Typefaces and the Perception of Humanness in Natural Language Chatbots

Heloisa Candello
IBM Research | Brazil
hcandello@br.ibm.com

Claudio Pinhanez
IBM Research | Brazil
csantosp@br.ibm.com

Flavio Figueiredo
Universidade Federal de Minas Gerais
flaviovdf@dcc.ufmg.br

ABSTRACT
How much do visual aspects influence the perception of users about whether they are conversing with a human being or a machine in a mobile-chat environment? This paper describes a study on the influence of typefaces using a blind Turing test-inspired approach. The study consisted of two user experiments. First, three different typefaces (OCR, Georgia, Helvetica) and three neutral dialogues between a human and a financial adviser were shown to participants. The second experiment applied the same study design but OCR font was substituted by Bradley font. For each of our two independent experiments, participants were shown three dialogue transcriptions and three typefaces counterbalanced. For each dialogue typeface pair, participants had to classify adviser conversations as human or chatbot-like. The results showed that machine-like typefaces biased users towards perceiving the adviser as machines but, unexpectedly, handwritten-like typefaces had not the opposite effect. Those effects were, however, influenced by the familiarity of the user to artificial intelligence and other participants' characteristics.

Author Keywords
User experience; Typography; Dialogue systems; Chatbots.

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous;

INTRODUCTION
The widespread popularity of messenger systems like *Whatsapp* combined with the explosion in automatic dialog technology sparked by the arrival of *Watson*, *Siri*, *Cortana* and similar systems, has materialized the dreams of the AI researchers of the 70s into the everyday lives of most users. New automatic conversational robots, or *chatbots*, are appearing every day in the western world, after widespread acceptance since 2014 in the Chinese messenger platform *WeChat* [45]. Although some of the chatbots are not more than textual representational of menus, the availability of natural language processing open source libraries and API services is enabling the deployment of chatbots which converse using sophisticated natural language capabilities.

It has been shown in many studies that users easily resort to different kinds and modes of personification when interacting with computers (for instance, in the classical study of Reeves and Nass [33]), with significant impacts on the user experience and adoption of those systems. As the linguistic capabilities of chatbots increase, it is expected that their users will become even more likely to ascribe human traits to chatbots, to the point that distinguishing a chatbot from a human being by simply looking into the generated dialog, or by interacting with it, becomes a daunting task.

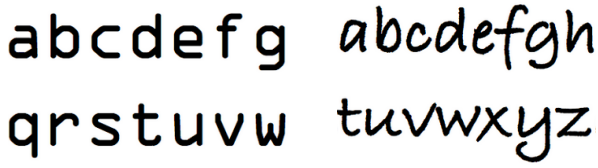
In some ways, users may soon be living in an environment where they are performing involuntary Turing Tests many times a day, having to remind themselves they are interacting with machines not people constantly. Besides some clear ethical implications, user awareness of the humanness of a dialog interlocutor have strong implications in the quality of the interaction. For instance, users tend to have less trust in machine advisers than human advisers, especially when advisers they are perceived as erring [12].

There have been many studies which have looked into the many different factors which affect the perception of humanness by users, such as reciprocity, expressions of emotion, etc. [20,32,26,35]. The absolute majority of those studies are concentrated in how *content* influences the machine-likeness of a conversational system. The aim of this study is to examine whether the perception of humanness in chatbots is also affected by *form*, here translated in one of its simplest elements in a chat, the *typeface* used by the bot.

Typeface elements are known to affect the way people perceive books, packages, signs, and screen interfaces. This study focuses on validating whether the choice of typefaces play a significant role in how users perceive a chatbot. To explore this question, we designed two experiments where users were exposed to 3 typography styles in three, counterbalanced, neutral static dialogues between a human

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI 2017, May 6–11, 2017, Denver, CO, USA.
© 2017 ACM ISBN 978-1-4503-4655-9/17/05...$15.00.
DOI: http://dx.doi.org/10.1145/3025453.3025919

user and a *dialog agent,* which the participants had to determine whether it/she was as human or machine. The first experiment included one dialogue where the dialog agent used a machine-like typeface (OCR-A, Figure 1), while employing neutral typefaces (Helvetica and Georgia) in the two other dialogues. In the second experiment we substitute the OCR typeface by a typeface that mimics handwriting (Bradley, Figure 2). The experiments included collection of demographics and were followed by a small survey.



Figures 1 and 2: OCR-A and Bradley typefaces.

However, based on our previous experience and some previous research [32,13,2,22,27,28,1,24,17] we were also concerned that asking participants about human traits could be very susceptible to (i) the familiarity of the participant with natural language and artificial intelligence systems; (ii) knowledge of design, and (iii) to the repetition of the task. Thus, we designed the two experiments in a way (detailed later) that could let us also explore whether there are significant changes in the behavior of the participants along those three lines. Formally, these are the two key questions we aimed to explore in this study:

RQ1: Are machine-like typefaces (OCR-A) more perceived as machines in a chat?

RQ2: Are typefaces which mimic human handwriting (Bradley) more perceived as human in a chat?

The practical implication of this study is to understand different connotations of typographic styles in the context of chat environments. By understanding typeface effects and what aspects influence those effects, we can expect to help chatbot designers to make conscious visual element choices for creating chatbots dialogues.

Our results from those two research questions led us to further explore our datasets in regards to: (i) the effect of successive dialogues; (ii) the effect of familiarity with technology and design. The first issue isolates the initial perception of participants, filtering out any cognitive learning that may arise from successive exposition. The second is related to past experiences that have been shown to affect perceptions [22].

This paper is organized as follows. First, we give an overview of related research about user perception of chatbots. Second, we detail how the experiment was conducted and describe the participants' profiles. Third, we describe our data analysis process and show our results followed by a discussion section. Finally, we present our conclusions and further work.

RELATED WORK

The act of reading is a complex cognitive process which involves language, memory, thought and intelligence, and perception. Reading comprehends two main types of processes: lexicon processes and comprehension processes. First, people identify letters and words, then activate relevant information from memory related to those words. Second, people process all the text to understand it. To be more specific, people detect letters presented in diverse typography styles and sizes, translate letters into sounds, string letters together to create a word, identify the word, and guess what this word means [37]. Subsequently, meaning might be affected by environment, knowledge, and syntactic elements such as typographic styles.

The effect of typefaces on human's perception

Choosing the best suitable elements for an interface is part of the design process but it can also be a dynamic process. Ishizaki [23] proposed a multi-agent model of dynamic design that facilitates designers to think along the process. It considered design elements as multi-agents. In his framework, design element agents would adapt to dynamic changes in information and user's intention. This framework is suitable in nowadays context, but the big challenge is to understand how those elements should behave to fit context and provide optimal user experience.

In this paper, we focus on perception of a particular design element: typeface. Typography has been researched to best understand reader's comprehension [29] and legibility [4,42]. Perceptions of typefaces have been investigated to understand emotion created into the reader [18], semantic associations on brand perceptions [10], and effects onscreen documents [16]. Typeface's personality [27,28,44], feelings associated to a typeface's tone [11], type connotative meaning [34], and perception relation to taste have also been explored [41]. With the aim to improve visual cues in text-communication, authors investigated *Kinect* typography in instant messaging and animated text [5,43,46,8,25]. Text and typefaces are one of the main elements of current chatbots, but other elements might clearly also affect the conversation experience with machines. In the next subsection, we point out previous literature in this subject.

Conversation experience with machines

The conversation experience with human and non-human beings is often convoluted. Humans are not aware of machine capabilities and machine interfaces not always show what they know from humans. Hidden interactions do not set knowledge limits and user expectations. Ness and Moon [31] performed a set of experiments during 10 years showing people behave with computers as they would behave with real people. They identified that people rely on social categories when interacting with computers. Novielli et al.[32] evaluated people's perception of interactive virtual agents, examining the relationship between input (written vs speech-based) and social attitude users towards the agent, seeing that tech-based interaction warms-up the

user's attitude towards the agent. This effect was more evident for users with background in humanities, and computer scientists tended to be colder and formal to agents, testing the limits of the artificial intelligence with tricky questions. Li et al. [35] performed an experiment that added strength to the hypothesis of people expecting robots to behave like humans. They run an experiment where a robot instructed subjects to touch its body parts, and showed that participants were more emotionally aroused when the robot asked to be touched in areas where people usually do not touch people, like eyes and buttocks.

Those seem to point to a general trait where users' experience with machines is expected to happen like it would with real people. The challenge is to address this complexity of communication with all the characteristics people expect from interactions with humans, such as greetings, understanding, and politeness. This is compounded with complexity of a dialogue where visual cues of traditional graphical interfaces (e.g. icons and buttons) are not the main channel of communication and are not usually explored in text visual styles.

Since 1994, researchers have conducted studies to understand parameters that would affect perception of text messages in dialogues. Brennan et al.[6] tested the effects of message style on dialog and on people's mental models of computer agents. Hill [20] compared human-human online conversations and human conversations with the chatbot *Cleverbot*. Researchers found that people adapt their communication styles accordingly to the other conversant, regardless they are a machine or human.

The Computer Human Interaction research community has been exploring perception effects of intelligent systems for a long time. Terada [39] found that perceptions of familiarity and intelligence of agent appearance are key factors in persuasion buying behavior. Muralidharan [30] found that human speech have higher ratings of trust than machine-like speech and that lower pitch range and greater time delay are identified as more machine-like speech. Smith [36] compared the use of implicit and explicit gift emoticons in care scenarios dialogues and did not find any difference of perception. Strait [38] validated the Masahiro Mori's uncanny valley hypothesis that "too human-like" agents often create repulsion and avoidance.

Chatbot perceptions: Human-like or machine-like?

The most famous test of a machine humanness is the Loebner Prize in Artificial Intelligence which the most recent incarnation of the Turing Test [40]. In this competition, human participants and machines try to convince judges about their humanness dialoguing through computer terminals without seeing each other. The judges should decide which terminals were controlled by humans and machines. Several researchers [26] identified some key reasons why human participants have been perceived as a machine by judges of Loebner Prize. Equilibrium in reciprocity of the exchanges between humans and chatbots; the use of more articles and sophisticated words and also depending on judges' pattern behavior were identified as dependent factors to judge humanness.

Several parameters might affect the way chatbots' humanness is perceived by humans. In our study, we consider a modified Turing test where people selected whether financial advisers were machine-like or human-like based on chat dialogues between them and a customer. To the best of our knowledge, there is no research focusing on human perception of text and typographic styles in chatbot advisers. Thus, more research is needed in order to provide studies to guide designers on those matters.

MATERIALS AND METHODS

Based on previous research in other areas, as discussed in the previous section, we expected that typography styles would affect people's identification of a text as written by a machine or human in a dialogue. We expected that calligraphic styles (like Bradley) would be recognized as more human-like than machine-like typefaces (like OCR). We also hypothesized that people more exposed to artificial intelligence would be less willingly to attribute humanness to chatbots and that designers would be more aware of typefaces than people with a different background.

Experimental Procedure

To explore the validity of our expectations in the context of the two research questions listed in the introduction, a quantitative and a qualitative approach was taken. The experiment was run as a between-subject design.

Two quantitative experiments were conducted where participants were exposed to three different dialogues of about 10 utterances between a human and a financial adviser. For each dialogue, the participant was given time to read and explore it and, at the end, had to classify each financial adviser as *human* or *machine*. Each participant repeated the procedure for the three different dialogues and answered a survey at the end. Our participants were native speakers of Portuguese. All dialogues and survey questions were written in Portuguese. None of our subjects participated in both experiments.

Participants were not made aware that in each different dialogue the typeface of each financial adviser was different. In the first experiment, referred here as the *OCR experiment,* we employed the typefaces OCR-A, Georgia, and Helvetica. The three dialogues were constructed by hand aiming to be as neutral as possible in terms of bias towards human or machine (later verified experimentally). Participants were exposed to three random combinations of the three dialogues and of the three typefaces, counterbalanced. The second experiment applied the same study design, except that the OCR-A font was substituted by Bradley font, and referred here as the *Bradley experiment*.

Those were blind experiments where we masked the real intent of testing typefaces, since setting people's

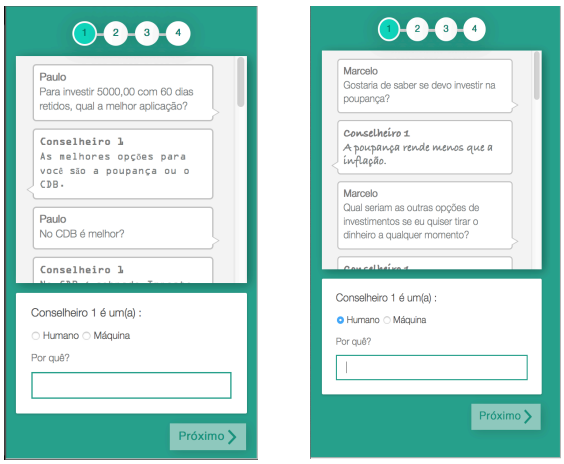


Figure 3: The study environment for the experiments.

expectations before experiments can change actual sensory experience shaping unreal concepts [1,31]. People were asked solely to identify whether the financial adviser was a human or a machine. Participants also answered a short survey after seeing all the dialogues counterbalanced. The survey questions were: 1) What is your age? 2) Do you have formal education in design? If not, what is your background? 3) Have you ever talked to any system that uses artificial intelligence? 4) How do you like to be identified? Male/Female/I prefer not. 5) Did you notice that the typeface (fonts) of advisers were different? Given that the definition of AI is a complex subject, in question 3, we considered the subjects' own interpretation and beliefs about what an intelligent system are. Because of this, familiarity with AI was ultimately determined by the subjects' answer to this question.

Some of the participants did the experiment in a lab environment where qualitative aspects could be observed and measured. In the 15-30 minute qualitative sessions 39 participants were asked to think aloud and share their thoughts with researchers while doing the experiment on a laptop computer. During the experiment, researchers asked about the reasoning behind their choices and give examples from the dialogues to clarify their opinions in case participants did not mention them. Other participants (160), did the experiment in its purely quantitative form as a web-based online test. Both lab- and web-based participants experienced exactly the same web environment and answered the same survey.

Figure 3 shows samples of the web-based environment of the experiments. Participants were invited to read the text and give their opinions, thinking aloud in the qualitative lab tests, or in a form in case of the online experiments. Figure 3(left) shows an excerpt of one of the dialogues, where Paulo, a user, asks: "*What is the best way to invest 5,000.00 with 60 days without liquidity?*" in Helvetica font. The financial adviser answers: "*The best options for you are savings and CDB*" in OCR-A font (CDBs are a popular

Participants		OCR study	Bradley study
Overall	199	90	109
Age group	30-34	52	53
Designers	Yes	15	28
	No	75	81
Work in an AI company	Yes	46	45
	No	44	64
Gender	Male	49	57
	Female	39	49
	No Ans.	2	3
Used AI before	Yes	65	61
	No	25	48

Table 1: Participants' demographics.

Treasury bond). Paulo replies: "*Is CDBs the best choice?*" In Figure 3(right) Marcelo, another person in the dialogue asking for advice, says: "*Should I leave my money in savings?*" in Helvetica font. The financial adviser replies: "*Savings pays less than the rate of inflation.*" in Bradley font. Then Marcelo asks: "*What would be other investment options if I want to retrieve the money at any time?*" After exploring the dialogue, participants had to answer the question at the bottom of the screen. "Is the adviser Human or Machine" and were asked to type an explanation in the form.

Typefaces were chosen according to the type classification by Bringhurst [7]: Sans-serif (Helvetica); Serif (Georgia), Monospace (OCR-A) and Script (ITC Bradley Hand). The last two typefaces were chosen based on type shapes and context of use. *OCR (Optical character recognition)* typefaces were created to be machine-readable and human-readable. OCR-A is a typeface people are often exposed to on food packages, serial numbers, and electronics..

ITC Bradley Hand is a script typeface based on the handwriting of the designer Richard Bradley, also the typeface creator. It was chosen due to similarity of human hand-writing (see Figure 2). *Helvetica*, a sans serif font, was the typeface chosen for the human interacting with the adviser in the dialogue, since it is considered a neutral typeface [21]. We also used Helvetica in one of the combinations, in which participants saw Helvetica as the typeface of both the adviser and the human. The intent was not to have a high discrepancy of typeface form to avoid participants noticing the experiment aim. For this reason, *Georgia*, a serif typeface, was also used in the study, since Helvetica is not perceived as significantly different from Serif like typefaces [34].

In this study, we decided to contrast a more machine-like (OCR-A) and a more human like (Bradley) typeface. This contrast showcases two extreme cases, a hard-edged

Percentage of cases perceived as MACHINE in the OCR Experiment						
95% confidence intervals for an exact binomial test on the OCR Experiment.						
Statistically significant results indicating that average fractions are above random chance (50%) are marked with * (p < 0.1); ** (p < 0.05); *** (p < 0.01)						
		count	**OCR**	**Georgia**	**Helvetica**	**All Fonts**
Overall		90	63 +- 11 %*	51 +- 11 %	51 +- 11 %	55 +- 6 %
Designers	Yes	15	60 +- 28 %	53 +- 27 %	67 +- 28 %	60 +- 16 %
	No	75	64 +- 12 %*	51 +- 12 %	48 +- 12 %	54 +- 7 %
Work in an AI company	Yes	46	74 +- 15 %**	59 +- 15 %	54 +- 15 %	62 +- 9 %**
	No	44	52 +- 16 %	43 +- 15 %	48 +- 15 %	48 +- 9 %
Gender	Male	49	57 +- 15 %	55 +- 15 %	57 +- 15 %	56 +- 8 %
	Female	39	72 +- 17 %*	46 +- 16 %	46 +- 16 %	55 +- 9 %
	No Ans.	2	50 +- 49 %	50 +- 49 %	0 +- 0 %	33 +- 29 %
Used AI before	Yes	65	66 +- 13 %*	51 +- 13 %	58 +- 13 %	58 +- 7 %*
	No	25	56 +- 21 %	52 +- 21 %	32 +- 17 %	47 +- 12 %
Perceived different fonts	Yes	71	59 +- 12 %	48 +- 12 %	55 +- 12 %	54 +- 7 %
	No	11	91 +- 32 %*	73 +- 34 %	27 +- 21 %	64 +- 19 %
	Now	8	62 +- 38 %	50 +- 34 %	50 +- 34 %	54 +- 21 %
All pvalues are computed after a bejamini hochbergh correction with FDR=0.1						

Percentage of cases perceived as MACHINE in the Bradley Experiment						
95% confidence intervals for an exact binomial test on the OCR Experiment.						
Statistically significant results indicating that average fractions are above random chance (50%) are marked with * (p < 0.1); ** (p < 0.05); *** (p < 0.01)						
		count	**Bradley**	**Georgia**	**Helvetica**	**All Fonts**
Overall		109	48 +- 10 %	57 +- 10 %	61 +- 10 %*	55 +- 6 %
Designers	Yes	28	54 +- 20 %	68 +- 20 %	43 +- 18 %	55 +- 11 %
	No	81	46 +- 11 %	53 +- 11 %	68 +- 11 %**	56 +- 6 %
Work in an AI company	Yes	45	42 +- 15 %	58 +- 16 %	78 +- 15 %***	59 +- 9 %
	No	64	52 +- 13 %	56 +- 13 %	50 +- 13 %	53 +- 7 %
Gender	Male	57	47 +- 13 %	61 +- 14 %	63 +- 14 %	57 +- 8 %
	Female	49	47 +- 14 %	53 +- 15 %	61 +- 15 %	54 +- 8 %
	No Ans.	3	67 +- 57 %	33 +- 32 %	33 +- 32 %	44 +- 31 %
Used AI before	Yes	61	49 +- 13 %	51 +- 13 %	62 +- 13 %	54 +- 8 %
	No	48	46 +- 14 %	65 +- 15 %	60 +- 15 %	57 +- 9 %
Perceived different fonts	Yes	98	46 +- 10 %	59 +- 10 %	61 +- 10 %	55 +- 6 %
	No	6	67 +- 44 %	33 +- 29 %	67 +- 44 %	56 +- 25 %
	Now	5	60 +- 45 %	40 +- 35 %	60 +- 45 %	53 +- 27 %
All pvalues are computed after a bejamini hochbergh correction with FDR=0.1						

Tables 2 and 3: Percentage of cases perceived as MACHINE in OCR (left) and Bradley (right) experiments.

typewriter-style font against another that mimics human writing. This also led to a more controlled experiment that compares two ends of a spectrum (more machine like or more human like).

Participants

Participants were initially recruited by a snowball sample under three conditions: less than forty years old; well-educated (university student or higher) and limited knowledge of finance. Further steps in the snowball maintained those conditions. By choosing subjects under forty we tried to avoid subjects who could not be familiar with chat systems. Additionally, having subjects with strong financial knowledge could impact results, i.e., experts could judge the agents by the quality of the advice. The nature of the study requires avoiding expertise effects.

Participants were also initially recruited based on their professional condition - working in an AI company or not - since we wanted to understand if this factor affected participants' perceptions. According to previous experiments [22] designers have different perceptions of typefaces so we also asked participants about their design experience.

Table 1 depicts the main data on the demographics of both experiments. Overall, 199 people participated in the study: 90 participants did the OCR experiment and 109 participants did the Bradley experiment. Thirty-nine participants did the experiments in the lab (which include qualitative data gathering as described before) and 160 performed the experiment online. Half of the participants were 30-34 years old in both studies but in the second study the other half of the people were younger. We had a similar proportion of males and females in both studies, although more males did both studies.

Designers were also part of the study since there is evidence that people from a background in visual design might show categorical perception of typefaces [13,22]. However, we did not have a high number of designers in our study (15 in the OCR and 28 in the Bradley). Most of the participants had previously used artificial intelligent systems (65 to 25 in the OCR, 61 to 48 in the Bradley). However, some of the participants who shared their thoughts with us considered artificial intelligence calls with a recorded speech helpdesk as an artificial intelligence system. About half of the participants were employees of a company that has artificial intelligence systems in its portfolio. Our demographic could not be made not exactly the same for both studies and for this reason we show the results of those two studies separately.

Data Analysis

To test for statistical significance, we used binomial tests where the null hypothesis states that ($p < 50\%$). Rejecting this hypothesis serves as evidence that the typeface leads to participants indicating that adviser messages were humans, when measuring p_h, or machines, when measuring p_m. To avoid false positives we made use of tests based on exact *Clopper-Pearson* confidence intervals. This approach is suitable for small sample sizes [15], as is our case. Moreover, we corrected our *p-values* using a *Benjamini-Hochberg* approach [19], setting the false discovery rate to 0.1. Finally, we present the 95% confidence intervals in all our results, marking the cases where the *p-value < 0.01* (***), *p-value < 0.05* (**), *p-value < 0.1* (*) for the hypothesis tests ($p < 50\%$).

In order to provide further evidences on the impact of typefaces and demographic features, we also resorted to logistic regressions [19]. Here, the dependent variable is the choice of a participant. Indication if an adviser is a machine is the positive class, while indication if the adviser is human accounts for the negative class. In order to select the best model, we made use of a subset selection approach [19].

Percentage of cases perceived as MACHINE for the OCR Experiment										
95% confidence intervals for an exact binomial test on the First Experiment.										
Statistically significant results indicating that average fractions are above random chance (50%) are marked with * (p < 0.1); ** (p < 0.05); *** (p < 0.01)										
			OCR		Georgia		Helvetica		All Fonts	
		count	First Chat	Other Chats	First Chat	Other Chats	First Chat	Other Chats	First Chat	Other Chats
Overall		90	88 +- 19 %***	54 +- 13 %	53 +- 19 %	50 +- 13 %	71 +- 18 %*	38 +- 13 %	70 +- 11 %***	48 +- 7 %
Design	Yes	15	100 +- 71 %	50 +- 29 %	50 +- 38 %	56 +- 34 %	83 +- 47 %	56 +- 34 %	73 +- 28 %	53 +- 19 %
	No	75	86 +- 21 %**	55 +- 14 %	54 +- 21 %	49 +- 14 %	69 +- 20 %	35 +- 13 %	69 +- 12 %**	47 +- 8 %
Work in an AI company	Yes	46	88 +- 25 %**	66 +- 20 %	73 +- 34 %	54 +- 18 %	72 +- 26 %	43 +- 18 %	78 +- 15 %***	54 +- 11 %
	No	44	88 +- 40 %	44 +- 17 %	42 +- 22 %	44 +- 20 %	71 +- 27 %	33 +- 17 %	61 +- 16 %	41 +- 10 %
Gender	Male	49	92 +- 28 %**	44 +- 17 %	50 +- 27 %	57 +- 18 %	77 +- 23 %*	41 +- 18 %	73 +- 15 %**	48 +- 10 %
	Female	39	82 +- 34 %	68 +- 20 %	60 +- 28 %	38 +- 19 %	62 +- 30 %	38 +- 18 %	67 +- 17 %	49 +- 11 %
	No Answ	2	100 +- 98 %	0 +- 0 %	0 +- 0 %	100 +- 98 %	0 +- 0 %	0 +- 0 %	50 +- 49 %	25 +- 24 %
Used AI before	Yes	65	89 +- 23 %**	57 +- 15 %	57 +- 23 %	48 +- 15 %	88 +- 19 %***	40 +- 15 %	78 +- 12 %***	48 +- 9 %
	No	25	83 +- 47 %	47 +- 23 %	44 +- 31 %	56 +- 26 %	30 +- 23 %	33 +- 22 %	48 +- 20 %	46 +- 14 %
Perceived different fonts	Yes	71	93 +- 25 %**	50 +- 14 %	48 +- 21 %	48 +- 15 %	73 +- 18 %*	39 +- 15 %	69 +- 12 %**	46 +- 8 %
	No	11	86 +- 44 %	100 +- 60 %	75 +- 56 %	71 +- 42 %	0 +- 0 %	27 +- 21 %	82 +- 34 %	55 +- 22 %
	Now	8	67 +- 57 %	60 +- 45 %	67 +- 57 %	40 +- 35 %	50 +- 49 %	50 +- 38 %	62 +- 38 %	50 +- 25 %
All pvalues are computed after a bejamini hochbergh correction with FDR=0.1										

Table 4: Percentage of cases perceived as MACHINE in OCR experiment separating First Chat.

That is, we tested every demographic feature, typefaces, and also on which dialogue the typeface was presented to the user as explanatory variables. Each variable was encoded as an indication factor, taking values of 0 or 1. We also added to our model the pairwise combinations of explanatory variables. Our subset selection considered every possible subset of variables as candidate models. In the end, results are all based on the model with the lowest AIC score [19] (lower means simpler but also explanatory).

The qualitative data analysis was guided by the main issues present in the notes during the lab sessions and by the open questions on the experiment form. The data was coded after transcribing all the sessions using *NVivo* software. Content analysis was applied as the main method [9]. The key qualitative results are shown as complementary to the analysis of the quantitative results.

RESULTS

We now present the main results of the two experiments in the context of the two research questions posed in the introduction section of this paper. To tackle such questions, we initially computed the percentages (p) of answers marked either as human (p_h) or as machine (p_m) in each experiment. We also computed such percentages for the different demographic groups as shown in Table 2 for the OCR experiment and Table 3 for the Bradley experiment.

On the Neutrality of Dialogues

Before presenting our results, we point out that we used the same tests described above the test whether the dialogues presented to the user where neutral (the three different dialogues fail to pass both the test for being perceived as machine and the test for being perceived as human). We tested the neutrality of dialogues based on a statistical analysis after the experiments were performed. We consider the order of appearance in which the dialogues were shown to the participants (three possible cases). Our results indicated that, on both experiments, dialogues were neutral, with fractions not being statistically different from 50% random chance. This result is valid regardless of the order of appearance of the dialogue. Furthermore, in the lab sessions, participants showed hesitation to choose one of the alternatives (machine or human), having difficulties to reason their choices. This outcome also increases our evidence of the neutrality of the dialogues. The neutrality of the three dialogues is a necessary validation of the experiments since we can then state that the outcomes of the experiments are not dependent on the order of the dialogues shown to the participant or on the pairing with the fonts.

On the Effect of Typefaces

We examine now the two research questions of whether the typeface choice of a dialogue agent affects the perception of its humanness. Next, we discuss other factors that emerged in the analysis.

RQ1: Are machine-like typefaces (such as OCR) more perceived as machines in a chat?

We show in Table 2 and 3 the percentage of cases perceived as MACHINE (p_m) for each of the two experiments. The tables show the fraction per demographic group (rows) with general results belonging to the *Overall* row. Moreover, we also show results per typeface and when considering all fonts combined (last column). Each cell represents a 95% confidence interval in the form "(expected value +- error)", and the level of statistical significance is indicated by the *, **, and *** markings as indicated.

We do not show the percentage of cases considered as humans (p_h). First, the percentage is easily calculated as $p_h = 100 - p_m$. But, more important, one of our most striking results is that we found **not a single context where**

Percentage of cases perceived as MACHINE for the Bradley Experiment separating the First Chat										
95% confidence intervals for an exact binomial test on the First Experiment.										
Statistically significant results indicating that average fractions are above random chance (50%) are marked with * (p < 0.1); ** (p < 0.05); *** (p < 0.01)										
			Bradley		Georgia		Helvetica		All Fonts	
		count	First Chat	Other Chats	First Chat	Other Chats	First Chat	Other Chats	First Chat	Other Chats
Overall		109	67 +- 19 %	41 +- 11 %	74 +- 18 %**	49 +- 12 %	75 +- 15 %**	52 +- 13 %	72 +- 9 %***	47 +- 7 %
Designers	Yes	28	88 +- 40 %	40 +- 21 %	89 +- 37 %	58 +- 24 %	55 +- 31 %	35 +- 21 %	75 +- 20 %*	45 +- 13 %
	No	81	59 +- 23 %	41 +- 13 %	69 +- 21 %	45 +- 13 %	82 +- 17 %***	58 +- 15 %	72 +- 11 %***	48 +- 8 %
Work in an AI company	Yes	45	54 +- 29 %	38 +- 16 %	75 +- 32 %	52 +- 18 %	90 +- 22 %***	68 +- 22 %	76 +- 15 %***	51 +- 11 %
	No	64	76 +- 26 %	43 +- 14 %	74 +- 22 %	46 +- 16 %	62 +- 22 %	42 +- 15 %	70 +- 13 %**	44 +- 9 %
Gender	Male	57	59 +- 26 %	42 +- 15 %	74 +- 25 %	55 +- 17 %	76 +- 23 %*	56 +- 17 %	70 +- 14 %**	51 +- 10 %
	Female	49	77 +- 31 %	36 +- 15 %	79 +- 29 %	43 +- 17 %	77 +- 23 %*	48 +- 19 %	78 +- 14 %***	42 +- 10 %
	No Answ	3	0 +- 0 %	67 +- 57 %	50 +- 49 %	0 +- 0 %	0 +- 0 %	50 +- 49 %	33 +- 32 %	50 +- 38 %
Used AI before	Yes	61	73 +- 28 %	41 +- 14 %	68 +- 23 %	41 +- 15 %	79 +- 21 %**	51 +- 17 %	74 +- 13 %***	44 +- 9 %
	No	48	60 +- 28 %	39 +- 16 %	85 +- 30 %*	57 +- 18 %	70 +- 24 %	54 +- 20 %	71 +- 15 %**	50 +- 10 %
Perceived different fonts	Yes	98	67 +- 24 %	40 +- 11 %	76 +- 18 %**	50 +- 13 %	77 +- 15 %***	49 +- 14 %	74 +- 10 %***	46 +- 7 %
	No	6	60 +- 45 %	100 +- 98 %	0 +- 0 %	40 +- 35 %	0 +- 0 %	67 +- 44 %	50 +- 38 %	58 +- 31 %
	Now	5	75 +- 56 %	0 +- 0 %	0 +- 0 %	40 +- 35 %	0 +- 0 %	75 +- 56 %	60 +- 45 %	50 +- 31 %
All pvalues are computed after a bejamini hochbergh correction with FDR=0.1										

Table 5: Percentage of cases perceived as MACHINE in Bradley experiment separating First Chat.

the chance of messages being perceived as coming from humans (p_h > 50%) was statistically significant. In this sense, our numbers indicate that users in our experiments were either biased towards perceiving chat advisers as machines or they were in doubt. Based on this result we can argue that if typefaces affect perception, the effect biases users towards perceiving the advisers as machines. As a participant mentioned in the OCR experiment *"I will finish this study and you will say that all the dialogues are machines, right? Since machines can be programmed by humans. Will you tell me the result?"* Similarly, a user in the Bradley experiment suggested: *"This is like a game […] should be all robots"*.

Regarding the perception of messages as coming from machines, Table 2 shows results for the OCR experiment. In this case, we found evidence that the OCR typeface can impact perceptions (as shown in the *Overall* columns). In this case the value of p_m is of (63+-11%), and no effect is significant for the two other typefaces. Therefore, we can argue that the OCR experiment provide evidence that the answer for RQ1 is: Yes, machine-like typefaces (such as OCR) are more perceived as machines in a chat. It should be noticed that this effect is small, since the 95% confidence interval starts at 51%. In the worst case scenario, there is only a 2% increased chance over random choice.

RQ2: Are typefaces which mimic human handwriting (such as Bradley) more perceived as human in a chat?

As commented before, all the statistical tests showed that there was no evidence in the experiments that any typeface, including the Bradley typeface, influenced the users toward judging the financial advisers that used that font as human. So, the answer for RQ2 is: No, typefaces that mimic human handwriting (such as Bradley) are not more perceived as human in a chat. In fact, we also found no evidence of Bradley biasing perceptions towards machines as well.

However, the qualitative results presented some counter-evidence here. Some participants claimed the Bradley experiment was trying to fool them showing a handwritten-like typeface. For instance, a participant thinking aloud while reading a text with Bradley typeface commented: *"This one just because changed the font thinks that fools me, I think this one wants to fool me."* This might explain why a script typeface might be considered written by a machine. Were participants being defensive against being fooled by a machine?

Moreover, if we look into the perception of the financial adviser being a machine in the Bradley experiment, as depicted in Table 3, we find some interesting and surprising findings. Participants did not perceive the Bradley dialogues as human, as expected, but showed a statistically significant tendency to perceive the advisers portrayed with the "neutral" Helvetica font as machines (61+-10%*). We did not observe this effect on the Helvetica-rendered dialogues in the OCR experiment (51+-11%). A possible explanation is that users could be considering Helvetica more machine-like in comparison with Bradley. But before going further into this issue, we have to explore more the issue of whether the participants' behavior changed as they went through the three dialogues.

On the Effect of Successive Dialogues

As we started to analyze the data from the two experiments, we were surprised by how judging the adviser as a machine seemed to be more common in the first chat the participants evaluated than the next two. Notice that in our experiments participants evaluate all three typefaces and all three dialogues in a random order and pairing, counterbalanced.

	Coeff.	*p-value*
Used AI before	-0.88	0.035**
Work in an AI company **AND** Used AI before	1.33	0.016**
Used AI before **AND** First Chat	1.23	0.049**
First Chat **AND** OCR	1.12	0.098*
First Chat **AND** Helvetica	1.5	0.073*

Table 6: Best Logistic Regression Model for the OCR Experiment. This model was statistically best than a null model (p-value < 0.001 under a log-likelihood ratio test).

	Coeff.	*p-value*
First Chat	0.98	0.037**
Work in an AI company **AND** Helvetica	1.85	0.01**

Table 7: Best Logistic Regression Model for the Bradley Experiment. This model was statistically best than a null model (p-value < 0.001 under a log-likelihood ratio test).

But the fact that each participant evaluated the first dialogue before being exposed to any other contact with the financial advisers seemed likely to have some interesting and unexpected effects, leading us to raise some interpretations.

In Tables 4 and 5, we detail the results based on when each typeface was evaluated on the first position (*First Chat*) or on the second and third position (*Other Chats*), and also for all three typefaces together (*All Fonts*) Overall, we can see that the first dialogue increases the chance of typefaces being perceived as machines independently of the font being used: in the OCR experiment, the overall bias is (70+-11%***), and (72+-9%***) in the Bradley experiment. On the other hand, there is no statistically significant tendency, either to human or machine, in the second and third chats, as shown by the *Other Chats* columns of Tables 4 and 5. So we can conclude that the two experiments provide evidence that the perception of humanness of dialogue agents' change as participants go through successive evaluations.

From Table 4, we can also see a significant effect not only for the OCR font (which in fact increases compared to the effect for all chats), but also for the Helvetica font. In this sense, our results show that users seem to perceive hard-edged fonts (OCR and Helvetica) as more machine-like when presented on the first chat. The expected value of p_m is of (88+-19%***) for OCR, and of (71+-18%*) for Helvetica. In the Bradley experiment (Table 5) both Helvetica (75+-15%**) and Georgia (74+-18%**) had a significant impact on the perception of dialogues as machines (p_m of about 75%) on the first chat. We will dwell later on the reasons and consequences of this tendency towards machine in the first contact.

However, there is no tendency towards machine or human in the first chat of Georgia in the OCR experiment (53+-19%) or, as expected, in the Bradley first chat of the Bradley experiment (67+-19%). The latter result could be regarded as evidence that Bradley is not able to "humanize" the dialogue agent. This is further evidence towards refuting RQ2. We conjecture later that the Bradley font seems to *at least* confuse users towards believing they are conversing with human agents, perhaps due to its form.

On the Effect of Familiarity with Technology and Design

Finally, we explore impacts of user demographic features on the perception of humanness, particularly in the case of prior exposure to both artificial intelligence and design.

Tables 2, 3, 4 and 5 already contain rows with detailed results for each item of our survey. We summarize these findings using logistic regression based on a subset selection approach [19] to pick the best model based on the AIC score. As our dependent variables, we had each feature, the order of the typeface on the experiment (first or second and third combined), and also the typeface itself. All variables were encoded as an indicator number (0 or 1). Our responses were encoded as 1 for chats perceived as machines and 0 for chats perceived as humans. On the final model, variables with positive signs determine that such effect can be used to explain why a user perceived a chat as machine. Negative ones explain perception as humans. Table 6 shows the results of our best model for the OCR experiment while Table 7 shows the best model for the Bradley experiment. On both tables we only show variables that were deemed significant (*p-value < 0.1).*

Starting with Table 6, the only negative effect is having previously used artificial intelligences (-0.88). Although significant, this effect had the smallest coefficient. It shows that previous conversations with AI can somewhat bias perceptions towards human. Conversely, when combined with the participant being part of an AI company, the tendency reverses towards a machine bias (1.33). This indicates that working at a company with an AI portfolio can bias results towards machines in some cases. As shown in Table 6, there are also first chat effects.

For the Bradley experiment, Table 7, we can see that simply having the Bradley typeface on the first dialogue is the most important effect (*p-value < 0.05)* towards stating that messages were machine generated. Also, workers of an AI company tend mostly to state that Helvetica messages come from a machine. Overall, combining evidences from both tables, we can argue that there is evidence that familiarity with artificial intelligence technology does impact the perception of dialogue agents as produced by a machine. The single counter-example of the negative effect depicted in the first row of Table 6 stems from users that are acquainted with AI but do not work at an AI company and are not evaluating the first chat. In these cases, acquaintance with AI, combined with other unmeasured factors, can somewhat bias perceptions towards humans.

Finally, the results show no significant evidence that design experience [3] affects the perception of humanness. Unfortunately, our sample size (43 participants with design background) might have an impact on this result. Larger samples can mitigate this issue.

Before concluding, from Tables 2, 3, 4 and 5 we also can note that variables like gender can have some impact as seen on when analyzing at a typeface in isolation. Since these effects did not show up in our regressions, understanding them in more details is left as future work.

DISCUSSION

The overall purpose of this study was to better understand the effect of typography choices have on user's perceptions of chatbots. Our qualitative findings helped us to understand some of the quantitative results and reflect on the way this kind of study might be conducted in the future. Our intent to mask the aim of the study, not telling participants we were testing typefaces, was successful. None of the participants who did the experiment in the lab noticed the aim of the study; they seem to be very focused on deciding how to classify the financial advisers as human or machine. In the end of the study, participants often requested to know the truth, asking: *"now tell me, what was the right answer?"* Of course in the end of the session, researchers explained the real aim of the study.

For experiment purposes, we wrote the dialogues to be perceived as neutral, as they were perceived according to our results. It is remarkable that not a single possible context combination (typefaces + dialogues) had a bias to be perceived as coming from humans with statistically significance, in spite of all the messages being actually written by human beings. We expected participants would select OCR as more machine-like (RQ1) but we did not expect Bradley would not be selected as human-like (RQ2). We felt in the qualitative studies that participants seem to be suspicious when they read the text in Bradley, as mentioned before. Trust was a concern participants showed while answering this part of the experiment.

The bias to machine-like was even greater from participants exposed to artificial intelligence technology, defined here as the ones that already used artificial intelligence systems or that worked in an artificial intelligence-related company. These findings align with similar results found by researchers evaluating perception of speech virtual agents by people with technical background knowledge [32]. We should be careful, however, since in the qualitative experiments some participants mentioned they considered simple web-based assistants as artificial intelligence technology; some even considered voice-based recorded helpdesks as artificial intelligence systems. Interestingly, most of them commented that they did not enjoy their previous experiences with such AI technologies.

We also expect designers would perceive typefaces differently. However, no impact was found in our study. It is true that designers who participated in the lab studies recognized more often the difference of typeface styles between dialogues, likewise in previous research [22,13]. But the overall perceptions of the dialogue as being from a human or a machine did not appear to have strong connections to typeface styles for designers.

The most striking results were related to the prominence of the first chat. Participants changed their perceptions or their decision making process as the experiments progressed, having a bias to select machine-like for the adviser in the dialogue they first experienced, independent of typeface. Following, in the second and third chats, the results showed that not even in a single context there was statistically significant bias in either direction. This finding could be basically telling that successive evaluations are necessary to test chat perception; a single exposure might not assure perception results. In the literature, there are divergent opinions in relation to the power of first impression. Gladwell [17] defends that first impressions are accurately and stands the test of time. On the contrary, Ariely [1] argues first impressions not only cannot be trusted but also experiment participants make subsequent judgments based on the first judgment they make. In other words, experiment participants tend to enter into a "comparison" mode after the first impression. Daniel Kahneman [24] offers an explanation based on a dual process model of the brain considering instinctive and immediate judgments and reasoning to check emotional responses.

In general, the overall results of the three chats agree with the results of the first chat, with some bias towards machines in the cases discussed before. But when we consider the participant's perceptions only in the second and third chats there is virtually no difference in any dimension or context. First, an argument can be made that part of the problem is that, when compared to the overall results for the three chats, a smaller number of participants may have rendered some real biases statistically insignificant. Although certainly possible, we observed how significant some of the biases are when considering only the first dialogue, which has an even smaller number of participants. The key difference could be that in the first chat participants did not have any reference to compare with, except their own previous experiences with chatbots. As participants progressed through the experiment, they may have entered into a decision criteria based on comparing chats, that is, trying to determine the humanness of the chat they have in front of them by comparing with the previously seen chats.

We can argue that the above comparison had to be based solely on the memory of the previous chat, given that participants were not allowed to revisit previous chats. Arguments can also be made regarding how participants' memories from previous chats affect subsequent ones. Considering the language of our participants is Portuguese, inferences could also be drawn from cross-language studies

in cognitive psychology [14] . Portuguese is a more syllabic language, in which reading memory is more auditory than logographic languages (e.g. Chinese or Japanese), which have strong visual components. Investigating language effects on perceptions is left for future work.

LIMITATIONS

There were several limitations in this study. First, we have to be careful when we compare the results from the two experiments since the OCR experiment was more well-balanced than the Bradley experiment. In the Bradley experiment, the number of people exposed to artificial intelligence concepts and practice were higher than in the OCR. Additionally, demographic characteristics were similar but not the same. Second, although dialogues were perceived as neutral statistically, in our qualitative evaluation we did find some cases where users picked either human or machine based on extracts of the dialogue text. Nevertheless, to minimize those issues, participants were asked to give their choices based on the overall perception of each dialogue. Third, our dialogues were not interactive (participants evaluated only static dialogues). If subjects could interact freely there would be a lot of variability in the experience as a result of limited answer accuracy, different unanswered questions, etc. which would impact the study in uncontrollable ways. By reading a recorded dialogue, all subjects had basically the same experience and therefore their answers are comparable.

CONCLUSION AND FURTHER WORK

Understanding better chatbot perceptions by humans is an essential element towards building trust and reliance in effective conversations between human and machines. However, identifying which design elements affect people's perceptions and how to select them is a challenge since several attributes can affect the way people perceive those elements. Our results provide a better understand of how typefaces are perceived by people and highlight the impact of first impressions on dialogue perception studies. Our results show evidences and design implications (DI) in regard of:

The effect of typefaces: OCR, a machine-like typeface was more identified as machine in a chat, as we expected. Contrary to our expectations, script typefaces (Bradley) were not more perceived as human in a chat.

DI (1): As some people have biases to perceive typefaces in chatbot dialogues as machine-like, designers, if desired, should use machine-like fonts such as OCR-A to reinforce machine-like perception as well as other dialogue features such as repeated answers, perfect grammatical standards, and tailored content.

The effect of successive dialogues: the effect of the OCR typeface is enhanced in the first dialogue seen by participants. Surprisingly, even messages with more neutral typefaces like Helvetica and Georgia, were biased towards being perceived as created by machines in the first dialogue. But a human writing typeface like Bradley yielded no effect, even considering the first dialogue alone. But the most surprising result here is that in the second and third chats seen by the participants, there is not a single stance where bias was detected for human or machine. As discussed before, this could be caused by participants mentally comparing the second and third chats with the previous ones. This may lead to the loss of the typeface effect due cognitive factors.

DI (2): Typeface effects are often restricted to initial perceptions, so designers might choose typefaces that set the initial dialogue tone and user expectation of chatbots. Although our results indicate that typefaces have limited impact on the overall perception, contrarily to common sense, designers can use our findings to avoid typefaces that try to mimic humans. It could lead to mistrust and affect the adoption of chatbots.

The effect of familiarity with technology and design: previous knowledge and exposure to artificial intelligence chatbots seem to bias users to perceive chats as created by machines. No evidence was found that design experience produces typeface bias.

DI (3): Designers might choose different typefaces according to the AI familiarity of target users, in order to have a desirable first impact. People with less contact with AI might need other elements or content reinforcement to perceive chatbots as machines. Contrary to previous literature on typography [22], designers did not have typeface bias. This issue could be tested further since the number of designers in our experiment was relatively small.

We are seeing a lot of chatbots being deployed to real users in a vacuum of user research and tested design guidelines. The methodology and results of this paper are thus a contribution to improve the quality of chatbots. Our results clearly suggest directions for further investigation on selecting suitable typefaces for chatbots. Further research could explore which kinds of other visual elements, such as color or *emojis*, affect the perception of humanness and trust of chatbots users. Other demographic variables (e.g gender) could also be explored in future studies as to understand their effect on typography perceptions.

ACKNOWLEDGEMENTS

We would like to thank Alan Braz, Breno Antunes, Bruna Andrade and Caio Americo, for helping developing the test environment. We also thank Matthias Kormaksson and David Millen for insightful feedbacks on the data analysis and all the volunteers who participated in the experiments.

REFERENCES

1. D. Ariely. 2010. *Predictably Irrational: The Hidden Forces That Shape Our Decisions*. HarperCollins.
2. Syed Z. Arshad, Jianlong Zhou, Constant Bridon, Fang Chen, and Yang Wang. 2015. Investigating User Confidence for Uncertainty Presentation in Predictive Decision Making. In *Proceedings of the Annual Meeting of the Australian Special Interest Group for*

Computer Human Interaction (OzCHI '15), 352–360. https://doi.org/10.1145/2838739.2838753

3. D. Bartram. 1982. The perception of semantic quality in type: Differences between designers and non-designers. *Information Design Journal* 3, 1: 38–50. https://doi.org/10.1075/idj.3.1.04bar
4. Michael L. Bernard, Barbara S. Chaparro, Melissa M. Mills, and Charles G. Halcomb. 2003. Comparing the Effects of Text Size and Format on the Readibility of Computer-displayed Times New Roman and Arial Text. *Int. J. Hum.-Comput. Stud.* 59, 6: 823–835. https://doi.org/10.1016/S1071-5819(03)00121-6
5. Kerry Bodine and Mathilde Pignol. 2003. Kinetic Typography-based Instant Messaging. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '03), 914–915. https://doi.org/10.1145/765891.766067
6. Susan E. Brennan and Justina O. Ohaeri. 1994. Effects of Message Style on Users' Attributions Toward Agents. In *Conference Companion on Human Factors in Computing Systems* (CHI '94), 281–282. https://doi.org/10.1145/259963.260492
7. Robert Bringhurst. 2002. *The Elements of Typographic Style. Second Edition*. Hartley & Marks Publishers.
8. Barbara Brownie. 2015. *Transforming Type: New Directions in Kinetic Typography.* Bloomsbury Academic.
9. Alan Bryman. 2015. *Social Research Methods*. Oxford University Press.
10. Terry L. Childers and Jeffrey Jass. 2002. All Dressed Up With Something to Say: Effects of Typeface Semantic Associations on Brand Perceptions and Consumer Memory. *Journal of Consumer Psychology* 12, 2: 93–106. https://doi.org/10.1207/S15327663JCP1202_03
11. R. C. Davis and H. J. Smith. 1933. Determinants of feeling tone in type faces. *Journal of Applied Psychology* 17, 6: 742–764.
12. B. J. Dietvorst, J. P. Simmons, and C. Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1: 114–126. https://doi.org/10.1037/xge0000033
13. M. C. Dyson. 2011. Do designers show categorical perception of typefaces? *Visible Language* 45.3, 15.3: 193–220.
14. Eysenck, M.W. and Keane, M.T. 2015. *Cognitive Psychology: A Student's Handbook. Psychology Press*. Taylor & Francis.
15. Joseph Fleiss. et al. 2003. *Statistical Methods for Rates and Proportions*. Wiley.
16. Doug Fox, A. Dawn Shaikh, and Barbara S. Chaparro. 2007. The Effect of Typeface on the Perception of Onscreen Documents. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 51, 5: 464–468. https://doi.org/10.1177/154193120705100508
17. Malcolm Gladwell. 2007. *Blink: The Power of Thinking Without Thinking*. Back Bay Books.
18. J. E. Gump. 2001. The Readability of Typefaces and the Subsequent Mood or Emotion Created in the Reader. *Journal of Education for Business* 76, 5: 270–273. https://doi.org/10.1080/08832320109599647
19. Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. 2009. *The elements of statistical learning : data mining, inference, and prediction*. Springer, New York. Retrieved from http://opac.inria.fr/record=b1127878
20. J. Hill, W. R. Ford, and I. G. Farreras. 2015. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior* 49: 245–250. https://doi.org/http://dx.doi.org/10.1016/j.chb.2015.02.026
21. Hustwit, G. 2007. Helvetica Video. (2007). Retrieved August 15, 2016 from `http://www.hustwit.com/category/helvetica`
22. Sarah Hyndman. 2016. *Why Fonts Matter*. Virgin Books.
23. Suguru Ishizaki. 1996. Multiagent Model of Dynamic Design: Visualization As an Emergent Behavior of Active Design Agents. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '96), 347–354. https://doi.org/10.1145/238386.238566
24. Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.
25. Jun Kato, Tomoyasu Nakano, and Masataka Goto. 2015. TextAlive: Integrated Design Environment for Kinetic Typography. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3403–3412. https://doi.org/10.1145/2702123.2702140
26. C.L. Lortie and M. J. Guitton. 2011. Judgment of the Humanness of an Interlocutor Is in the Eye of the Beholder. *PLOS ONE* 6, 9: 1–7. https://doi.org/10.1371/journal.pone.0025085
27. Jo MacKiewicz. 2005. How to Use Five Letterforms to Gauge a Typeface's Personality: A Research-Driven Method. *Journal of Technical Writing and Communication* 35, 3: 291–315. https://doi.org/10.2190/LQVL-EJ9Y-1LRX-7C95

28. J. Mackiewicz and R. Moeller. 2004. Why people perceive typefaces to have different personalities. In *International Professional Communication Conference, 2004. IPCC 2004. Proceedings.*, 304–313. https://doi.org/10.1109/IPCC.2004.1375315
29. Sara J. Margolin. 2013. Can Bold Typeface Improve Readers' Comprehension and Metacomprehension of Negation? *Reading Psychology* 34, 1: 85–99. https://doi.org/10.1080/02702711.2011.626107
30. Laya Muralidharan, Ewart J. de Visser, and Raja Parasuraman. 2014. The Effects of Pitch Contour and Flanging on Trust in Speaking Cognitive Agents. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '14), 2167–2172. https://doi.org/10.1145/2559206.2581231
31. Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1: 81–103. https://doi.org/10.1111/0022-4537.00153
32. Nicole Novielli, Fiorella de Rosis, and Irene Mazzotta. 2010. User attitude towards an embodied conversational agent: Effects of the interaction mode. *Journal of Pragmatics* 42, 9: 2385–2397. https://doi.org/10.1016/j.pragma.2009.12.016
33. Byron Reeves and Clifford Nass. 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. Cambridge University Press, New York, NY, USA.
34. Camille L. Rowe. 1982. The connotative dimensions of selected display typefaces. *Information Design Journal* 3, 1: 30–37. https://doi.org/http://dx.doi.org/10.1075/idj.3.1.03row
35. Touching a robot can elicit physiological arousal in humans: Participants were more hesitant to touch a robot's intimate parts when instructed. *ScienceDaily*. Retrieved August 8, 2016 from https://www.sciencedaily.com/releases/2016/04/160405093057.htm
36. Kirsten A. Smith. 2015. Assessing the Supportiveness of Gift Emoticons in Care Scenarios. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15), 151–156. https://doi.org/10.1145/2702613.2726969
37. Robert J. Sternberg and Karin Sternberg. 2016. *Cognitive psychology*. Nelson Education.
38. Megan Strait, Lara Vujovic, Victoria Floerke, Matthias Scheutz, and Heather Urry. 2015. Too Much Humanness for Human-Robot Interaction: Exposure to Highly Humanlike Robots Elicits Aversive Responding in Observers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15), 3593–3602. https://doi.org/10.1145/2702123.2702415
39. Kazunori Terada, Liang Jing, and Seiji Yamada. 2015. Effects of Agent Appearance on Customer Buying Motivations on Online Shopping Sites. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (CHI EA '15), 929–934. https://doi.org/10.1145/2702613.2732798
40. Turing, A.M. 1950. Computing machinery and intelligence. *Mind.* 59, 236 (1950), 433–460.
41. C. Velasco, A. T. Woods, S. Hyndman, and C. Spence. 2015. The Taste of Typeface. *i-Perception* 6, 4. https://doi.org/10.1177/2041669515593040
42. Jean-Luc Vinot and Sylvie Athenes. 2012. Legible, Are You Sure?: An Experimentation-based Typographical Design in Safety-critical Context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '12), 2287–2296. https://doi.org/10.1145/2207676.2208387
43. Vinot, J.-L. and Athenes, S. 2012. Legible, Are You Sure?: An Experimentation-based Typographical Design in Safety-critical Context. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA, 2012), 2287–2296.
44. Hua Wang, Helmut Prendinger, and Takeo Igarashi. 2004. Communicating Emotions in Online Chat Using Physiological Sensors and Animated Text. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '04), 1171–1174. https://doi.org/10.1145/985921.986016
45. Timothy Tien-Lou Wang. 2013. *Fonts and Fluency: The Effects of Typeface Familiarity, Appropriateness, and Personality on Reader Judgments*. Master thesis. University of Canterbury, Christchurch, New Zealand.
46. Wang, Y. et al. 2015. Dwelling and Fleeting Encounters: Exploring Why People Use WeChat - A Mobile Instant Messenger. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (New York, NY, USA, 2015), 1543–1548.
47. Zhiquan Yeo. 2008. Emotional Instant Messaging with KIM. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '08), 3729–3734. https://doi.org/10.1145/1358628.1358921