** • Analyse Face Reader – **100%** • Analyse EDA – **100%** • Retranscription et analyse des entrevues **– 100%** • Analyses statistiques **– 75%** • Le statisticien du laboratoire a fourni une précieuse aide pour procéder aux analyses complexes des résultats.
Rédaction	• Écriture des articles – **100%** • Les co-auteurs ont apporté des commentaires, des précisions et des corrections au contenu, à des fins d'améliorations. • Écriture du mémoire – **100%**

Chapitre 1
A new approach to measure user experience with conversational user interface: An Exploratory Study

Félix Le Pailleur[1], Bo Huang[1], Pierre-Majorique Léger [1], Sylvain Sénécal[1]

[1] HEC Montréal, Montréal, Canada

felix.le-pailleur@hec.ca, bo.huang@hec.ca, pierre-majorique.leger@hec.ca, sylvain.senecal@hec.ca

Abstract

Voice-controlled intelligent assistants use a conversational user interface (CUI), a system that relies on natural language processing and artificial intelligence to have verbal interactions with end users. In this research, we propose a multi-method approach to assess user experience with a smart voice assistant through triangulation of psychometric and psychophysiological measures. The approach aims to develop a richer understanding of what users experience during the interaction, which could provide new insights to researchers and developers in the field of voice assistant. We apply this new approach in a pilot study, and we show that each method captures a part of emotional variance during the interaction. Results suggest that emotional valence is better captured with psychometric measures, whereas arousal is better detected with psychophysiological measures.

Keywords: Human-Computer Interactions, Conversational user interface, User Experience, Vocal Assistant, Arousal, Valence, Emotion.

1.1 Introduction

Voice assistants (e.g., Alexa, Google Assistant, Siri) are voice-controlled devices that allow consumers to use their voice to make queries such as listening to music, accessing the latest news, or answering general questions. In the U.S., it is estimated that conversational user interface (CUI) users will surpass 123 million by 2021, which represents an increase of 44% since 2017 (Petrock, 2019). In addition, a recent study shows that Amazon has sold 75 million dollars' worth of smart speakers around the globe in 2018, a growth of 600% over the last year (Tung, 2018).

Although voice assistants have become omnipresent in our phones, vehicles, and homes, to date, academic research that aims at developing methods to study these increasingly popular technologies are still lacking (Nass 2006; Sciuto et al. 2018; Lopatovska & Oropeza 2018; Lopatovska & Williams 2018, Jiang 2015). In fact, not all traditional methods for evaluating the user experience appears to be suited to the context of interaction with intelligent voice assistants. For instance, the "Think Aloud" method (Fonteyn et al. 1993) where the researcher asks the participant to verbalize what he or she is doing and thinking while performing a task does not apply in this context since the participant is already using his/her voice to interact with the device.

Therefore, the goal of this paper is to propose a new approach to evaluate user experience during vocal interactions with voice assistants. Specifically, we propose to bonify self-reported measures used *before* and *after* the task with psychophysiological measures (i.e., electrodermal activity and micro facial expressions) to investigate the automatic and non-conscious reaction *during* the interaction. To test the feasibility of our new approach, we conducted an laboratory experiment in which participants (N=11) were instructed to interact with Alexa. To elicit emotional reactions from participants, we designed a set of tasks likely to generate a wide range of discrete emotions.

The article is structured as follows. We first review existing research using self-reported measures in the context of a voice assistant, then we discuss related work on

psychophysiological measurement in the Human-Computer Interaction (HCI) literature. Next, we explain our research methodology as well as summarize the results and their interpretations in the discussion.

1.2 Current research on voice assistant using self-reported measures

Past researches on user interaction with voice assistants has been using both qualitative and quantitative research methods such as questionnaires, diaries, and interviews.

Questionnaires are a widely used tool since they allow researchers to manage a large amount of data from participants qIUCkly and inexpensively (De Singly 2016). There are several forms in which questionnaires can be presented. For example, using Likert scales, questionnaires can be qIUCkly presented to participants before or after completing a task without hindering the flow of the experiment. In a study conducted by Jiang (2015), participants were asked to complete a sequence of 10 tasks using the vocal assistant Cortana on a smartphone, and a questionnaire was used to assess frustration, success, effort, and reuse intentions. For every task, the participant only had to answer a questionnaire regarding their experience using a standard 5-point Likert scale, the most used question model for measuring affective variables (Brown 2000).

Similarly, diaries have also been frequently used as a method for qualitative research because it provides access to users' subjective impressions and more importantly, reflections on their interactions. This technique is advantageous since studies have shown that the presence of a stranger, e.g., researcher, might affect the way a user will interact with a voice assistant since it is mainly used in a private or comfortable context (e.g., home, with friends or alone) (Easwara et al. 2014). Hence, diaries offer a suitable alternative or an addition to qualitative research that might be affected by the presence of a researcher in a laboratory (Nicholl 2010). Researchers have used this method in a variety of contexts to measure user experience with a voice assistant. For instance, Lopatovska and Williams (2018) used a diary log in studying user personification of Alexa. The study data were collected primarily through a structured online diary, which

participants were asked to complete once a day for four days. The diary was also the primary method in Lau et al. (2018)'s study on users' privacy concerns when interacting with the voice assistant. Through the analysis of the diary logs, they found that many non-users did not see the utility of smart speakers or did not trust speaker companies. Other studies went further. They found innovative ways to conduct data collections to understand how Alexa was used in participants' households with multiple members for a long period of time. This was a more natural way to gather data without having to report their interactions in a diary (Sciuto et al. 2018; Lopatovska & Oropeza 2018). For example, a recent study by Porcheron et al. (2018) used a Conditional Voice Recorder (CVR), a device that is activated when Alexa is turned on, to record the interaction. That way, the voice assistant allows to record multiple interactions with family members within a more natural context of use.

As a common tool in the HCI literature, interviews are often used as a complementary method in conjunction with the above-discussed methods. For example, to study user sharing practices of voice assistants, Gary and Moreno (2019) used semi-structured interviews in addition to diary logs. In a similar vein, in-depth interviews were conducted to have a better understanding of the collected conversational logs with voice assistants in investigating Alexa's in-homer usage pattern (Sciuto et al. 2018).

Finally, observations are the only traditional methods allowing to record user behavior during the interaction directly. For instance, in a recent study examining user interaction with voice assistants in public spaces, the area around Alexa was observed at different times of the day and different days for one week, totalling 5.5 observation hours and 132 persons observed (Lopatovska and Oropeza 2019). However, observation provides little insight on how the user feels emotionally and cognitively during the interaction.

1.3 Psychological measures in HCI

As presented above, most studies used qualitative or quantitative methods, mostly relying on self-reported measures. Although they provide extensive and informative results on

user interaction with voice assistants, these methods alone may suffer from not precisely measuring what the user really experienced during the interaction. Researchers are calling for multi-method approaches that consider what the users really experience and perceive (Vom Brocke et al. 2020). For instance, it is possible that these results mainly "assess the user's reflection on the interaction, but not the interaction itself" (Georges et al. 2017, p. 91). Therefore, we posit that what users have really experienced might be different from their subjective evaluation of their experience.

Research in HCI has used psychophysiological measures as a viable indicator of cognitive and emotional states such as cognitive effort or frustration (Rowe et al. 1998; De Guinea et al. 2013; De Guinea et al. 2014; Giroux-Huppé et al. 2019; Beauchesne et al. 2019; Lourties et al. 2018; Agourram et al. 2019; Maunier et al. 2018). The literature has shown that user's emotional and cognitive states can also be inferred using psychophysiological signals, such as electrodermal activity (EDA), heart rate, eye tracking, and facial expressions (see Riedl and Léger 2016 and Riedl et al., Forthcoming; for a review).

By using self-reported measures only, researchers can face various cognitive biases such as social desirability (Ortiz de Guinea et al. 2014). For example, psychologists suggest that the presence of a stranger (e.g., researcher) can change the way one will interact and, in our case, use a voice assistant to respond to the most socially desirable way (Piedmont 2014, p. 6036-6037). For example, by asking participants their likelihood to use a voice assistant in multiple environments (e.g., alone at home, in the metro or at work), Easwara and Vu (2015) found that the social context in which the interaction occurs, influences the information transmitted to the vocal assistant. Hence, psychophysiological measurement tools can contribute to overcoming bias coming from self-reported measures or observations (Xiong 2019).

Thus, in the context of assessing the experience of users while they are interacting with a voice assistant, psychophysiological tools are an interesting add-on because they make it possible to complement traditional means of measurements (e.g., questionnaires, interviews), but especially to bring precision on a specific emotional state, in time, to

which a user cannot remember (Lourties et al. 2018). For example, it might be difficult for a participant, in the context of evaluating an intelligent voice assistant, to remember how he/she felt at a particular moment of the interaction (e.g., when he/she felt frustrated after the CUI gave an irrelevant answer to his/her question).

How users react at the moment of interacting with a device comes from unconscious and automatic mechanisms (De Guinea et al. 2013). The most accurate way to assess how they felt at one particular moment is with the psychophysiological response to the stimuli rather than their perception of what motivates their reaction (Dijksterhuis & Smith, 2005).

In this research, we contribute to the literature on human interaction with voice assistants by proposing a multi-method approach to study user experience with a voice assistant by combining both psychological and psychophysiological measures, which could provide insights to researchers and developers in the field of intelligent assistants. Specifically, this study leverages electrodermal activity and micro facial expressions based on Ekman's universal facial expressions (Ekman, 1997) (happy, sad, angry, surprised, scared, disgusted) and emotional valence (positive-negative) in studying user experience with intelligent assistants. In the next section, we show how psychophysiological measures can offer interesting additional information to
conventional self-reported measures.

1.3.1 Arousal

Arousal is an emotional state related to psychophysiological activity, which is linearly manifested from "calm" to "aroused" (Deng & Poole, 2010; Russell, 2003). Being aroused by a specific stimulus results typically in a feeling of alertness, readiness, or mobility (e.g. body movement, deep breath) (Boucsein, 2012). This emotional state can be measured with Electrodermal Activity (EDA), which can assess the changes in the skin conductance response (SCR) from the nervous system functions (Braithwaite et al., 2013; Dawson et al., 2000.; Bethel 2007). It is an easy to use and reliable psychophysiological measure that has been widely used in NeuroIS research (Léger et al., 2014; Brocke et al., 2013; Giroux-Huppé et al., 2019; Lamontagne et al. (2019). Arousal can also be measured perceptually

by using the self-reported measure such as the Self-Assessment manikin rating (SAM), in which users report their perceived emotional state for a specific stimulus, such as excited, wide-awake, neutral, dull, calm (Bradley& Lang 1994.)

However, the main advantage of using a psychophysiological measure to assess arousal is that it is not invasive, requires no overt behaviour to be recorded, and offers an ecologically valid portrait of the user's arousal, at any time during an experiment (Dirican & Göktürk 2011). For instance, in a study on child-robot interaction, Leite et al. (2013) measured user's arousal through skin conductance and found that such a method is valuable and reliable for capturing interaction with social robots. Also, it can be used to complement and validate traditional survey methods (e.g. questionnaires).

Moreover, in a study measuring the effects of time pressure and accuracy using a computer mouse, participants were asked to paint rectangles with a decreasing time limit. Heiden et al. (2005) found that there was a significant difference in electrodermal data between task difficulty levels. Finally, in a study providing a systematic assessment of IS construct validity, de Guinea et al. (2013) found that the convergent validity of arousal was evidenced by the significant correlation between the SAM scale and the electrodermal data.

1.3.2 Valence

Emotional valence refers to the emotional response, with negative emotions (e.g., fear, anger, sadness) on one side of the spectrum and positive emotions (e.g., joy, surprise) on the other, to a specific stimulus (Lane et al. (1999). Valence can easily be measured perceptually with self-reported measure (e.g., SAM Scale) as the intensity of positive emotions minus the intensity of negative emotions expressed within a range from -1 to 1 (Bradley & Lang 1994). Another way to measure valence is by interpreting facial expressions, which are expressed by the micro-movements of facial muscles (e.g. frowning when angry) (Ekman 1993). It used to be that the only way to interpret facial expressions was via a trained observer who would observe and note changes in facial

expressions based on the Facial Action Coding System (FACS) by Ekman & Friesen (1997).

Today, this time-consuming method is replaced with automatic facial analysis tools (AFA), which can automatically recognize the small changes in facial action units (e.g. raising a brow, chin raise, jaw drop, etc.) and interpret data based on the FACS (Cohn & Kanade, 2007, Ekman 19997).

This technology allows us to accurately detect facial expressions in real time by distinguishing between a set of discrete emotions such as angry, happy, disgusted, sad, scared, surprised. For example, Danner et al. (2014) used this technology to examine participants' facial reactions when tasting orange juice samples to compare implicit measures from the tool with explicit measures from the questionnaire. They found that the software was accurate to report changes in the participant's micro facial expressions between the different samples. Zaman and Shrimpton-Smith (2006) found that, compared to a user's questionnaire, data captured by facial micro-expressions is more effective in measuring instant emotions and fun of use. Also, their results suggest that questionnaire data was instead a reflection of the outcome of a task, rather than a genuine self-reflection of how the user felt when accomplishing the task. Similarly, in a recent study, Lourties et al. (2018) explored the convergent validity of self-reported measures with psychophysiological measures. Their results suggest that the experience lived by a participant is not the same as it is reported. Users self-evaluate their emotional valence more accurately at the end than at the beginning of a task, while they evaluate their arousal more accurately only at the beginning of a task.

To the best of our knowledge, no studies have yet used automatic facial analysis in conjunction with the precise triangulation of electrodermal activity to study user experience with a voice assistant. The proposed triangulated method could provide new insights for this learning or evaluation context using voice only.

1.4 Method

To test the feasibility of using psychophysiological measures in conjunction with psychometric measures to evaluate user experience with voice assistants, we conducted a pilot laboratory experiment where participants were invited to actively interact with Alexa through Amazon's (Amazon Inc, Seattle, WA) Echo Dot (3rd generation) devices by completing a series of tasks. A total of 11 subjects participated in the experiment (4 males, 7 females, mean age=24; sd=5.48) and received a $20 gift card as compensation. This project was approved by the IRB of our institution.

1.4.1 Participants and Design

Since this is a feasibility study, and we wanted to generate as much as variance in the data, we designed a within-subject experiment where each participant was instructed to perform a sequence of interactions. The experiment has one factor with two conditions: impossible tasks (i.e., queries that Alexa was unable to complete) and possible tasks (i.e., queries that Alexa was able to complete) in order to induce negative emotions such as frustration. Participants were randomly assigned to two different sets of tasks wherein one condition, they completed possible tasks before impossible tasks and in the other condition, we reversed the sequence (see Fig. 2). During the experiment, participants completed a set of 8 interactions in total.

1.4.2 Procedure and Measures

Participants were informed that they would have to complete a total of 8 tasks. The goal of each task was explained under the form of pictograms on a tablet.

Participants completed a short questionnaire after each interaction as well as a final questionnaire at the end of the study, followed by a brief interview. To measure user perceptions, the 5-point Self-Assessment Manikin (SAM) scale (Bradley & Lang 1994) was used. The tool allows to directly measure a person's perceived emotional reaction to

a stimulus, such as valence and arousal. Respectively, the scales range from sad (1) happy (5) and calm (1) to excited (5).

For the psychophysiological arousal measure, we collected EDA with Biopac MP-160 (Biopac, Goleta, USA) devices with pre-gel sensors placed on the palm of the participant's non-dominant hand to capture changes in skin conductivity. Electrodermal measures were standardized using as a reference a baseline captured on each participant before the experiment. The baseline consists of measuring the normal electrodermal activity unique to each participant, so that variations from the baseline can be compared. Also, results were rescaled from -1 to 1 for analysis purposes.

Finally, psychophysiological emotional valence was captured via micro facial expressions with the software FaceReader (Noldus, Wageningen, Netherlands). This non-obtrusive method can detect up to six emotions: happy, sad, angry, surprised, scared, and disgusted. Valence value was calculated by subtracting the value of the "happy" emotion and the value of the highest negative emotion (Noldus, FaceReader).

Since the objective of this study is to investigate user experience at the moment of interaction with a voice assistant, only psychophysiological measures that were captured at the moment of listening to Alexa's answers were retained for analysis. It is the participant's reactions to the response given by the voice assistant that interests us.

1.4.3 Material and Apparatus

The apparatus was installed in a quiet room with a mirror window, to reduce noise or external stimulation to make sure there was no interruption and that our psychophysiological data would be of good quality (see Fig. 1 for a detailed setup).

Fig. 1. Experimental setup

Our experimental setup was composed of an Alexa device, a microphone, mounted with a camera, and a digital tablet was installed. During the experiment, participants were interacting with the device. Facial expressions during the experiment were captured using a Logitech camera (Newark, USA), and recorded with the software Media Recorder (Noldus, Wageningen, Netherlands). The software Observer XT (Noldus, Wageningen, Netherlands) and CubeHX (Montréal, Canada) was used to precisely and temporally synchronize all psychophysiological measurements, in line with the guidelines proposed by Léger and colleagues (Léger et al. 2014, Léger et al. 2019, Courtemanche et al. 2018). Statistics were performed using the Statistical Analysis System 9.4 (SAS Inst. U.S.A.).

1.5 Results

Given our within-subject experimental design with repeated measure, as well as the nature of our variable, we performed several linear mixed-effect regressions where each of the measures was entered as a dependent variable, as suggested in the neuro science literature (Rield et al. 2014) (see Table 1 for detailed results). For self-reported measures, namely the valence and arousal, we found that participants reported significantly more positive valence in the possible tasks, compared to impossible tasks ($t(76) = -3.77$, $p < .001$), which was expected. This suggests that participants felt more positive emotions than negative emotions when having successful interactions with the voice assistant. However, arousal did not show a significant difference ($t(76) = 0.54$, $p = .59$, NS) between the two task sets.

For psychophysiological measures, arousal results suggest that impossible tasks generate much higher EDA than possible tasks (t (2638) =7.46, p<.0001). This means that participants experienced a much higher aroused emotional state when they were having difficulties during their interactions. However, in terms of the valence, we did not find a significant difference between possible and impossible tasks (t (1776) =-0.94, p=.35, NS). The following table presents the descriptive statistics and regression results.

Table 1. Summary of results: Means Standard Deviation and Linear Regression

	Possible tasks	**Impossible tasks**	Estimate	Std.Error	t-value	p-value
Valence (self-reported)	3.65 (0.96)	3 (0.83)	-0.65	0.18	-3.77	p<.001
Arousal (self-reported)	2.45 (0.96)	2.36 (0.93)	-.09	0.17	-0.54	p=.59
Arousal (Psychophysiological)	- 0.01 (0.33)	0.07 (0.30)	0.08	0.01	7.46	p<.0001
Valence (Psychophysiological)	0.03 (0.35)	0.004 (0.31)	0.01	0.01	-0.94	p=0.35

Note: Standard deviations are reported in parentheses.

In order to understand the relationship between the two self-reported measures and the psychophysiological measures, we conducted two additional linear mixed-effects regression analyses. The results showed that the self-reported arousal is positively correlated with psychophysiological arousal (t (2638) =3.82, p<.0001). However, surprisingly, our analysis revealed that self-reported valence was negatively correlated with psychophysiological valence (t (1776) = -5,09 ρ <.0001).

1.6 Discussion

Mis en forme: Anglais (Canada)

Our main contribution with this methodological paper is through the triangulation of psychological and psychophysiological measures since, to the best of our knowledge, this study is the first to compare results from both psychophysiological and self-reported

measures in the context of user interaction with a voice assistant. Specifically, we found that for arousal, results from EDA showed a significant difference between possible tasks and impossible tasks (but the self-reported measure did not capture such difference). In contrast, for valence, the self-reported measure was more effective than the AFA in detecting variance in valence. Since previous studies mainly used self-reported measures in studying user interaction with voice assistants, our study contributes by showing the benefit of a multimethod approach in this context, as each method captures a distinct emotional dimension. This suggests that during interaction with a voice assistant, what users experienced might not be the same as reported by themselves. We note that this finding is in line with previous research that combines both methods in studying similar emotional states (i.e., arousal and valence) (Lourties et al. 2018).

Also, the results suggest that the self-perceived arousal was consistent with the psychophysiological responses measured with electrodermal activity when combining both task sets, as they showed a significant positive correlation. These results support previous findings in HCI research using EDA and extend these findings in user interaction context with voice assistants. For example, De Guinea et al. (2013) found that the convergent validity of arousal was evidenced by the significant correlation between the SAM scale measure and the electrodermal measure. Such correlation was evidenced in the current research as well.

Moreover, our results indicate that the emotions inferred from the user's facial expressions by AFA during the interaction complement the self-perceived emotional valence reported by the users. However, we note that there is a discrepancy between valence inferred based on AFA and what is reported by the questionnaire. For example, they are negatively correlated in general when combining both tasks. To investigate this surprising result, we conducted further observation by analyzing the video recordings of our participants performing the tasks. We found a tendency of several participants smiling when they were not able to complete an impossible task, but a smile emanating from frustration rather than joy, which would be aligned with self-reported valence results.

As a future research avenue, researchers have found a way to overcome this kind of situation by focusing on a new set of emotions called epistemic. For example, D'Mello and Calvo (2013) report in their E-learning study with students that" boredom," "confusion," "curiosity," "happiness," and "frustration" where the most common affective states felt during learning and reading situations. In particular, the affective state of "confusion" might be interesting to test in our context since there can be much discrepancy between what the participant expects to get as an answer and the actual answer given by the intelligent voice assistant since speech recognition is not yet optimal. We are currently running a new study where we are considering the affective states, "boredom," "confusion", and "curiosity".

Our experience is limited by the fact that it took place in a user experience laboratory. Thus, the user experience may have been slightly different than if it had taken place in a more natural setting. Future research could extend the current study to other real-life settings such as home and office where interaction with voice assistants is more frequent. In addition, our experiment only measured EDA and facial expressions, while many other tools and measurements suggested by the literature still need to be tested in our specific study context. Hence, it would be interesting for future research to consider a more natural set up and to add more psychophysiological tools. Also, rarely do voice assistant users use their device without performing other tasks at the same time. The main advantage of this tool is that it allows the user to perform a vocal command when he can perform something else simultaneously (e.g. walking, driving or watching television). In our opinion, the idea of adding pupillometry to measure cognitive load (Sirois & Brisson, 2014; Léger et al. 2018) in a multitasking context using a vocal assistant would be an excellent contribution to the research in HCI.

1.7 Références

1. Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019, July). The Relationship Between Technology Self-Efficacy Beliefs and

User Satisfaction–User Experience Perspective. In *International Conference on Human-Computer Interaction* (pp. 389-397). Springer, Cham.

2. Beauchesne, A., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., Di Fabio, M. L., & Léger, P. M. (2019, July). User-centred Gestures for Mobile Phones: Exploring a Method to Evaluate User Gestures for UX Designers. International Conference on Human-Computer Interaction (pp. 121-133). Springer, Cham.
3. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.
4. Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analyzing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology, 49(1), 1017-1034.
5. Brocke, J. V., Riedl, R., & Léger, P. M. (2013). Application strategies for neuroscience in information systems design science research. Journal of Computer Information Systems, 53(3), 1-13.
6. Brown, J. D. (2000). What issues affect Likert-scale questionnaire formats. Shiken: JALT Testing & Evaluation SIG Newsletter, 4(1) Burns N, Grove SK (2005) The Practice of Nursing Research: Conduct, Critique and Utilization
7. C. L. Bethel, K. Salomon, R. R: Murphy, J. L. Burke, Survey of Psychophysiology Measurements Applied to Human-Robot Interaction, in16th IEEE International Symposium on Robot & Human Interactive Communication. (2007)
8. Clifford Nass. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship (Emplacement du Kindle 3471). Édition du Kindle.
9. Cohn, J. F., & Kanade, T. (2007). Use of automated facial image analysis for measurement of emotion expression. Handbook of emotion elicitation and assessment, 222-238.
10. Courtemanche, François, Pierre-Majorique Léger, Aude Dufresne, Marc Fredette, Élise Labonté-LeMoyne, and Sylvain Sénécal. "Physiological heatmaps: a tool for visualizing users' emotional reactions." Multimedia Tools and Applications 77, no. 9 (2018): 11547-11574.

11. Danner, L., Sidorkina, L., Joechl, M., & Duerrschmid, K. (2014). Make a face! Implicit and explicit measurement of facial expressions elicited by orange jIUCes using face reading technology. *Food Quality and Preference*, *32*, 167-172.
12. De Guinea, A. O., Titah, R., & Leger, P. M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. Journal of Management Information Systems, 30(4), 179-210.
13. De Guinea, Ana Ortiz, Ryad Titah, and Pierre-Majorique Léger. "Measure for measure: A two study multi-trait multi-method investigation of construct validity in IS research." Computers in Human Behavior 29, no. 3 (2013): 833-844.
14. De Singly, F. (2016). Le questionnaire-4e édition. Armand Colin.
15. Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human-computer interaction. Procedia Computer Science, 3, 1361–1367.
16. Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. International Journal of Human-Computer Interaction, 31(4), 307–335.
17. Ekman, P. (1993). Facial expression and emotion. *American psychologist*, *48*(4), 384.
18. Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27-46.
19. Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative health research*, *3*(4), 430-441.
20. Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Pourchon, R. (2017, July). The adoption of psychophysiological measures as an evaluation tool in UX. In International Conference on HCI in Business, Government, and Organizations (pp. 90-98). Springer, Cham. Chicago
21. Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019, July). Identifying psychophysiological pain points in the online user

journey: the case of online grocery. In International Conference on Human-Computer Interaction (pp. 459-473). Springer, Cham.

22. Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference. https://doi.org/10.1145/2736277.2741669
23. Kepuska, V., & Bohouta, G. (2018). Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). Presented at the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). https://doi.org/10.1109/ccwc.2018.8301638
24. Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016, March). Understanding user satisfaction with intelligent assistants. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 121-130). ACM.
25. Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map?. In International Conference on Human Interaction and Emerging Technologies (pp. 136-142). Springer, Cham.
26. Lane, R. D., Chua, P. M., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. *Neuropsychologia*, *37*(9), 989-997.
27. Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-311.
28. Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In Information Systems and Neuroscience (pp. 93-99). Springer, Cham.

29. Léger, P. M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neuropsychophysiological correlates of cognitive absorption in an enactive training context. *Computers in Human Behavior*, *34*, 273-283.
30. Léger, P. M., Sénecal, S., Courtemanche, F., de Guinea, A. O., Titah, R., Fredette, M., & Labonte-LeMoyne, É. (2014, October). Precision is in the eye of the beholder: Application of eye fixation-related potentials to information systems research. Association for Information Systems.
31. Léger, P. M., Charland, P., Sénécal, S., & Cyr, S. (2018). Predicting Properties of Cognitive Pupillometry in Human–Computer Interaction: A Preliminary Investigation. In *Information Systems and Neuroscience* (pp. 121-127). Springer, Cham.
32. Lopatovska, I., & Oropeza, H. (2018). User interactions with "Alexa" in public academic space. Proceedings of the Association for Information Science and Technology, 55(1), 309–318. https://doi.org/10.1002/pra2.2018.14505501034
33. Lopatovska, I., & Williams, H. (2018, March). Personification of the Amazon Alexa: BFF or a mindless companion. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 265-268). ACM.
34. Lopatovska, I., & Williams, H. (2018, March). Personification of the Amazon Alexa: BFF or a mindless companion. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval* (pp. 265-268).
35. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., … Martinez, A. (2018). Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, 51(4), 984–997. https://doi.org/10.1177/0961000618759414
36. Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018, July). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In *International Conference on HCI in Business, Government, and Organizations* (pp. 132-144). Springer, Cham.
37. Maunier, B., Alvarez, J., Léger, P. M., Sénécal, S., Labonté-LeMoyne, É., Chen, S. L., ... & Gagné, J. (2018, July). Keep calm and read the instructions: factors for

successful user equipment setup. In International Conference on HCI in Business, Government, and Organizations (pp. 372-381). Springer, Cham.

38. Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. In John T Cacioppo, Louis G Tassinary, and Gary G Berntson, editors, Handbook of Psychophysiology. Cambridge University Press, Cambridge UK, 2 editions, 2000.
39. Nicholl, H. (2010). Diaries as a method of data collection in research. Paediatric Care, 22(7), 16–20. https://doi.org/10.7748/paed2010.09.22.7.16.c7948
40. Noldus FaceReader methodology. https://info.noldus.com/free-white-paper-on-facereader-methodology.
41. Petrock, V. (2019, August 15). Voice Assistant Use Reaches Critical Mass. Retrieved from e-Marketer database
42. Piedmont, R. L. (2014). Social Desirability Bias. In Encyclopedia of Quality of Life and Well-Being Research (pp. 6036–6037). https://doi.org/10.1007/978-94-007-0753-5_2746
43. Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2853-2859). ACM.
44. Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. the 2018 CHI Conference. https://doi.org/10.1145/3173574.3174214
45. Riedl, R., Fischer, T., Léger, P.-M., & Davis, F. D. (Forthcoming). A Decade of NeuroIS Research: Progress, Challenges, and Future Directions. The Data Base for Advances in Information Systems, In Press.
46. Riedl, R., Davis, F. D., & Hevner, A. R. (2014). Towards a NeuroIS research methodology: intensifying the discussion on methods, tools, and measurement. Journal of the Association for Information Systems, 15(10), 4.

47. Riedl, R., Randolph, A. B., vom Brocke, J., Léger, P. M., & Dimoka, A. (2010). The potential of neuroscience for human-computer interaction research. SIGHCI 2010 Proceedings.
48. Rowe, D. W., Sibert, J., & Irwin, D. (1998, January). Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 480-487). ACM Press/Addison-Wesley Publishing Co.
49. Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?" Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18. Presented at the the 2018. https://doi.org/10.1145/3196709.3196772.
50. Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(6), 679-692.
51. Tung, L. (2018, December 20). Amazon : We sold tens of millions of Echo devices in 2018, and Alexa has now 70 000 skills. Retrieved from ZDnet database.
52. Vom Brocke, Jan, Alan Hevner, Pierre Majorique Léger, Peter Walla, and René Riedl. "Advancing a neurois research agenda with four areas of societal contributions." European Journal of Information Systems (2020): 1-16
53. Xiong, J., & Zuo, M. (2020). What does existing NeuroIS research focus on? Information Systems, 89, 101462. https://doi.org/10.1016/j.is.2019.101462
54. Zaman, B., & Shrimpton-Smith, T. (2006, October). The FaceReader: Measuring instant fun of use. In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles (pp. 457-460). ACM. Chicago

Chapitre 2

Can Warmth Save the Day When Interacting with Conversational User Interface? An Exploratory Study

Félix Le Pailleur[1], Bo Huang[1], Pierre-Majorique Léger [1], Sylvain Sénécal[1]

[1] HEC Montréal, Montréal, Canada

felix.le-pailleur@hec.ca, bo.huang@hec.ca, pierre-majorique.leger@hec.ca, sylvain.senecal@hec.ca

Abstract.

In the last few years, Smart Assistants, which are internet-connected devices using a Conversational User Interfaces (CUI) system, have become a center of interest of Human-Computer Interaction (HCI). These devices have been developed to increase users' productivity when performing a variety of tasks that would have otherwise required gestures or visual attention. However, even the most experienced users may still experience frustration when interacting with such technology via spoken language. In this study, we investigate the impact of emotional warmth in the response of CUI on emotional intensity, emotional valence and frustration, following a failed feedback to requests. This approach aims to add to our knowledge on how CUIs can be designed to alleviate frustration by better understanding how users experience vocal interaction with a CUI. To do so, we performed an experiment with 17 participants, where a CUI simulation was incapable of giving adequate feedback to the participant's request. The results suggest that the perceived warmth in the feedback given by the CUI leads to an increase in emotional valence, emotional intensity as well as a decrease in perceived frustration, following a negative feedback.

Keywords: Conversational User Interface, warmth, emotional valence, emotional intensity, frustration, user experience

2.1 Introduction

Recent technological advances in speech recognition have contributed to the development of a new era of voice-controlled electronic devices known as Conversational User Interface (CUI). Amazon's Alexa, Google Assistant, Microsoft's Cortana, and Apple's Siri, are the most popular CUIs, embedded in modern speakers, phones, cars, or homes. Users of such technology can now more efficiently multitask using voice command, since it is now possible, for instance, to change the radio station while keeping eyes on the road when driving or dim the lights from the comfort of their couch (Novak and Hoffmann 2019). In addition to their built-in capabilities, CUIs can be improved by downloading "skills" (equivalent of apps on a smartphone), which allow a third party (e.g., media organization) to extend their services, such as listening to a podcast or a radio, to smart assistants as a new channel to reach consumers (Hoy, 2018). More specifically, CUIs use Natural Language Processing (NLP). This technology allows computers to process human spoken language and to understand its meaning, in order to quickly provide the most accurate feedback, via access to a vast amount of cloud-based information (Liddy, 2001). Compared to previous voice-activated technology, which could only answer pre-established built-in commands, CUIs are internet-connected devices, which means that all the interactions are processed through an external system (e.g., cloud), which enhances the capabilities of these devices to give meaningful feedback to the request (Hoy, 2018).

Despite the potential of CUIs, human-computer interactions are still not as natural as human to human interactions since speech is complex. Unlike computer code language, which does not leave room for interpretation, humans use many approaches to express themselves orally, such as word choice, intonation, amplitude, or range. Hence, one of the significant challenges of speech recognition by computers resides in difficulty to interpret the correct intent of the request (Schneiderman, 2000). The three main problems encountered by users when interacting with CUIs are: 1) Unfamiliar intent (e.g., when the users make a query whose answer is not supported by the device), 2) NLP errors (e.g.,

when the device understands the wrong thing) and 3) Failed feedback (e.g., the feedback is unclear or does not meet the purpose of the query) and system error (Myers et al. 2018).

Therefore, how can we reduce the negative impact of facing these inevitable obstacles on the user's experience? Since interactions with CUI are verbal exchanges, we argue that it should build upon social interaction principles from psychology derived from human-human relationships and dialogues. Research in social psychology comes to the consensus that humans' perception and judgment of others are based on two dimensions: Warmth and competence (Cuddy et al. 2008). Even though both dimensions are considered important, when it comes to making social impressions, the weight of emotional warmth in affective and behavioral reactions in the social perceptions balance is heavier than competence. In other words, one's perceived warmth, i.e., good or bad intentions, will be judged before competence, i.e., perceived ability of the other party to act on those intentions (Fiske et al. 2007).

To this day, the impact of emotional warmth in voice-only interactions with CUIs on the user experience, which is a person's perceptions and responses that result from the use or anticipated use of a product, system, or service (ISO, 2008), is still unclear. Past studies on CUIs, have mainly focused on their characteristics (Hoy, 2018, Lopatovska et al. 2018, Berdasco et al., 2018), satisfaction (Kiseleva et al., 2016; Purington et al., 2017), and personification (Lopatovska et al. 2018) and to better understand how smart assistants are used in the natural context of home (Sciuto et al. 2018 Porcheron and al. 2018). Although warmth has been shown to affect human to human interactions (Fiske et al., 2007, Cuddy et al. 2008), the literature does not provide an answer when it comes to human-CUI interactions. Given the constantly growing number of CUI users (Petrock, 2019), we aim to address this gap in the literature.

In this paper, we argue that a better understanding of how emotional warmth affects the user's affective experience could help CUI and skill designers in offering a better user experience. Hence, we adopt a multi-method approach using psychometric and psychophysiological measures to investigate how the variation in the perceived warmth level in the feedback given by the CUI impacts frustration, emotional intensity and

emotional valence (Riedl & Léger 2016). To do so, we conducted a within-subject experiment in which seventeen participants were instructed to interact with a CUI prototype, hereafter Olia, for the purpose of information retrieval, which was giving either a high-warmth or low-warmth negative feedback to requests.

Our results suggest that following a negative interaction, the warm responses given by the CUI have a positive effect on the physiological emotional intensity response. In addition, participants report significantly more positive emotional valence and less frustration when they perceive a higher level of emotional warmth in the feedback. Thus, results from this exploratory study contribute to the existing literature in several ways. First, they allow us to fill the research gap on how to cope with failed feedback using emotional warmth, since most of the previous studies focus on successful interaction (Kiseleva et al. 2016; Purington et al. 2017). Second, this study contributes to the literature about computers as social actors (Nass et al. 1995), more specifically about a brand tone of voice (Barcelos & Sénécal 2018) and anthropomorphism (Nass and Moon 2000).

This article is structured as follows: First, we discuss the extant literature on warmth and competence and how these constructs have been measured in the HCI context. Then, we present our research methodology using the Wizard of Oz technique, which will be followed by our results and discussion.

2.2 Literature Review

Many researchers in psychology tried to find what factors influence the perception of others in a social context (Fiske et al., 2007; Cuddy et al. 2008). Most of them agreed that the basic dimensions explaining the concept of social behaviors, i.e., judgments and impressions of others, are warmth and competence. These two dimensions alone are responsible for 82% of the variations of our daily social behaviors (Wojciszke et al. 1998).

The term "competence" can be defined as the ability to mobilize with efficacy and efficiency skills and behaviour in a specific context (Le Boterf, 1994). Hence, competence

is not the resource itself (e.g., to know something) but how intelligence, skillfulness, capability, and effectivity are put forward (Aaker, Vohs, and Mogilner 2010; Judd et al. 2005;). In other words, competence is the utilitarian and functional dimension of another party (Yzerbyt et al. 2008)

and is closely related to the ability to mobilize resources to solve a problem or accomplish a task. Hence, we propose that when a smart assistant is unable to answer a (simple) query, it could therefore be associated with a low level of competence, since it is not able to demonstrate the skills required to complete such a task and that they are expected to be fully functional systems.

On the other hand, the concept of warmth can be described as *"mild, volatile emotion involving physiological arousal and precipitated by experiencing directly or vicariously love, family, or friendship relationship*" (Aaker et al., 1986, p.377), implied in a relationship context involving one or more social objects (e.g., a person, a group, an animal, an organization) and associated with being kind, sociable, trustworthy, or helpful (Fiske et al., 2007). The level of perceived warmth in a human-human relationship influences how we will perceive the intentions of the other party as being well-intentioned or not (Fiske et al., 2007). For example, it would be possible for one to feel the warmth through empathy, by watching a pleasant scene of friends having a good. In other words, warmth includes the other party's perceived intentions (e.g., intention to help or harm), and competence refers to the individual's perceived behavioural abilities (e.g., the ability to conduct those intentions or not) (Fiske et al., 2007).

In the literature, several studies state that a high level of perceived warmth and competence is associated with positive emotions, whereas low perceived warmth and competence tend to be judged negatively overall (Fiske et al. 2007, Bergmann., 2007, Kulms et al., 2018). A wealthy businessperson can be perceived as cold but competent, whereas an elderly could be considered having more warmth than competence. In a study measuring the importance of warmth and competence in customer service, participants were given the hypothetical scenario in which they were checking in a hotel at the front

desk; Smith et al. (2015) found that competence and warmth significantly predicted perceived satisfaction from the service given by the employee.

The relation between warmth and competence have been studied jointly in the literature at the level of human-to-human relations. Most of the literature shows that there is no agreement about how warmth and competence are correlated, or if they are correlated at all. However, some work points toward a positive relationship between the two constructs, which means that an individual perceived as positive on one of the two constructs will also tend to be perceived as more positive overall (Kervyn et al. 2010). This effect has been reported from Kelley's (1950) experiment, where two groups of students had a short description of their instructor before the class, one as being very warm and considerate and the second as being very cold. Results show that 52% of the students in the warm condition participated in the class discussion compared to 32% for the cold condition. This result illustrates that perceiving the other party as warm, has a positive effect on the experience. Also, students in the warm condition qualified for the instructor to be more considerate, more sociable, and more humane than those in the other condition, even though it was the same instructor for both groups. However, even though this study shows no significant impact of warmth on competence related traits, such as intelligence, this study clearly shows that students in the high warm condition had a more engaging experience by participating more in class and had a more favourable judgment of the teacher.

In addition to psychology, the concept of warmth has already been studied with computer-based tasks (Kulms et al. 2018), in brand image (Wu et al. 2017), with embodied (e.g., agents that can transmit warmth via media other than voice) (Nakanishi et al. 2019; Kim et al. 2019). For instance, a recent study conducted by Wu et al. (2017) investigated the perception of the brand image through interaction with a smart object being either friendly-like (warmth) or engineer-like (competent). Their results suggest that users perceived the brand image as warmer in the warm interaction compared to the engineer-like. Also, concerning CUI, Kim et al. (2019) show that making a robot more human-like through behaviors and physical appearance increases user's perception of warmth while perceived competence stays unchanged. In other words, anthropomorphized robots should

be perceived as warmer. Moreover, in a recent study, Barcelos & Sénécal (2018) looked at the impact of using different voice tones in communications between a brand and its users on social networks, on their behavior. Their results suggest that, in a hedonic context where there is a low level of engagement and risk among consumers (e.g. when interacting with a smart assistant), a more human tone of voice should be prioritized, whereas in the opposite case, a more corporate tone of voice is preferred.

In sum, the literature shows the positive impact of emotional warmth in interpersonal relationships, but also on the experience with brands and robots, when predicting user satisfaction and perceived warmth. Therefore, although the use of CUIs is now reaching the mainstream market and allows consumers to simplify the execution of everyday tasks through the use of voice commands, the reality remains that the quality of the experience is still frequently spoiled by errors related to the imperfect NLP (Myers, 2018; Luger & Sellen, 2016). Thus, there is currently a gap between consumers' expectations of CUIs and the imperfect user experience, which will be filled once these systems genuinely understand the intent and purpose behind the requests (Moore, 2012). Systems using speech as the only communication medium could, therefore, greatly benefit from the use of emotional warmth, since the technology is not perfect yet and users are still frequently facing issues.

2.3 Hypotheses Development

According to the Computer as Social Agent (CASA) paradigm (Lee & Nass 2010), human beings view and similarly respond to computers as they would to other people, such as by perceptions of personality or using forms of politeness in their interactions, even if they are fully aware that they are interacting with an object. The reason is that Smart Assistants, also called "Social Agents" predispose users to personify or anthropomorphize them since they are speech-based, they need social interaction in order to function, they have gender (e.g., male or female voice) and a personality (e.g., telling jokes). Also, activating the device by saying its name (e.g., Alexa, Siri, Cortana) also predisposes users to anthropomorphize and socialize with the device and, by doing so, facilitates its integration

into a social context. Anthropomorphism can be defined as a conscious mechanism where human-like characteristics are given to non-human agents or objects (Nass and Moon 2000, Purington et al. 2017). Hence, compared to the relationship a user may have with any other type of product, human-CUI relationships share the characteristics that they are dynamic (Foehr & Germelmann, 2019). For instance, studies report users having different types of relationships with CUI's, from the agent being perceived as a servant, a partner, or a master (Novak & Hoffman 2019; Foehr & Germelmann, 2019).

Since warmth is a priori dimension applied to humans, we suggest that giving more human-like characteristics to CUIs through a warmer dialogue could reduce negative emotions, such as frustration, when a problem occurs. Past research on CUIs found supporting evidence that the emotional and behavioral responses of CUI users that shape our impressions are almost identical to those of humans. For example, Kulms et al. (2018) investigated the impact of warmth with computer-based tasks where participants were asked to play a game similar to Tetris on a computer, but with an intelligent agent as a partner. Their results suggest that the level of perceived warmth from the agent determines whether he appeared to be cooperative or selfish during the game and hence how trustworthy the agent is. Moreover, Nakanishi et al. (2019) found that a social robot perceived as being welcoming, polite, and cute creates a heartwarming feeling for the users, making them feel like the robot was more engaged to help them.

Also, by analyzing Amazon product reviews, Purington et al. (2017) found that personification of Alexa is associated with increased levels of satisfaction, regardless of technical problems or function. In other words, the quality of the experience itself of interacting with a smart device is more important than the outcome, i.e., if the answer meets the initial objective of the request (Lopatovska et al. 2018) since users seek for more enjoyable human-like interactions with the CUI, over the correctness of the feedback. This suggests that a warm interaction will have a mitigating effect on the negative emotions experienced as a result of negative feedback.

Emotions can be described as changes in the body and brain in response to specific stimuli of one's perceptions relative to a specific object or event (Damasio, 1994). In their

experiment, Aaker et al. (1986) used the warmth monitor to measure the perceived emotional responses in combination with Electrodermal Activity (EDA), which can assess variance in skin conductance response when users were watching advertisements. This self-reported tool consists of a paper separated from "no warmth" to "emotional," on the horizontal axis, where participants must continuously draw a pencil on a vertical line reflecting the temporal flow of the advertisement, and the horizontal variations express the perceived warmth. Their results suggest warmth has a physiological component since emotional intensity, i.e., arousal measured since EDA was correlated (0.67) with the perceived warmth from the warmth monitor (Aaker et al. 1896). Therefore, we expect that a negative, but emotionally warm response will evoke greater emotional intensity as a physiological response from the body, following a failed interaction with the CUI (i.e., an EDA increase):

H1: A warmer response leads to greater emotional intensity following a failed interaction with CUI.

On the other hand, valence is used in psychology is referring to the pleasant or unpleasant quality of a stimulus or situation, with positive emotions (e.g., joy) on one side of the spectrum and the other, negative emotions (e.g., fear, anger, sadness) (Lane et al. (1999). Since warmth is here defined as a mild positive emotion (Aaker et al. 1986) and emotional warmth is useful to predict interpersonal judgment's valence in a social context, i.e., whether it will be more positive or negative emotions involved (Fiske et al. 2007), we propose:

H2: A warmer response will lead to less perceived negative emotions following a failed interaction with CUI.

Finally, since we are in an information retrieval context, we consider that users are frustrated when their search is interrupted by an inability to obtain a response to the query. To this day, CUI users can use several tactics to solve unfamiliar intent, NLP errors, and failed feedback, but their usage does not guarantee a positive outcome since many of them may face negative emotions such as frustration, following a failed interaction (Myers et

al. 2018). Thus, we suggest that a negative, but emotionally warm response will be perceived as less frustrating by users, even though the outcome is still negative:

H3: A warmer response will lead to less perceived frustration following a failed interaction with CUI.

2.4 Method

2.4.1 Sample & Design

To test our hypotheses, we conducted a within-subject experiment with one factor and two conditions: high warmth and low warmth of the answer given by our CUI prototype named "Olia." A total of 17 subjects participated in our study (7 males, 9 females, mean age=41), which were recruited via a research panel. Our experiment, which was approved by the IRB of our institution, consisted of interacting with a simulation of a vocal assistant using the Wizard of Oz (WoZ) method (Kelley 1984). The WoZ, which consists of a human, the "wizard," manually controlling the responses of the simulated CUI (i.e., Olia) with pre-recorded answers. It allows the researchers to overcome the technological and financial constraints of creating a functional application. Since many vocal assistants have female voices (e.g. Alexa, Siri, Google Assistant), we did the same with Olia, to make the simulated interaction as realistic as possible. Then, we created different typical scenarios that CUI users face in everyday life when using such devices, such as finding the opening hours of a store or adding appointments to the calendar. For a similar question, we varied the type of answer given by Olia as either being high-warmth, i.e., Olia was responding with empathy as well as a form of politeness, or low-warmth, i.e., Olia was responding with short and direct answers. For instance, we asked participants to tell Olia that they are not feeling well and to ask her to find the nearest hospital. The high-warmth negative feedback was: "I hope everything is fine, however, I'm afraid what you're asking is beyond my capabilities at the moment, can I help you with something else??" and the low warmth answer was "I do not know". Conditions' order was counterbalanced to eliminate the possibility of order effects. Finally, both conditions started with two successful

interactions, where Olia gave positive feedback to the request, to make the prototype more truthful (see annexe for full dialogue).

2.4.2 Procedure

After we welcomed the participants, they were informed to read and sign the consent form, and then a research assistant installed psychophysiological instruments (See Apparatus below). Participants were then invited to sit comfortably and were informed that they would have to complete a total of 6 interactions with the prototype and that an interaction ends when Olia provides an answer. Tasks were: 1) Ask Olia to book your dentist appointment for tomorrow at 10:00 am, 2) Ask Olia to remind you to call your friend Felix tomorrow at 8:00 pm, 3) You are planning your outing next weekend at the Montreal Museum of Fine Arts. Ask Olia about opening hours, 4) Tell Olia that you don't feel well and ask her to find the nearest hospital. 5) You are planning your outing next weekend to the Musée Grévin. Ask Olia for more information about business hours. 6) Tell Olia that you have a headache and ask her to find the nearest pharmacy. After receiving feedback from each of the conditions, i.e. high-warmth and low-warmth, participants were asked to complete a questionnaire measuring perceived warmth, emotions felt (e.g. frustration, excitation, surprise, confusion), and self-reported emotional valence and emotional intensity. At the end of the tasks, participants were interviewed to learn more about the strengths and weaknesses of the experience of interacting with Olia.

2.4.3 Measures

Since user's emotional reactions following the failed interaction comes from unconscious automatic mechanism (De Guinea et al. 2013) and that the objective of this study is to measure what users experienced when they receive the negative feedback from the CUI, we used a multi-method approach which provides a new, richer and more accurate perspective on the actual user experience (Vom Brocke et al. 2020). Because it is difficult to retroactively recall a specific emotional state at a particular point in the experience without interrupting it, Dijksterhuis & Smith (2005) suggest that the most accurate way to understand how one feels at a particular moment of the experience is through the

measurement of psychophysiological response to specific stimuli. The psychophysiological measures have mainly been used for research in Human-Computer Interaction (HCI) and have proven to be a reliable indicator of cognitive and emotional states (Rowe et al. 1998; De Guinea et al. 2013; De Guinea et al. 2014; Giroux-Huppé et al. 2019; Beauchesne et al. 2019; Lourties et al. 2018; Agourram et al. 2019; Maunier et al. 2018). The main advantages of psychophysiological measures are that they can be captured in a non-obtrusive way without distracting the user's attention during the experience to reduce the ecological validity of the results (Dirican & Göktürk 2011). Also, the literature shows that psychophysiological signals such as electrodermal activity (EDA) can measure a user's emotional intensity, precisely, in real-time (Riedl & Léger 2016; Riedl et al. 2020).

In their experiment, investigating the level of perceived warmth in advertising viewing, Aaker et al. (1986) found that the perceived level of warmth was correlated with EDA. EDA is an easy to use, reliable and widely used psychophysiological indicators in HCI research (Léger et al., 2014; Brocke et al., 2013; Giroux-Huppé et al., 2019; Lamontagne et al. (2019) that assess changes in the skin conductance response (SCR) from the nervous system functions (Braithwaite et al., 2013; Dawson et al., 2000.; Bethel 2007). Hence, we measured emotional intensity, which is the physiological response to a warm stimulus, with a psychophysiological measure of EDA. General changes in autonomic arousal are linked to changes in the tonic level of electrical conductivity of the skin, which varies with the state of sweat glands. Therefore, we measured EDA with sensors placed on participants' palms, with a Biopac MP-160 (Biopac, Goleta, USA) device, in order to capture changes in SCR. Standardized physiological means were then rescaled from -1 to 1 for analysis. Perceived warmth and frustration were assessed using a one item 7-point Likert scale, from strongly agree to disagree strongly, following the questions: "after completing these tasks, I find that Olia's voice is warmth" and "after completing these tasks, I feel frustrated." At the end of each interaction, the participant had to complete a short questionnaire to measure their self-perceived valence; we used the 5-point Self-Assessment Manikin (SAM) scale (Bradley & Lang 1994) to measure perceived emotional valence in reaction to a stimulus. More specifically, for the warmth manipulation question, we used the following items: To what extent do the following traits

describe the answer: warm, friendly and well-intentioned (Fiske et al., 2002). Frustration was measured by one item, "frustrated." We summarized all our measures and instruments in Table 1. Statistical analysis was performed using Statistical Analysis System 9.4 (SAS Inst., USA), and for analysis purposes, we rescaled EDA's results from -1 to 1.

Table 1: Summary of measures and instruments.

Variables	**Measurement instrument**
Emotional intensity (Arousal)	Changes in the skin conductance response measured with Electrodermal activity (-1 to 1) (Boucsein, 2012)
Perceived warmth	Participants assess perceived warmth using a 7-point (1 to 7) Likert Scale with 3 items (warm, friendly and well-intentioned) after each task: "after completing these tasks, I find that Olia's voice is warmth"
Perceived frustration	Participants assess perceived frustration using a one item 7-point (1 to 7) Likert Scale after the tasks: "after completing these tasks, I feel frustrated"
Perceived valence	Participants assess their perceived valence using the SAM Scale (-1 to 1) (Bradley & Lang 1994) after the tasks

2.4.5 Apparatus

The apparatus was installed in a closed and quiet room to reduce external distractions that could affect the quality of the psychophysiological data or interrupt the experiment. The experimental set up is composed of a Bluetooth speaker to simulate the voice, a tablet on a stand with instructions mounted on a camera and a second tablet placed on the table on which the questionnaires are completed. To capture changes in skin conductivity, we collected EDA using a Biopac MP-160 (Biopac, Goleta, USA) device with sensors placed

on the participant's non-dominant hand. Finally, we used Amazon Polly (Amazon Inc, Seattle, WA), which uses cloud technology to turn textual entry into like-like speech, to pre-record Olia's answers.

During the experiment, the software Observer XT (Noldus, Wageningen, Netherlands) was used to precisely and temporally synchronize psychophysiological measurements from skin conductance following the guidelines proposed by Léger and colleagues (Léger et al. 2014, Léger et al. 2019, Courtemanche et al. 2018).

2.5 Results

The following table presents descriptive statistics of our results, where for each condition, the mean and standard deviation are presented.

Table 2: Summary of descriptive results

	Emotional intensity (Mean and (std))	**Emotional valence (Mean and (std))**	**Frustration (Mean and (std))**
High warmth	-0.099 (0.25)	2.82 (0.88)	2.29 (1.4)
Low warmth	-0.116 (0.256)	2.24 (0.66)	3.59 (1.8)

Descriptive results suggest that participants experienced greater emotional intensity, had more positive emotional valence and less frustration. These differences are tested below. Also, we performed a manipulation to asses the perceived warmness in both conditions. In terms of the warmth manipulation, results show that it was successful: Participants in the high-warmth condition perceived Olia as warmer than those in the low-warmth condition (Mhigh-warmth=4.491, Mlow-warmth= 4.392, t (32) =3.63, p= 0.0029).

A t-test was used to analyze the results of self-reported and psychophysiological measures (see Table 2). Concerning Hypothesis 1 on the effect of warmth on emotional intensity (**H1)**, our results suggest that EDA is significantly higher in the high-warmth condition (*M*high-warmth=-0.099, *M*low-warmth= -0.116, t (32) = 3.37, p=0.015). In other words, participants felt more emotional warmth when they received a warm (high warmth) answer from the smart assistant when facing a negative answer than when receiving a colder (low warmth) answer. **Thus, H1b is supported.**

For Hypothesis 2, suggesting a positive relationship between warmth and valence, results suggest that perceived valence is more positive in the high-warmth condition than for the low-warmth condition (*M*high-warmth=2.824, *M*low-warmth= 2.235, t (32) = 2.11, p=0.0431). This result suggests that the participant reported feeling more positive emotions when perceived warmth was higher after a failed interaction. **Thus, H2 is supported.**

For the effect of warmth on perceived frustration (**H3**), results suggest that perceived frustration is significantly lower in the high-warmth condition (*M*high-warmth= 2.294, *M*low-warmth= 3.588, t (32) = -3.45, p=0.0032). That means that even though the outcome of both conditions was negative since the interaction failed, the warmer response has a significant negative effect on the perceived frustration. **Thus, H13 is supported.**

Table 3: Summary of results: Means Standard Deviation

	High Warmth	**Low Warmth**	p-value
Emotional intensity (Psychophysiological)	-0.10	-0.12	<0.001
Emotional valence (Self-reported)	2.82	2.24	0.043

Frustration (Self-reported)	2.29	3.59	0.003

2.6 Discussion

The goal of this paper was to investigate the effect of warmth in responses of CUI following a failed interaction, on emotional intensity (H1), as well as subjective frustration and valence (H2-3). Overall, we found that, following a failed interaction, warmth had a positive impact on the user's experience, as the participants reported fewer negative emotions and less frustration, as expected. This result indicates that warmth in the answer given by the CUI, even though participants were facing a negative outcome after their interactions in both conditions, felt more positive emotions, which confirms that a warmer experience is more important than the outcome (Lopatovska et al. 2018). Moreover, results from EDA suggest that warmth had a positive physiological effect on the participant's affective experience.

This study is making three theoretical contributions. First, it investigates interactions with CUI in a failure context. Past research on user experience (UX) with CUI mainly focused on successful interactions while overlooking the unavoidable situations where CUI fails to deliver a satisfactory outcome (Kiseleva et al. 2016; Purington et al. 2017). Therefore, we contribute to this stream of research by showing the positive effect of warmth on user experience with CUI following failed interaction since our results suggest that participants reported feeling fewer negative emotions and less frustration in the high-warmth condition. We hence offer a possible solution to mitigate the negative effect when a failure occurs.

Second, past studies on emotional warmth have mainly focused on brands, advertising and embodied agents (Aaker et al. 1987; Wu et al. 2017; Nakanishi et al. 2019). For instance, Aaker and al. (1986) measured perceived warmth when watching advertising with electrodermal activity in combination with self-reported arousal (Aaker et al. 1986),

but to the best of our knowledge, this study is the first to investigate emotional intensity, in a failure context, when interacting with CUI.

Third, the current research adds to human-computer interaction research by highlighting the effect of warmth with voice-only interaction with CUI, using a multi-method approach. Extant research mainly used interviews, observation, questionnaires and diaries in order to investigate the interactions with CUI (Easwara et al., 2014; Jiang, 2015; Lopatovska and Williams, 2018). By using physiological data, we look at real-time interaction instead of the reflection of the interaction (Georges et al. 2017, p. 91), which is a limitation for the usage of self-reported measures only, like most of the previous studies on CUI. Researchers are calling for multi-method approaches to assess what the users really perceived and experienced, (Vom Brocke et al. 2020), where both psychometric and psychophysiological measures are used.

Managerially, this study makes one main contribution. Companies might benefit from our study's results. CUI such as Alexa, Siri, Google Assistant and Cortana are very well adopted by the consumers, with 83.1 million smart speakers in the US market in 2020, representing a growth of 13.7% compared to the previous year (He, 2020). Moreover, tools are available online such as Amazon Developer (Amazon Inc, Seattle, WA) offer the possibility for designers to build and set up their own "skills" and then extend their content through smart speakers. To this day, businesses are already using voice interaction as a competitive advantage. For instance, some of them, such as media organizations, now allow their users, which own and daily use their smart speakers, to personalize their news access experience by specifically asking their intelligent voice assistant to play the podcast or radio show of their choice. Hence, such organizations can reach a more significant number of consumers through these popular devices. Also, companies with high-volume call centers are already using or might consider using voice recognition technology for their customers to save time to navigate through the various options of the menus in a more autonomous and intuitive way than by dialling numbers on their phone. In such scenarios, skill designers for CUI can benefit from our results. The creation user paths can be more complex for CUI than for traditional interfaces since the possibilities are not limited to the options displayed on the screen, but rather to the multiple ways a

user can formulate his or her query. It is also much more difficult to situate a user in the digital business environment without a physical and visual appearance. Using emotional warmth in programmed responses could help reduce frustration from speech recognition errors, missing content or unclear responses, and preserve a good brand image. The rise in popularity of the user-centric experience is now well established within companies and the use of neuroscience is a considerable added value in order to offer well-designed products. Some banks now offer the possibility for their customers to perform tasks related to their secure customer area via Alexa, such as checking their balance, their payment history or paying bills (Amazon Inc, Seattle, WA). This kind of action generates higher stakes than when it comes to knowing the weather, and that is why it is important for companies to offer a good experience to their clients so that they can be confident and use CUI in such serious context. Our findings regarding the benefits of using emotional warmth in the feedback given by CUI can help create products that will be better adopted by consumers.

Our experience is limited by the fact that it took place in a controlled environment where participants sat in front of a tablet and had to interact with our prototype. In a more natural context of use, consumers tend to interact with these products in a multitasking context in the comfort of their homes. In addition, the fact that the researchers were in the same room as the participant may have caused a cognitive bias of social desirability with the participant. This means that their behaviours and responses may have been biased to be the most socially acceptable in an evaluation context (e.g., not showing frustration). Thus, for future research, it would be interesting to reproduce the usual environment where interactions with CUI take place. Participants would be alone in a room and would have to interact with the CUI to obtain certain information while having their attention shared with another task (e.g. playing a game on their cell phone). In addition, this type of experiment would deepen the concept of attention in the context of multitasking with CUI. Also, our experience is limited by the use of the Wizard of Oz technique, which allowed us to reproduce voice interactions that would have had a CUI user but allowed no more than one exchange between the participant and the device. This limitation meant that the interactions were of short duration and, therefore, less authentic. It would be interesting in future research to create a skill using a tool such as Amazon Developer (Amazon Inc,

Seattle, WA) to reproduce a more engaging and realistic experience with several possible exchanges per task. Finally, in this experiment, we limited ourselves to measuring the perceived warmth of responses. However, it would be interesting to assess the perception of the second dimension, competence, i.e., whether CUI appears more competent when they are warm.

In this study, we addressed the concept of emotional warmth to understand its impact on the user experience when CUIs are unable to provide adequate feedback to the request. We hope the current research provides findings and insights that bring value to the understanding of UX with CUI by enabling interactions that generate the least amount of frustration for users, until technological advances and research provide a consistent vocal user experience.

2.7 Annexe

Table 4: Summary of interactions with Olia by condition.

	Condition 1	**Condition 2**
Instruction	**Successful feedback**	**Successful feedback**
Ask Olia to remind you to call your friend Felix tomorrow at 8:00 pm.	*Of course, your reminder is scheduled for tomorrow at 8:00 pm.*	*Of course, your reminder is scheduled for tomorrow at 8:00 pm.*
Ask Olia to schedule your dentist appointment for tomorrow at 10:00 am.	*Perfect, I've booked your appointment at the dentist, tomorrow at 10:00 am*	*Perfect, I've booked your appointment at the dentist, tomorrow at 10:00 am*
Instruction	**High-warmth feedback**	**Low-warmth feedback**
Tell Olia that you are not feeling well and ask her to find the nearest hospital	*I hope everything is going well, however, I'm afraid what you're asking is beyond my capabilities at the moment, can I help you with something else?*	*I hope everything is going well, however, I'm afraid what you're asking is beyond my capabilities at the moment, can I help you with something else?*

You are planning your outing next weekend at the Montreal Museum of Fine Arts. Ask Olia about the opening hours.	*Forgive me, I've tried but I can't find the opening hours of the Montreal Museum of Fine Arts, is there anything else I can do to help you?*	*Forgive me, I've tried but I can't find the opening hours of the Montreal Museum of Fine Arts, is there anything else I can do to help you?*
Instruction	**Low-warmth feedback**	**High-warmth feedback**
Tell Olia that you have a headache and ask her to find the nearest pharmacy	*I don't know*	*I hope everything is going well, however, I'm afraid what you're asking is beyond my capabilities at the moment, can I help you with something else?*
You are planning your outing next weekend at the Grévin Museum. Ask Olia for more information about business hours	*I cannot help you with this request*	*Forgive me, I've tried but I can't find the opening hours of the Grévin Museum, is there anything else I can do to help you?*

2.8 References

1. Aaker, D. A., Stayman, D. M., & Hagerty, M. R. (1986). Warmth in advertising: Measurement, impact, and sequence effects. Journal of consumer research, 12(4), 365-381
2. Aaker, J., Vohs, K. D., & Mogilner, C. (2010). Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter. Journal of Consumer Research, 37(2), 224-237
3. Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019, July). The Relationship Between Technology Self-Efficacy Beliefs and User Satisfaction–User Experience Perspective. In International Conference on Human-Computer Interaction (pp. 389-397). Springer, Cham.

4. B. Schneiderman. The limits of speech recognition. Communications of the ACM, 43:63–65, 2000.
5. Barcelos, R. H., Dantas, D. C., & Sénécal, S. (2018). Watch your tone: How a brand's tone of voice on social media influences consumer responses. Journal of Interactive Marketing, 41, 60-80.
6. Beauchesne, A., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., Di Fabio, M. L., & Léger, P. M. (2019, July). User-centred Gestures for Mobile Phones: Exploring a Method to Evaluate User Gestures for UX Designers. International Conference on Human-Computer Interaction (pp. 121-133). Springer, Cham.
7. Berdasco, López, Diaz, Quesada, & Guerrero. (2019). User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. Proceedings, 31(1), 51. https://doi.org/10.3390/proceedings2019031051
8. Bergmann, K., Eyssel, F., & Kopp, S. (2012, September). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In International conference on intelligent virtual agents (pp. 126-138). Springer, Berlin, Heidelberg.
9. Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.
10. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), 49-59.
11. Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analyzing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology, 49(1), 1017-1034.
12. Brocke, J. V., Riedl, R., & Léger, P. M. (2013). Application strategies for neuroscience in information systems design science research. Journal of Computer Information Systems, 53(3), 1-13.
13. C. L. Bethel, K. Salomon, R. R: Murphy, J. L. Burke, Survey of Psychophysiology Measurements Applied to Human-Robot Interaction, in16th IEEE International Symposium on Robot & Human Interactive Communication. (2007)
14. Courtemanche, François, Pierre-Majorique Léger, Aude Dufresne, Marc Fredette, Élise Labonté-LeMoyne, and Sylvain Sénécal. "Physiological heatmaps: a tool for

visualizing users' emotional reactions." Multimedia Tools and Applications 77, no. 9 (2018): 11547-11574.

15. Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In Advances in Experimental Social Psychology (pp. 61–149). Elsevier. https://doi.org/10.1016/s0065-2601(07)00002-0
16. Damasio, A. R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain, New York: Avon Books.
17. De Guinea, A. O., Titah, R., & Leger, P. M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. Journal of Management Information Systems, 30(4), 179-210.
18. De Guinea, Ana Ortiz, Ryad Titah, and Pierre-Majorique Léger. "Measure for measure: A two study multi-trait multi-method investigation of construct validity in IS research." Computers in Human Behavior 29, no. 3 (2013): 833-844.
19. Dijksterhuis, A., & Smith, P. K. (2005). What do we do unconsciously? And how? Journal of Consumer Psychology, 15(3), 225–229
20. Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human-computer interaction. Procedia Computer Science, 3, 1361–1367.
21. Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. International Journal of Human-Computer Interaction, 31(4), 307–335.
22. Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. Trends in Cognitive Sciences, 11(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005
23. Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of personality and social psychology, 82(6), 878
24. Foehr, J., & Germelmann, C. C. (2019). Alexa, can I trust you? Exploring consumer paths to trust in smart voice-interaction technologies.

25. fundamental dimensions of social judgment. Personality and Social Psychology Bulletin, 34(8), 1110-1123.
26. Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Pourchon, R. (2017, July). The adoption of psychophysiological measures as an evaluation tool in UX. In International Conference on HCI in Business, Government, and Organizations (pp. 90-98). Springer, Cham. Chicago
27. Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Pourchon, R. (2017, July). The adoption of psychophysiological measures as an evaluation tool in UX. In International Conference on HCI in Business, Government, and Organizations (pp. 90-98). Springer, Cham. Chicago
28. Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019, July). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In International Conference on Human-Computer Interaction (pp. 459-473). Springer, Cham.
29. He, A. (2020, February 18). Amazon Maintains Convincing Lead in US Smart Speaker Market. Retrieved from e-Marketer database
30. Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly, 37(1), 81–88. https
31. J. Kelley. An iterative design methodology for user-friendly natural language office information applications. ACMTransactions on Information Systems (TOIS), 2(1):26–41, 1984.
32. Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference. https://doi.org/10.1145/2736277.2741669
33. Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between judgments of competence and warmth. Journal of Personality and Social Psychology, 89(6), 899–913. https://doi.org/10.1037/0022-3514.89.6.899

34. Kelley, H.: The warm-cold variable in first impressions of persons. Journal of Personality 18, 431—439 (1950)
35. Kervyn, N., Yzerbyt, V., & Judd, C. M. (2010). Compensation between warmth and competence: Antecedents and consequences of a negative relation between the two fundamental dimensions of social perception. European Review of Social Psychology, 21(1), 155–187. https://doi.org/10.1080/13546805.2010.517997
36. Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Marketing Letters, 30(1), 1–12. https://doi.org/10.1007/s11002-019-09485-9
37. Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016, March). Understanding user satisfaction with intelligent assistants. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 121-130). ACM.
38. Kulms, P., & Kopp, S. (2018). A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers. Frontiers in Digital Humanities, 5. https://doi.org/10.3389/fdigh.2018.00014
39. Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map?. In International Conference on Human Interaction and Emerging Technologies (pp. 136-142). Springer, Cham.
40. Lane, R. D., Chua, P. M., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. Neuropsychologia, 37(9), 989-997.
41. Le Boterf, G. (1994). De la compétence. Essai sur un attracteur étrange.
42. Lee, J. E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives (pp. 1-15). IGI Global

43. Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In Information Systems and Neuroscience (pp. 93-99). Springer, Cham.
44. Léger, P. M., Sénecal, S., Courtemanche, F., de Guinea, A. O., Titah, R., Fredette, M., & Labonte-LeMoyne, É. (2014, October). Precision is in the eye of the beholder: Application of eye fixation-related potentials to information systems research. Association for Information Systems.
45. Liddy, E. D. (2001). Natural language processing.
46. Lopatovska, I., & Williams, H. (2018, March). Personification of the Amazon Alexa: BFF or a mindless companion. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 265-268). ACM.
47. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., … Martinez, A. (2018). Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, 51(4), 984–997. https://doi.org/10.1177/0961000618759414
48. Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018, July). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In International Conference on HCI in Business, Government, and Organizations (pp. 132-144). Springer, Cham.
49. Luger, E., & Sellen, A. (2016, May). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 5286-5297).
50. Maunier, B., Alvarez, J., Léger, P. M., Sénécal, S., Labonté-LeMoyne, É., Chen, S. L., ... & Gagné, J. (2018, July). Keep calm and read the instructions: factors for successful user equipment setup. In International Conference on HCI in Business, Government, and Organizations (pp. 372-381). Springer, Cham.
51. Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. In John T Cacioppo, Louis G Tassinary, and Gary G Berntson, editors, Handbook of Psychophysiology. Cambridge University Press, Cambridge UK, 2 editions, 2000.

52. Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). Patterns for How Users Overcome Obstacles in Voice User Interfaces. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. the 2018 CHI Conference. https://doi.org/10.1145/3173574.3173580
53. Nakanishi, J., Baba, J., & Kuramoto, I. (2019). How to Enhance Social Robots' Heartwarming Interaction in Service Encounters. Proceedings of the 7th International Conference on Human-Agent Interaction - HAI '19. the 7th International Conference. https://doi.org/10.1145/3349537.3352798
54. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of social issues, 56(1), 81-103.
55. Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995, May). Can computer personalities be human personalities? In Conference companion on Human factors in computing systems (pp. 228-229).
56. Novak, T. P., & Hoffman, D. L. (2019). Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. Journal of the Academy of Marketing Science, 47(2), 216-237
57. Novak, T. P., & Hoffman, D. L. (2019). Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. Journal of the Academy of Marketing Science, 47(2), 216-237
58. Petrock, V. (2019, August 15). Voice Assistant Use Reaches Critical Mass. Retrieved from e-Marketer database
59. Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. the 2018 CHI Conference. https://doi.org/10.1145/3173574.3174214
60. Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2853-2859). ACM.

61. Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Berlin, Heidelberg.
62. RIEDL, René, FISCHER, Thomas, LÉGER, Pierre-Majorique, et al. A Decade of NeuroIS Research: Progress, Challenges, and Future Directions. Data Base for Advances in Information Systems, 2020, vol. 51.
63. Roger Moore. 2012. Spoken Language Processing: Where do we go from Here? In Your Virtual Butler: The Making- of. Robert Trappl (ed). Springer, 119-133 doi: 10.1007/978-3-642-37346-6_10
64. Rowe, D. W., Sibert, J., & Irwin, D. (1998, January). Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 480-487). ACM Press/Addison-Wesley Publishing Co.
65. Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?" Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18. Presented at the the 2018. https://doi.org/10.1145/3196709.3196772.
66. Smith, N. A., Martinez, L. R., & Sabat, I. E. (2016). Weight and gender in service jobs: The importance of warmth in predicting customer satisfaction. Cornell Hospitality Quarterly, 57(3), 314-328.
67. Vom Brocke, Jan, Alan Hevner, Pierre Majorique Léger, Peter Walla, and René Riedl. "Advancing a neurois research agenda with four areas of societal contributions." European Journal of Information Systems (2020): 1-16
68. Wojciszke, B. et al. (1998) On the dominance of moral categories in impression formation. Pers. Soc. Psychol. Bull. 24, 1245–1257
69. Wu, J., Chen, J., & Dou, W. (2016). The Internet of Things and interaction style: the effect of smart interaction on brand attachment. Journal of Marketing Management, 33(1–2), 61–75. https://doi.org/10.1080/0267257x.2016.1233132
70. Yzerbyt, V. Y., Kervyn, N., & Judd, C. M. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1110-1123.

Conclusion

Les relations humain-machine lors des dernières années ont grandement été marquées par les avancées technologiques en termes d'intelligence artificielle, permettant aux IUC de passer de la science-fiction, au quotidien des détenteurs d'appareils intelligents. Ce gain en popularité et l'adoption de produits utilisant des IUC tels que les assistants vocaux intelligents permettent de justifier l'intérêt afin d'adapter les méthodes d'évaluations de ces produits n'utilisant aucune interface visuelle.

Tout d'abord, ce mémoire a pour objectif de proposer une nouvelle approche pour mesurer l'expérience utilisateur lors d'interactions purement verbales avec les IUC, en proposant l'utilisation de mesures psychophysiologiques, notamment l'activité électrodermale et l'analyse automatisée des expressions faciales. Ensuite, la méthodologie utilisée dans le cadre de l'étude exploratoire a été utilisée afin d'investiguer l'impact de la chaleur émotionnelle dans les réponses données par l'IUC, dans un contexte où il est impossible pour l'utilisateur d'obtenir réponse à sa requête, sur la frustration, la valence ainsi que la réaction physiologique à la chaleur évoquée par la tâche.

À l'hiver 2020, deux expériences intra-sujets ont été réalisées en laboratoire. Dans le cadre de la première expérience, un total 27 participant ont interagi le « skill » de Radio-Canada, par l'entremise d'Alexa, où chaque tâche était une fonctionnalité offerte par l'entreprise à découvrir. Afin de générer de la variance émotionnelle chez le participant pour comparer la valence et l'activation entre les mesures psychométriques et psychophysiologiques, nous avons créé des tâches impossibles, c'est-à-dire des tâches n'étant pas programmée pour être répondues générant en autre, de la frustration chez les participants. Cette première collecte a permis de comprendre que l'expérience vécue par les utilisateurs au moment d'interagir avec l'IUC, peut ne pas être exactement ce qu'ils ont eux-mêmes rapporté dans les questionnaires. En effet, les résultats suggèrent que la valence

émotionnelle est mieux saisie avec des mesures psychométriques, alors que l'activation est mieux détectée par des mesures psychophysiologiques.

Ensuite, nous avons réalisé une seconde expérience auprès de 17 participants, dans laquelle nous avons utilisé la technique du Wizard of Oz afin de simuler les réponses d'un assistant vocal intelligent, et ce, en manipulant le niveau de chaleur des réponses données par celui-ci. Chaque participant devait interagir avec l'IUC simulé, dans le but d'obtenir de l'assistance afin de réaliser diverses tâches. Toutefois, les participants ont reçu des réponses négatives à leur requête, dans un premier temps de façon chaleureuse et ensuite, de façon neutre. Cette seconde collecte nous a permis d'approfondir l'impact de la chaleur émotionnelle, dans un contexte où le résultat d'interaction est négatif, sur la valence, la frustration ainsi que réaction physiologique à la chaleur évoquée par la tâche. Les deux collectes de données ont été effectuées en milieux contrôlés et avec l'utilisation de mesures physiologiques en continu ainsi que psychométriques post-tâches.

Ce chapitre rappelle ainsi les principales questions de recherche de ce mémoire ainsi que des principaux résultats tirés de chacun des deux articles. De plus, les contributions théoriques et managériales, ainsi que les limites et les avenues de recherche futures seront abordées.

Rappel des questions de recherche et principaux résultats

Les résultats de ces deux articles ont permis de répondre aux questions de recherche suivante :

QR1*: De quelles façons la valence émotionnelle et l'intensité émotionnelle permettent-elles de mesurer l'expérience utilisateur au moment d'interagir avec un IUC?*

Pour cette première question de recherche exploratoire, nous nous intéressons à savoir si les mesures physiologiques permettent de capturer une autre dimension de l'expérience. Plus précisément, nous nous intéressons à savoir si la perception subjective des participants quant à l'expérience vécue lors de l'interaction avec l'IUC va dans la même direction que ceux rapportées par les réactions physiologiques du corps. Nos résultats

suggèrent que l'activation perçue par les participants était cohérente avec les réponses psychophysiologiques mesurées avec l'EDA uniquement. À l'inverse, les résultats provenant de l'analyse automatisée des micro-expressions faciales démontrent des résultats contraires à ceux rapportées par le Sam Scale. Cela suggère que lors d'une interaction vocale avec un IUC, l'état émotionnel que les utilisateurs ont objectivement vécu peut ne pas être exactement semblable à celui qu'ils ont subjectivement rapporté.

Nos résultats suggèrent que l'EDA capture une différence d'activation entre les tâches possibles et impossibles, alors que ce n'est pas le cas pour les mesures auto rapportées. Au contraire, l'analyse automatisée des micros-expressions faciales fut meilleure pour détecter la variance de la valence entre les deux conditions. Ainsi, utiliser une approche multi méthode permet de capturer deux dimensions émotionnelle distinctes.

QR2: *Quel est l'impact de la chaleur émotionnelle perçue sur la valence émotionnelle, l'intensité émotionnelle et la frustration, lorsque que la réponse d'un IUC est non pertinente à la requête d'un utilisateur?*

Plus précisément, 3 hypothèses ont découlé de cette troisième question de recherche en se basant sur la littérature en Interaction Humain-Machine, notamment avec les IUC ainsi qu'en psychologie, en lien avec les dimensions sociales expliquant les relations entre les humains.

En premier lieu, **H1** postulait qu'au moment d'interagir avec un IUC, qui, basé sur les principes psychologiques des relations humaines, est plus chaleureux dans sa rétroaction, conduit à une plus grande réponse physiologique évoquée par la tâche. En effet, nos résultats suggèrent que l'EDA est significativement plus élevée dans la condition de chaleur élevée. Autrement dit, les participants ont ressenti plus de chaleur émotionnelle, mesurée grâce à la sudation, lorsqu'ils ont reçu une réponse chaleureuse de l'IUC face à une réponse négative.

En second lieu, **H2** soutenait qu'une réponse plus chaleureuse entraîne une diminution des émotions négatives perçues à la suite de l'échec d'une interaction avec un IUC, les résultats suggèrent que la valence perçue est plus positive pour la condition de chaleur élevée que pour la condition de chaleur faible. En d'autres termes, ces résultats indiquent que les participants ont déclaré ressentir significativement plus d'émotions positives, lorsque la chaleur perçue était plus élevée, et ce, même après une interaction ratée.

Finalement, **H3** soutenait qu'une réponse plus chaleureuse entraînerait moins de frustration perçue suivant l'échec d'une interaction avec un IUC. Les résultats suggèrent que la frustration perçue est significativement plus faible lorsque le niveau de chaleur est plus élevé. Ce résultat suggère que, même si le résultat des deux conditions, soit une rétroaction chaleureuse ou neutre, était négatif puisque l'interaction a échoué, une réponse plus chaleureuse a un effet négatif significatif sur la frustration perçue.

Contributions du mémoire

Méthodologiquement parlant, ce mémoire contribue à combler l'écart dans la littérature à ce qui a trait aux méthodologies utilisées afin de mesurer l'expérience entre les utilisateurs et les IUC. Les recherches passées étudiant ces appareils sans interfaces visuelles utilisent principalement des méthodes traditionnelles de collecte de données, lesquelles sont rapportées par les participants et font ainsi face à plusieurs biais. Avec la popularité croissante de ces dispositifs, il est important de se demander si les méthodes traditionnelles, utilisées principalement pour les interfaces visuelles telles que mobile et desktop, sont appropriées dans ce contexte. Par exemple, la méthode "Think Aloud" (Fonteyn et al. 1993) où le chercheur demande au participant de verbaliser ses pensées pendant l'exécution d'une tâche ne s'applique pas dans ce contexte puisque le participant utilise déjà sa voix pour interagir avec le dispositif. Ce mémoire permet donc de mettre de l'avant une approche multi méthode pour étudier l'expérience de l'utilisateur avec un IUC par la triangulation des mesures psychologiques et psychophysiologiques. Ainsi, les principaux résultats de ce mémoire au niveau méthodologique pourront permettre aux

chercheurs en expérience utilisateur désirant faire l'évaluation de produits vocaux, d'utiliser cette nouvelle approche afin de capturer autant les dimensions perçues que celles vécues de l'expérience.

D'un point de vue théorique, ce mémoire permet de combler l'écart quant à l'utilisation de la chaleur émotionnelle, dans un contexte d'interaction humain-machine. En effet, les recherches antérieures sur l'expérience des utilisateurs (UX) avec les IUC se sont principalement concentrées sur les interactions réussies, tout en négligeant les situations inévitables où les IUC ne donnent pas de résultats satisfaisants (Kiseleva et al. 2016 ; Purington et al. 2017). Ce mémoire contribue à combler cet écart en montrant que la chaleur émotionnelle, soit cette dimension centrale des relations sociales entre humains, permet notamment aux utilisateurs d'être moins frustrés et plus heureux à la suite d'une interaction où l'IUC n'est pas en mesure d'offrir une rétroaction adéquate à la requête. De plus, des études antérieures ont mesuré l'activité électrodermale en combinaison avec l'excitation auto déclarée (Aaker et al. 1986), mais à notre connaissance, cette étude est la première à étudier la chaleur ressentie dans un contexte d'échec, lors d'une interaction avec un IUC. Ce mémoire contribue à la littérature en démontrant qu'une réponse chaleureuse donnée par l'IUC évoque de la chaleur émotionnelle via une réponse physiologique du corps.

Pour l'industrie, les résultats de ce mémoire permettent d'enrichir les connaissances actuelles sous plusieurs aspects. Premièrement, la conception de « skill » pour les assistants vocaux intelligents est désormais accessible grâce à des outils comme Amazon Developper (Amazon Inc, Seattle, WA), permettant à des entreprises, de petite ou grande envergure, d'utiliser les commandes vocales pour améliorer l'expérience utilisateur offerte. En effet, tel que mentionné dans ce mémoire, les IUC font désormais partie de notre quotidien, mais la technologie permettant le traitement du langage naturel cause encore fréquemment de la frustration chez les utilisateurs. Ainsi, les résultats de ce mémoire notamment quant à la triangulation des données ainsi qu'à l'utilisation de la chaleur émotionnelle offrent aux chercheurs, développeurs et designers des outils afin d'offrir constamment une expérience centrée sur l'utilisateur.

Deuxièmement, au-delà d'offrir son contenu via les assistants vocaux, les entreprises pourraient bénéficier des résultats de ces recherches afin d'améliorer l'expérience utilisateur des centres d'appels, autant pour les clients que pour les employés. L'utilisation du traitement du langage naturel permettant actuellement aux IUC de donner des réponses pertinentes pourrait certainement bénéficier aux centres d'appels, permettant un meilleur tri des appels entrants, d'exprimer un besoin n'étant pas présent dans la liste offerte et de rendre l'expérience personnalisable. En effet, ces entreprises qui prendront le virage vers les commandes vocales pourraient certainement bénéficier de nos résultats en faisant un usage de la chaleur émotionnelle, notamment dans les messages codés lors de situations d'erreurs, afin de limiter la frustration des utilisateurs dans ces circonstances.

Finalement, alors que certaines entreprises sont déjà dotées de laboratoire de recherche en expérience utilisateur doté d'outils neurophysiologiques, la méthodologie présentée dans ce mémoire offre aux chercheurs en entreprise les fondations nécessaires afin de créer une expérience permettant de tester de futurs produits utilisant le traitement du langage naturel. Il sera tout d'abord important de considérer que les émotions déduites des expressions faciales de l'utilisateur par l'analyse automatisée des expressions faciales pendant l'interaction complètent la valence émotionnelle auto rapportée par les utilisateurs.

Limites du mémoire

Les recherches effectuées dans le cadre de ce mémoire comportent plusieurs limites qui se doivent d'être mises de l'avant. Premièrement, les expériences se sont déroulées dans laboratoire de recherche en expérience utilisateur, soit un environnement contrôlé. Ainsi, l'expérience de l'utilisateur peut avoir été légèrement différente de celle qui aurait eu lieu dans un cadre plus naturel.

Dans cet environnement, les participants étaient assis devant l'assistant vocal intelligent ainsi qu'une tablette, ce qui diffère grandement du contexte naturel d'utilisation de ce genre de produit, soit à la maison et généralement dans un contexte multitâche.

De plus, dans la seconde étude, le fait que les chercheurs étaient dans la même pièce que le participant peut avoir causé un biais cognitif de désirabilité sociale. Cela signifie que leurs comportements et leurs réponses peuvent avoir été biaisés pour être les plus socialement acceptable, ne reflétant pas nécessairement le véritable fond de leur pensée. De plus, de la technique du Wizard of Oz, laquelle nous a permis de reproduire des rétroactions vocales qu'auraient eu un utilisateur avec un véritable IUC, ne permettait qu'un seul échange. Cette limitation signifie que les interactions étaient de courte durée et donc moins authentiques.

Ensuite, les deux recherches effectuées dans le cadre de ce mémoire ont utilisé que l'activité électrodermale ainsi que les expressions faciales, alors que de nombreux autres outils et mesures sont suggérés par la littérature, notamment la pupillométrie qui offre des mesures d'attention et de charge cognitive, pour mesurer l'expérience utilisateur dans un contexte multitâche.

Finalement, une dernière limitation de ce mémoire concerne la taille de l'échantillon. En effet, nos recherches comportent respectivement de petits échantillons de N=27 et N=11, pouvant rendre nos résultats plus difficilement généralisables.

Avenues pour recherche future

En ce qui a trait aux avenues de recherche, il serait tout d'abord intéressant de reproduire des expériences similaires tout en agrandissant l'échantillon, afin de savoir si les résultats se maintiennent et sont donc généralisables à l'ensemble de la population utilisant les IUC.

Ensuite, pour de futures recherches, il serait intéressant de reproduire l'environnement habituel dans lequel se déroulent les interactions avec les IUC, soit la maison. Les participants pourraient être seuls dans une pièce ressemblant à un salon et devraient interagir avec l'IUC afin d'obtenir certaines informations tout en ayant leur attention partagée avec une autre tâche (par exemple, jouer à un jeu sur leur téléphone portable). Cela permettrait d'approfondir le concept d'attention et de charge cognitive dans le

contexte du multitâche avec ce genre d'appareil mesuré grâce à des lunettes oculométriques.

Finalement, il serait intéressant de créer un « skill » spécifiquement conçue pour la recherche afin de reproduire une expérience plus engageante et plus réaliste avec plusieurs échanges possibles.

Bibliographie

1. Aaker, D. A., Stayman, D. M., & Hagerty, M. R. (1986). Warmth in advertising: Measurement, impact, and sequence effects. Journal of consumer research, 12(4), 365-381
2. Aaker, J., Vohs, K. D., & Mogilner, C. (2010). Nonprofits are seen as warm and for-profits as competent: Firm stereotypes matter. Journal of Consumer Research, 37(2), 224-237
3. Agourram, H., Alvarez, J., Sénécal, S., Lachize, S., Gagné, J., & Léger, P. M. (2019, July). The Relationship Between Technology Self-Efficacy Beliefs and User Satisfaction–User Experience Perspective. In *International Conference on Human-Computer Interaction* (pp. 389-397). Springer, Cham.
4. B. Schneiderman. The limits of speech recognition. Communications of the ACM, 43:63–65, 2000.
5. Barcelos, R. H., Dantas, D. C., & Sénécal, S. (2018). Watch your tone: How a brand's tone of voice on social media influences consumer responses. Journal of Interactive Marketing, 41, 60-80.
6. Beauchesne, A., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., Di Fabio, M. L., & Léger, P. M. (2019, July). User-centred Gestures for Mobile Phones: Exploring a Method to Evaluate User Gestures for UX Designers. International Conference on Human-Computer Interaction (pp. 121-133). Springer, Cham.
7. Berdasco, López, Diaz, Quesada, & Guerrero. (2019). User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. Proceedings, 31(1), 51. https://doi.org/10.3390/proceedings2019031051
8. Bergmann, K., Eyssel, F., & Kopp, S. (2012, September). A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In International conference on intelligent virtual agents (pp. 126-138). Springer, Berlin, Heidelberg.
9. Boucsein, W. (2012). Electrodermal activity. Springer Science & Business Media.

10. Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. Journal of behavior therapy and experimental psychiatry, 25(1), 49-59.
11. Braithwaite, J. J., Watson, D. G., Jones, R., & Rowe, M. (2013). A guide for analyzing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology, 49(1), 1017-1034.
12. Brocke, J. V., Riedl, R., & Léger, P. M. (2013). Application strategies for neuroscience in information systems design science research. Journal of Computer Information Systems, 53(3), 1-13.
13. Brown, J. D. (2000). What issues affect Likert-scale questionnaire formats. Shiken: JALT Testing & Evaluation SIG Newsletter, 4(1) Burns N, Grove SK (2005) The Practice of Nursing Research: Conduct, Critique and Utilization
14. Business Insider (2016). Capital One increases Alexa functionality for Amazon Echo. Retrieve from: https://www.businessinsider.com/capital-one-increases-alexa-functionality-for-amazon-echo-2016-7
15. C. L. Bethel, K. Salomon, R. R: Murphy, J. L. Burke, Survey of Psychophysiology Measurements Applied to Human-Robot Interaction, in16th IEEE International Symposium on Robot & Human Interactive Communication. (2007)
16. Clifford Nass. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship (Emplacement du Kindle 3471). Édition du Kindle.
17. Cohn, J. F., & Kanade, T. (2007). Use of automated facial image analysis for measurement of emotion expression. Handbook of emotion elicitation and assessment, 222-238.
18. Courtemanche, François, Pierre-Majorique Léger, Aude Dufresne, Marc Fredette, Élise Labonté-LeMoyne, and Sylvain Sénécal. "Physiological heatmaps: a tool for visualizing users' emotional reactions." Multimedia Tools and Applications 77, no. 9 (2018): 11547-11574.
19. Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In Advances in Experimental Social Psychology (pp. 61–149). Elsevier. https://doi.org/10.1016/s0065-2601(07)00002-0

20. Damasio, A. R. 1994. Descartes' Error: Emotion, Reason, and the Human Brain, New York: Avon Books.
21. Danner, L., Sidorkina, L., Joechl, M., & Duerrschmid, K. (2014). Make a face! Implicit and explicit measurement of facial expressions elicited by orange jIUCes using face reading technology. *Food Quality and Preference*, *32*, 167-172.
22. De Guinea, A. O., Titah, R., & Leger, P. M. (2014). Explicit and implicit antecedents of users' behavioral beliefs in information systems: A neuropsychological investigation. Journal of Management Information Systems, 30(4), 179-210.
23. De Guinea, Ana Ortiz, Ryad Titah, and Pierre-Majorique Léger. "Measure for measure: A two study multi-trait multi-method investigation of construct validity in IS research." Computers in Human Behavior 29, no. 3 (2013): 833-844.
24. De Singly, F. (2016). Le questionnaire-4e édition. Armand Colin.
25. Dijksterhuis, A., & Smith, P. K. (2005). What do we do unconsciously? And how? Journal of Consumer Psychology, 15(3), 225–229
26. Dirican, A. C., & Göktürk, M. (2011). Psychophysiological measures of human cognitive states applied in human-computer interaction. Procedia Computer Science, 3, 1361–1367.
27. Easwara Moorthy, A., & Vu, K.-P. L. (2015). Privacy Concerns for Use of Voice Activated Personal Assistant in the Public Space. International Journal of Human-Computer Interaction, 31(4), 307–335.
28. Ekman, P. (1993). Facial expression and emotion. *American psychologist*, *48*(4), 384.
29. Ekman, P., & Keltner, D. (1997). Universal facial expressions of emotion. *Segerstrale U, P. Molnar P, eds. Nonverbal communication: Where nature meets culture*, 27-46.
30. Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: warmth and competence. Trends in Cognitive Sciences, 11(2), 77–83. https://doi.org/10.1016/j.tics.2006.11.005

31. Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of personality and social psychology, 82(6), 878
32. Foehr, J., & Germelmann, C. C. (2019). Alexa, can I trust you? Exploring consumer paths to trust in smart voice-interaction technologies.
33. Fonteyn, M. E., Kuipers, B., & Grobe, S. J. (1993). A description of think aloud method and protocol analysis. *Qualitative health research*, *3*(4), 430-441. fundamental dimensions of social judgment. Personality and Social Psychology Bulletin, 34(8), 1110-1123.
34. Georges, V., Courtemanche, F., Sénécal, S., Léger, P. M., Nacke, L., & Pourchon, R. (2017, July). The adoption of psychophysiological measures as an evaluation tool in UX. In International Conference on HCI in Business, Government, and Organizations (pp. 90-98). Springer, Cham. Chicago
35. Giroux-Huppé, C., Sénécal, S., Fredette, M., Chen, S. L., Demolin, B., & Léger, P. M. (2019, July). Identifying psychophysiological pain points in the online user journey: the case of online grocery. In International Conference on Human-Computer Interaction (pp. 459-473). Springer, Cham.
36. He, A. (2020, February 18). Amazon Maintains Convincing Lead in US Smart Speaker Market. Retrieved from e-Marketer database
37. Hoy, M. B. (2018). Alexa, Siri, Cortana, and More: An Introduction to Voice Assistants. Medical Reference Services Quarterly, 37(1), 81–88. https
38. J. Kelley. An iterative design methodology for user-friendly natural language office information applications. ACMTransactions on Information Systems (TOIS), 2(1):26–41, 1984.
39. Jiang, J., Hassan Awadallah, A., Jones, R., Ozertem, U., Zitouni, I., Gurunath Kulkarni, R., & Khan, O. Z. (2015). Automatic Online Evaluation of Intelligent Assistants. Proceedings of the 24th International Conference on World Wide Web - WWW '15. Presented at the 24th International Conference. https://doi.org/10.1145/2736277.2741669
40. Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: Understanding the relations between

judgments of competence and warmth. Journal of Personality and Social Psychology, 89(6), 899–913. https://doi.org/10.1037/0022-3514.89.6.899

41. Kelley, H.: The warm-cold variable in first impressions of persons. Journal of Personality 18, 431—439 (1950)
42. Kepuska, V., & Bohouta, G. (2018). Next-generation of virtual personal assistants (Microsoft Cortana, Apple Siri, Amazon Alexa and Google Home). 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). Presented at the 2018 IEEE 8th Annual Computing and Communication Workshop and Conference (CCWC). https://doi.org/10.1109/ccwc.2018.8301638
43. Kervyn, N., Yzerbyt, V., & Judd, C. M. (2010). Compensation between warmth and competence: Antecedents and consequences of a negative relation between the two fundamental dimensions of social perception. European Review of Social Psychology, 21(1), 155–187. https://doi.org/10.1080/13546805.2010.517997
44. Kim, S. Y., Schmitt, B. H., & Thalmann, N. M. (2019). Eliza in the uncanny valley: anthropomorphizing consumer robots increases their perceived warmth but decreases liking. Marketing Letters, 30(1), 1–12. https://doi.org/10.1007/s11002-019-09485-9
45. Kiseleva, J., Williams, K., Jiang, J., Hassan Awadallah, A., Crook, A. C., Zitouni, I., & Anastasakos, T. (2016, March). Understanding user satisfaction with intelligent assistants. In Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (pp. 121-130). ACM.
46. Kulms, P., & Kopp, S. (2018). A Social Cognition Perspective on Human–Computer Trust: The Effect of Perceived Warmth and Competence on Trust in Decision-Making With Computers. Frontiers in Digital Humanities, 5. https://doi.org/10.3389/fdigh.2018.00014
47. Lamontagne, C., Sénécal, S., Fredette, M., Chen, S. L., Pourchon, R., Gaumont, Y., ... & Léger, P. M. (2019, August). User Test: How Many Users Are Needed to Find the Psychophysiological Pain Points in a Journey Map?. In International Conference on Human Interaction and Emerging Technologies (pp. 136-142). Springer, Cham.

48. Lane, R. D., Chua, P. M., & Dolan, R. J. (1999). Common effects of emotional valence, arousal and attention on neural activation during visual processing of pictures. Neuropsychologia, 37(9), 989-997.
49. Lau, J., Zimmerman, B., & Schaub, F. (2018). Alexa, are you listening? privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction*, *2*(CSCW), 1-311.
50. Le Boterf, G. (1994). De la compétence. Essai sur un attracteur étrange.
51. Lee, J. E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. In Trust and technology in a ubiquitous modern environment: Theoretical and methodological perspectives (pp. 1-15). IGI Global
52. Léger, P. M., Charland, P., Sénécal, S., & Cyr, S. (2018). Predicting Properties of Cognitive Pupillometry in Human–Computer Interaction: A Preliminary Investigation. In *Information Systems and Neuroscience* (pp. 121-127). Springer, Cham.
53. Léger, P. M., Courtemanche, F., Fredette, M., & Sénécal, S. (2019). A cloud-based lab management and analytics software for triangulated human-centered research. In Information Systems and Neuroscience (pp. 93-99). Springer, Cham.
54. Léger, P. M., Davis, F. D., Cronan, T. P., & Perret, J. (2014). Neuropsychophysiological correlates of cognitive absorption in an enactive training context. *Computers in Human Behavior*, *34*, 273-283.
55. Léger, P. M., Sénecal, S., Courtemanche, F., de Guinea, A. O., Titah, R., Fredette, M., & Labonte-LeMoyne, É. (2014, October). Precision is in the eye of the beholder: Application of eye fixation-related potentials to information systems research. Association for Information Systems.
56. Liddy, E. D. (2001). Natural language processing.
57. Lopatovska, I., & Oropeza, H. (2018). User interactions with "Alexa" in public academic space. Proceedings of the Association for Information Science and Technology, 55(1), 309–318. https://doi.org/10.1002/pra2.2018.14505501034

58. Lopatovska, I., & Williams, H. (2018, March). Personification of the Amazon Alexa: BFF or a mindless companion. In Proceedings of the 2018 Conference on Human Information Interaction & Retrieval (pp. 265-268). ACM.
59. Lopatovska, I., Rink, K., Knight, I., Raines, K., Cosenza, K., Williams, H., … Martinez, A. (2018). Talk to me: Exploring user interactions with the Amazon Alexa. Journal of Librarianship and Information Science, 51(4), 984–997. https://doi.org/10.1177/0961000618759414
60. Lourties, S., Léger, P. M., Sénécal, S., Fredette, M., & Chen, S. L. (2018, July). Testing the convergent validity of continuous self-perceived measurement systems: an exploratory study. In International Conference on HCI in Business, Government, and Organizations (pp. 132-144). Springer, Cham.
61. Luger, E., & Sellen, A. (2016, May). " Like Having a Really Bad PA" The Gulf between User Expectation and Experience of Conversational Agents. In Proceedings of the 2016 CHI conference on human factors in computing systems (pp. 5286-5297).
62. Maunier, B., Alvarez, J., Léger, P. M., Sénécal, S., Labonté-LeMoyne, É., Chen, S. L., ... & Gagné, J. (2018, July). Keep calm and read the instructions: factors for successful user equipment setup. In International Conference on HCI in Business, Government, and Organizations (pp. 372-381). Springer, Cham.
63. Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. In John T Cacioppo, Louis G Tassinary, and Gary G Berntson, editors, Handbook of Psychophysiology. Cambridge University Press, Cambridge UK, 2 editions, 2000.
64. Michael E Dawson, Anne M Schell, and Diane L Filion. The electrodermal system. In John T Cacioppo, Louis G Tassinary, and Gary G Berntson, editors, Handbook of Psychophysiology. Cambridge University Press, Cambridge UK, 2 editions, 2000.
65. Myers, C., Furqan, A., Nebolsky, J., Caro, K., & Zhu, J. (2018). Patterns for How Users Overcome Obstacles in Voice User Interfaces. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. the 2018 CHI Conference. https://doi.org/10.1145/3173574.3173580

66. Nakanishi, J., Baba, J., & Kuramoto, I. (2019). How to Enhance Social Robots' Heartwarming Interaction in Service Encounters. Proceedings of the 7th International Conference on Human-Agent Interaction - HAI '19. the 7th International Conference. https://doi.org/10.1145/3349537.3352798
67. Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of social issues, 56(1), 81-103.
68. Nass, C., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995, May). Can computer personalities be human personalities? In Conference companion on Human factors in computing systems (pp. 228-229).
69. Nicholl, H. (2010). Diaries as a method of data collection in research. Paediatric Care, 22(7), 16–20. https://doi.org/10.7748/paed2010.09.22.7.16.c7948
70. Noldus FaceReader methodology. https://info.noldus.com/free-white-paper-on-facereader-methodology.
71. Novak, T. P., & Hoffman, D. L. (2019). Relationship journeys in the internet of things: a new framework for understanding interactions between consumers and smart objects. Journal of the Academy of Marketing Science, 47(2), 216-237
72. Petrock, V. (2019, August 15). Voice Assistant Use Reaches Critical Mass. Retrieved from e-Marketer database
73. Piedmont, R. L. (2014). Social Desirability Bias. In Encyclopedia of Quality of Life and Well-Being Research (pp. 6036–6037). https://doi.org/10.1007/978-94-007-0753-5_2746
74. Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice Interfaces in Everyday Life. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. the 2018 CHI Conference. https://doi.org/10.1145/3173574.3174214
75. Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017, May). Alexa is my new BFF: social roles, user satisfaction, and personification of the amazon echo. In Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (pp. 2853-2859). ACM.

76. Riedl, R., & Léger, P. M. (2016). Fundamentals of NeuroIS. Studies in Neuroscience, Psychology and Behavioral Economics. Springer, Berlin, Heidelberg.
77. Riedl, R., Fischer, T., Léger, P.-M., & Davis, F. D. (Forthcoming). A Decade of NeuroIS Research: Progress, Challenges, and Future Directions. The Data Base for Advances in Information Systems, In Press.
78. Riedl, R., Randolph, A. B., vom Brocke, J., Léger, P. M., & Dimoka, A. (2010). The potential of neuroscience for human-computer interaction research. SIGHCI 2010 Proceedings.
79. RIEDL, René, FISCHER, Thomas, LÉGER, Pierre-Majorique, et al. A Decade of NeuroIS Research: Progress, Challenges, and Future Directions. Data Base for Advances in Information Systems, 2020, vol. 51.
80. Roger Moore. 2012. Spoken Language Processing: Where do we go from Here? In Your Virtual Butler: The Making- of. Robert Trappl (ed). Springer, 119-133 doi: 10.1007/978-3-642-37346-6_10
81. Rowe, D. W., Sibert, J., & Irwin, D. (1998, January). Heart rate variability: Indicator of user state as an aid to human-computer interaction. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 480-487). ACM Press/Addison-Wesley Publishing Co.
82. Sciuto, A., Saini, A., Forlizzi, J., & Hong, J. I. (2018). "Hey Alexa, What's Up?" Proceedings of the 2018 on Designing Interactive Systems Conference 2018 - DIS '18. Presented at the the 2018. https://doi.org/10.1145/3196709.3196772.
83. Smith, N. A., Martinez, L. R., & Sabat, I. E. (2016). Weight and gender in service jobs: The importance of warmth in predicting customer satisfaction. Cornell Hospitality Quarterly, 57(3), 314-328.
84. Vom Brocke, Jan, Alan Hevner, Pierre Majorique Léger, Peter Walla, and René Riedl. "Advancing a neurois research agenda with four areas of societal contributions." European Journal of Information Systems (2020): 1-16
85. Wojciszke, B. et al. (1998) On the dominance of moral categories in impression formation. Pers. Soc. Psychol. Bull. 24, 1245–1257

86. Wu, J., Chen, J., & Dou, W. (2016). The Internet of Things and interaction style: the effect of smart interaction on brand attachment. Journal of Marketing Management, 33(1–2), 61–75. https://doi.org/10.1080/0267257x.2016.1233132
87. Yzerbyt, V. Y., Kervyn, N., & Judd, C. M. (2008). Compensation versus halo: The unique relations between the fundamental dimensions of social judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1110-1123.