Forthcoming, *Information Systems Research*

Estimating the Impact of 'Humanizing' Customer Service Chatbots

Scott Schanke
University of Minnesota Carlson School of Management, schan067@umn.edu,

Gordon Burtch
University of Minnesota Carlson School of Management, gburtch@umn.edu,

Gautam Ray
University of Minnesota Carlson School of Management, gautamr@umn.edu,

We study the impacts of 'humanising' AI-enabled autonomous customer service agents (chatbots). Implementing a field experiment in collaboration with a dual channel clothing retailer based in the United States, we automate a used clothing buy-back process, such that individuals engage with the retailer's autonomous chatbot to describe the used clothes they wish to sell, obtain a cash offer, and (if they accept the offer) print a shipping label to finalize the transaction. We causally estimate the impact of chatbot anthropomorphism on transaction conversion by randomly exposing consumers to exogenously varied levels of chatbot anthropomorphism, operationalized by incorporating a random draw from a set of three anthropomorphic features: humor, communication delays and social presence. We provide evidence that, in this retail setting, anthropomorphism is beneficial for transaction outcomes, but that it also leads to significant increases in offer elasticity. We argue that the latter effect occurs because, as a chatbot becomes more human-like, consumers shift to a fairness evaluation or negotiating mindset. We also provide descriptive evidence suggesting that the benefits of anthropomorphism for transaction conversion may derive, at least in part, from consumers' increased willingness to disclose personal information necessary to complete the transaction.

Key words: chatbot, artificial intelligence, intelligence augmentation, human computer interaction, field experiment, customer service, anthropomorphism

1. Introduction

Researchers, the general public and organizations alike have become enamored with Artificial Intelligence (AI). With recent breakthroughs in the field, coupled with changes in public perception and advances in hardware, society has seen AI technologies move to the main stage. Organizations are looking to capitalize by putting these technologies into practice to both capture value, and to hedge against the possibility of disruption.[1] AI technologies have seen widespread implementation in a variety of domains, from fraud detection, to image recognition, voice recognition and natural language processing (Dale 2016). Gartner predicts that 2.3 million AI-related jobs will be created by the year 2020.[2]

Although media and public interest have caused AI to reach what Gartner refers to as a state of "inflated expectations", there is clear value in these technologies, if they are used appropriately and expectations are managed. One prominent example of an AI-based tool that has seen widespread adoption and value creation for firms of all sizes is the text-based 'chatbot'. Chatbots are autonomous software agents that support text-based exchanges with human users, drawing on tools and techniques from the domain of Natural Language Processing. Chatbots have the potential to automate basic, repeatable, standardized customer service interactions, relieving the need for those interactions to be handled by human employees.[3] Recognizing the potential of these sorts of AI-based autonomous agents, firms are adopting them at an extremely rapid pace. Google Search Trends indicates that interest in chatbots has grown by an order of magnitude in the last two years (see Figure 1), and industry estimates forecast that, by 2020, conversations with autonomous agents will be more common for the average individual than conversations with a spouse.[4]

The anticipated volume of customer interactions these digital agents will be expected to handle suggests that chatbots will soon become the main point of customer contact for many retail

[1] https://www.forbes.com/sites/joemckendrick/2018/01/25/artificial-intelligence-isnt-killing-jobs-its-killing-business-models/#64e5388a5ea0

[2] https://www.gartner.com/newsroom/id/3837763

[3] https://chatbotsmagazine.com/chatbots-vs-apps-the-final-frontier-a0df10861c48

[4] https://www.gartner.com/smarterwithgartner/gartner-predicts-a-virtual-world-of-exponential-change/

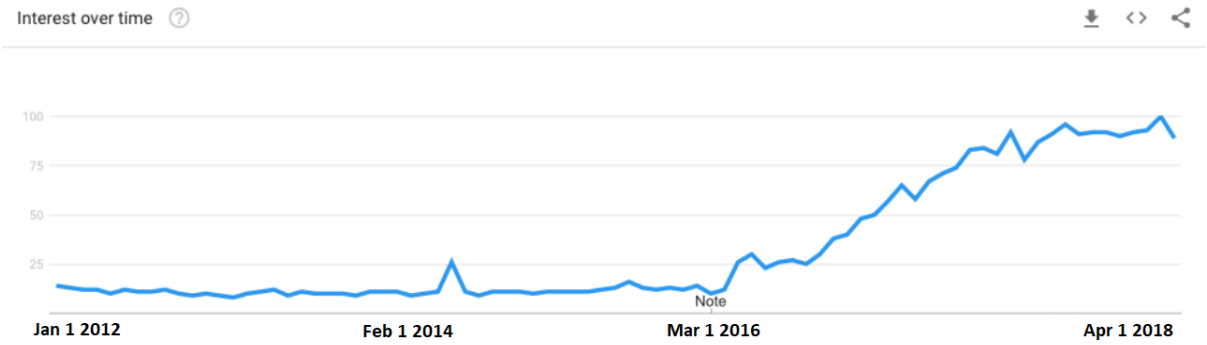

Figure 1 Google Trends Global Interest in the Term 'Chatbot'

organizations. Organizations therefore need to be careful in their design and deployment of these technology artifacts, to ensure that the experience that customers have is both effective and enjoyable. While many features warrant attention, one particularly important aspect to consider is the extent to which autonomous agents (and specifically chatbots) are designed with social interaction, and specifically anthropomorphism, in mind (Wilson et al. 2017).

Though anthropomorphism touches several academic disciplines, it can best be described as the attribution of human-like qualities to non-human entities like machines, animals and other objects (Duffy 2003). This phenomenon is a common occurrence when individuals interact with technology that possess certain elements associated with human-to-human interaction, like eye gaze (Kiesler et al. 1984), facial expressions (Kiesler et al. 2008) and conversational turn-taking (Cassell and Bickmore 2000). How individuals humanize technology has been an important topic of inquiry in both Human Computer Interaction (HCI) and Human Robot Interaction (HRI) literature for decades. In some cases, making technology more human-like has proven to be beneficial, increasing user trust and satisfaction with the interface. However, in other cases, adding human-like social cues has led to negative consequences, such as social anxiety (Sproull et al. 1996) and reduced cooperation (Kiesler et al. 1996). As we articulate in our review of prior literature in later sections, a common feature of much of the prior work in this space is the inconsistency of the relationship between anthropomorphism and desirable user outcomes. This inconsistency speaks to the myriad contextual factors that can shape the relationship. With that in mind, in this work, we seek

to understand the impact of integrating anthropomorphic features into AI-enabled autonomous customer service agents, i.e., chatbots, particularly within a retail environment. Specifically, we seek to empirically evaluate the effects of anthropomorphism on transaction conversion. Further, we explore the impact of anthropomorphism on consumer offer sensitivity, informed by prior work in the HCI literature which has drawn a connection from consumer perceptions of anthropomorphism to customer perceptions of fairness and trust. Formally, we evaluate the following two research questions:

- ***RQ1: How and to what degree does customer transaction probability depend on the anthropomorphism of AI-enabled automated customer service agents (chatbots)?***
- ***RQ2: To what degree does customer offer sensitivity vary with the anthropomorphism of AI-enabled automated customer service agents (chatbots)?***

We examine these questions via a field experiment, conducted in partnership with a dual channel clothing retailer based in the United States. Our retail partner has historically operated a used-clothing buy-back program through a web-based form, and employee conversations with customers over email and Facebook messenger. In the prior process, a customer would describe the clothes, obtain a offer estimate from an employee, provide mailing address info and print a shipping label, before sending the clothes to the retailer for final evaluation and payment. We insert ourselves into this process, automating the customer interactions with a Facebook Messenger chatbot, which is integrated with the retailer's Facebook business page. In implementing the chatbot, we integrate a framework that enables us to randomly assign customers into various treatment conditions, such that customers ultimately converse with a chatbot that bears a randomly assigned set of anthropomorphic features. This randomized design allows us to experimentally evaluate the causal relationship between the degree of a chatbot's anthropomorphism and the customer's probability of completing the buy-back process. Moreover, we simultaneously introduce random variation into the cash offer each customer receives, which further enables us to assess the moderating effect of chatbot anthropomorphism on customers' offer sensitivity.

We arrive at two notable findings. First, we find that incorporating anthropomorphism into autonomous customer service chatbots increases conversion rates. Second, we show that, in the presence of a sufficiently large degree of anthropomorphism (3 treatments), customers become more offer sensitive. This latter finding indicates that, as a chatbot becomes more human-like, consumers begin to scrutinize offers. This might occur because offers made by humans are more likely to be perceived as potentially opportunistic (price gouging) or inconsistent (noisy) by consumers, compared to computer-generated offers.

Our study contributes to a number of different streams of literature. First, we contribute to the literature in Information Systems by exploring the design and efficacy of an increasingly prevalent form of information system, the customer service chatbot. In so doing, we build on an extensive literature in HCI related to anthropomorphism by evaluating these features in a field setting. Second, we contribute to the Marketing literature by considering a variety of practical and theoretical issues in the AI-enabled automation of customer service job roles. Building on the work of Wirtz et al. (2018), we empirically evaluate anthropomorphism, a "critical design attribute" of service robots, demonstrating its value in customer service settings. Third, our work contributes to the burgeoning literature on individual's reactions to algorithmic forecasts and estimates (Kleinberg et al. 2017, Dietvorst et al. 2015, 2018, Tambe et al. 2019), and highlight how anthropomorphism could play a role. Finally, and more broadly, our work contributes to the literature on Intelligence Augmentation, or IA (Jain et al. 2018). In particular, our study demonstrates the potential to augment artificially intelligent agents with human-like social intelligence (Wang et al. 2007a). Whereas the literature on IA to date has primarily focused on the possible applications of technology to augment human decision-making abilities, our work highlights opportunities for the reverse; that incorporating human-like behavior and decision-making into autonomous agents can amplify their performance and efficacy as well.

2. Literature Review

2.1. Anthropomorphism & AI

Scholars of computer science and engineering have dedicated a great deal of attention to the efficient performance of AI-based systems, with an eye toward operational performance. However, when it

comes to the automation of job roles or processes that involve human touch-points, social factors are likely to play a particularly prominent role as well. Fortunately, designing autonomous agents to account for social factors has been a focal subject in the Human-Computer Interaction literature for many decades.

A central component in research on the effective design of autonomous agents has been the role of anthropomorphism. Anthropomorphism is a concept that touches several fields of study: psychology (Heider and Simmel 1944, Malle and Pearce 2001, Barrett and Keil 1996), marketing (Aaker 1997), computer science (Duffy 2003, Kiesler et al. 2008) and religion (Guthrie 1995). Although definitions within these fields vary slightly, anthropomorphism, at broad scope, is the attribution of human-like qualities to non-human entities like machines, animals and other objects (Duffy 2003). This attribution is generally the product of humans seeking to explain the actions and behaviors of non-human objects and beings in a way that they understand (Duffy 2003). Although assigning human-like qualities is a very common occurrence that pervades several disciplines, this phenomenon is viewed by several scientific disciplines like biology and psychology as a nuisance that confounds causal mechanisms and hampers scientific inquiry (Kennedy 2003).

While some disciplines view anthropomorphism as a hindrance, others, like HCI, view anthropomorphism as an inevitability that should be accounted for and acknowledged when designing the interface (Caporael 1986). A popular paradigm used in HCI is known as 'Computers Are Social Actors', or CASA, which suggests that people, when presented with technology that contains features like dialogue and turn taking, identify those pieces of technology as a social actor (Moon 2000, Nass and Lee 2001, Nass et al. 1994). It is this conceptualization of digital agents as social actors, that interface designers can apply theories from social sciences, which govern human to human interaction like politeness (Nass et al. 1994) and reciprocity (Moon 2000), and effectively carry these over to human machine interactions (Nass et al. 1994). As such, designers can strategically utilize social cues like small talk, greetings, and transitions to influence user trust with the interface and elicit specific behaviors like self-disclosure (Cassell and Bickmore 2000) and persuasion (Xu and Lombard 2017).

Although anthropomorphic social cues can help designers create a more effective user interface, these features can also lead to unintended negative consequences. More specifically, Ben Shneiderman, a critic of the use of anthropomorphic social cues in the technology interface (Don et al. 1992), contends that designers do not fundamentally understand the way users will perceive and interpret social cues. This lack of understanding can lead to unintended outcomes, namely undesirable perceptions of anthropomorphism (Duffy 2003). As a result, incorporating even minor social cues in an ad-hoc (and ill considered) manner may lead to user disappointment, when the human-like agent falls short of user expectations (Duffy 2003, Nass and Moon 2000). A delicate balance thus needs to be struck when it comes to the incorporation of social cues in chatbots. Accordingly, it should come as no surprise that so many chatbots on Facebook's messenger platform today are incapable of fulfilling the basic requirements of users.[5]

We seek to evaluate the effects of introducing anthropomorphism in chatbots via the three commonly used social cues: social presence, communicative delay, and humor. We will explore how user (customer) exposure to greater levels of anthropomorphism in a chatbot, i.e., greater numbers of features, influence transaction outcomes in a live customer service interaction, as well as any associated shifts in customer offer sensitivity. We discuss the three anthropomorphic features below, referencing relevant literature for each.[6]

Social Presence: A commonly discussed element in papers related to conversational agents is social presence (Sah and Peng 2015, Verhagen et al. 2014, Araujo 2018). In this technological context, adding social presence means to add "sensitive human contact" (Verhagen et al. 2014).

[5] https://www.fool.com/investing/2017/02/28/facebook-incs-chatbots-hit-a-70-failure-rate.aspx

[6] We opt to implement the intensity of anthropomorphism via introducing combinations of treatments, rather than manipulating the level of one treatment, because this enables us to abstract away from any specific cue, to infer effects from anthropomorphism more broadly. In our robustness checks section, we explore the pattern of effects that emerges when we estimate the effect of different combinations of specific cues. There, we demonstrate a pattern consistent with the idea that each cue has a directionally consistent effect on conversion, indicating that our abstraction away from particular cues to anthropomorphism more broadly is justified.

In interacting with a chatbot users have opportunities to make social presence attributions at the beginning (Araujo 2018, Holtgraves et al. 2007), middle (Sah and Peng 2015, Holtgraves et al. 2007) and end (Araujo 2018) of the conversation.

This social presence can prove to be a double edged sword for practitioners. The more socially-present the interactions are, the more engaging the interface; however, the more human-like the interface the higher expectations that the user has of the machine's communicative prowess (Mone 2016, Nowak and Biocca 2003). With this, designers of chatbots make a very important decision of how their conversational agent is perceived in the beginning of the interaction with a greeting (Araujo 2018, Gefen and Straub 2003). For example, a designer can either greet the user, by introducing itself with a real human name, or level expectations of communicative capability by using a generic machine-like name. By setting the tone with a human name the designer could elicit an anthropomorphic response to the chatbot leading to a more engaging customer experience. Alternatively, in giving the chatbot a human name, the designer could enforce unattainable human expectations on the chatbot, which could lead to frustration later in the experience.

In addition to the greeting, designers can influence anthropomorphic perceptions through the language choices they make in the conversation. For example, using more polite (Fussell et al. 2008), informal (Araujo 2018, Holtgraves et al. 2007) or social (Verhagen et al. 2014) language can help induce anthropomorphic perceptions and also perceptions of social presence. Slight differences in agent language have shown to greatly impact a chatbot's perceived personality (Holtgraves et al. 2007). It is with these linguistic features that designers help to enforce a feeling of social presence and further promote anthropomorphism in their chatbot.

Another method HCI designers use to achieve anthropomorphic attributions towards their machines is through physical social cues (Goetz et al. 2003, Fussell et al. 2008). Unlike embodied conversational agents, chatbots rely solely on text based computer mediated communication to communicate and cannot show physical non-verbal cues like facial expressions or gaze (Kiesler et al. 1984). In computer mediated communication, when these typical face to face social cues

are not present, communicators shift focus to alternative cues available and make social interpretations (Walther and Tidwell 1995, Walther 1992). This theory is known as Social Information Processing (SIP), typically this manifests itself in chronemic cues like timestamps (Walther and Tidwell 1995, Liebman and Gergle 2016). Due to the disembodied nature of chatbots that exist on messaging platforms like Facebook Messenger, Kik or Telegram, designers only have a couple of chronemic social cues at their disposal to enforce feelings of a real socially present human. These would include: read receipts and ellipses during typing messages. Although, these two features are common place when two humans are talking via Facebook messenger, these cues are not required from a chatbot as it neither types nor reads.

Although, these anthropomorphic perceptions could lead to the higher amounts of sociability between the chatbot and the customer, these deviations from a more task oriented style could lead to more difficulty and time for users to complete a self-service task. Additionally, it could also over promise the communicative prowess of the agent on the other end of the conversation. This could be counter-productive as users of self-service technologies do so because they are convenient, quick and a means to circumvent interacting with service individuals (Meuter et al. 2000). As such, there is a potential that these communicative features could lead to one of two outcomes. The first is that, the more anthropomorphic the chatbot becomes the more a customer is willing to engage with the artifact. This prolonged interaction would eventually lead to a resolution of the issue, and save labor costs for the company. Alternatively, these anthropomorphic additions to the chatbot obfuscate task oriented nature of the typical self-service interaction, and could lead to frustration and dissatisfaction as the features add overhead to the experience and also mislead the user about the chatbot's communicative prowess.

Communication Delays: In addition to language communication features, another social cue employed by both researchers and practitioners is delay (Holtgraves and Han 2007, Crozier 2017, Gnewuch et al. 2018). From one perspective, delays could be interpreted as the chatbot not working as expected. However, when implemented correctly, slight delays that are dynamic to the amount of

text can dictate levels of persuasion (Moon 1999) and chatbot personality perceptions (Holtgraves and Han 2007). At face value, this anthropomorphic effect of delays seems somewhat intuitive as humans do not read and respond to messages sent through text based mediums instantaneously.

Although these slight delays may lead to more anthropomorphic perceptions of the chatbot, they may also interrupt the service quality associated with the experience (Taylor 1994, Meuter et al. 2000). Thus delays in sending messages could lead to two different outcomes in a customer service interaction. If the anthropomorphic features of the interface lead to higher levels of trust in the interface, then potentially these slight delays would enhance the user experience and lead to higher levels of satisfaction with the experience. In contrast, delays can be viewed as an element that impedes the service encounter and prevents the customer from accomplishing the self service task.

Humor: In the fields of socio-linguistics and pragmatics, humor has been shown to introduce feelings of common ground between two communicating social actors (Holtgraves 2011, Brown and Levinson 1987). Similar to human to human interactions, humor can be an effective way to personify systems, and create a more engaging interaction (Niculescu et al. 2013, Morkes et al. 1999). Additionally, humor in task oriented communications has been shown to increase individuals satisfaction with the task (Morkes et al. 1999).

Although humor may be beneficial, it does appear that there is some nuance required in implementing humor. For instance in the medical field, humor helps improve reassurance for patients, but only in the correct context (Francis et al. 1999). This also has been shown in human and robot interaction, where robots with a more playful personality gains more compliance from humans in a non-serious task, and more serious robots perform better in serious task (Goetz et al. 2003). Similarly, humor in both business and customer service interactions requires a more nuanced approach (Malone 1980, Dolen et al. 2008). More specifically, Dolen et al. (2008) find that while humor in an electronic service encounter can help in some situations in which the process is to their liking, but when the process is not to their liking additions of humor exacerbates the negative feelings associated with the service experience. With this nuance of humor, in a customer service interaction, it is unclear whether humor will increase the satisfaction for users engaging with the chatbot or whether it will hinder the overall experience.

Humans and Algorithmic Decision Making. Several emerging studies in Human Resources (Tambe et al. 2019), Economics (Kleinberg et al. 2017) and Psychology (Dietvorst et al. 2015, 2018, Logg et al. 2019) have investigated how humans respond to algorithmic outcomes. Dietvorst et al. (2015) find that in general humans are averse to forecasts made by an algorithm, even when they outperform their less accurate human counter-parts. Dietvorst et al. (2018) further this line of inquiry and find that algorithmic aversion can be reduced when individuals have the ability to manipulate and make adjustments to the algorithm. Similarly, Tambe et al. (2019) theorize that employees will be less accepting of algorithmically determined shift decisions than those determined by a supervisor as they could potentially feel less involved in the decision. Interestingly, Tambe et al. (2019), further discusses an anecdote from Uber, describing that individuals negatively respond to surge pricing when they believe it is set by an algorithm.

Contrasting these findings Logg et al. (2019) find that individuals can be appreciative of algorithmic judgements in numeric forecasts and recommendations for dating and music, as opposed to those made by humans. In addition, Logg et al. (2019) find, similar to Dietvorst et al. (2018), that individuals prefer their own judgements over that of an algorithm. As this aforementioned research indicates, how individuals react to algorithmic outcomes is very dependent on context and human involvement.

Behavioral Economics has sought to understand how individuals reason through offers. One classic example is the *Ultimatum Game*, (Gurth et al. 1982). In this game, a proposer makes an offer of money, and the offer receiver is to accept or reject the offer. The rational expectation is that the proposer is to make a small offer, and the recipient should accept the offer, regardless of its fairness, because this is the utility maximizing response, i.e., take what you can get (Gurth et al. 1982). A fairly robust experimental finding, however, is that offers of 20% of the total funds available are rejected 50% of the time (Sanfey et al. 2003), because of perceived injustice or a lack of fairness.

Previous research has found that human players tend to reject unfair offers less when the actor making an offer is perceived as lacking intentionality, e.g., a computer, rather than a human. For

example, Sanfey et al. (2003), Moretti and Pellegrino (2010) report that recipient rejection rates for relatively low offers increase when the offer is made by a human, versus when the offer is made by a computer (notably, a computer that is totally absent of anthropomorphic features). These authors argue that this occurs because human proposers are more likely to induce recipient emotions, such as disgust (Moretti and Pellegrino 2010).

However, other work has documented contradictory evidence. Torta et al. (2013) found that individuals rejected computer generated offers in the Ultimatum Game more frequently than offers made by humans. Torta et al. (2013) theorize that this occurs because human actors have an easier time processing offers from other humans, but face some difficulty deciding how to respond to offers from computers. For example, the willingness to reject an offer may depend on the manifestation or conformity to social norms and etiquette. Thus, whereas a human actor may have no qualms about rejecting an offer from a non-human actor, off hand, social norms might dictate that the human be courteous and considerate when interacting with another human, imposing a sort of social friction on rejection.

More generally, the HCI literature has found that humans respond more socially when computer-based agents are more anthropomorphic (Kiesler et al. 1996, Nass et al. 1994). As one specific example, Kiesler et al. (1996) found that human participants presented with a Prisoner's Dilemma game tended to respond socially to 'humanized' computer actors, in a manner similar to the response they would exhibit with a true human partner. These findings further the notion that a potentially important element leading to offer receivers acceptance or rejection of offers is the level of anthropomorphism of the automated proposer.

As there is ample evidence to support the benefits and detriments of including anthropomorphism in customer service chatbots, we take on this study and look to its data to help us reach a conclusion.

3. Study Context

As described above, we conducted our field experiment in partnership with a dual channel clothing retailer based in the United States, similar to other businesses like Plato's Closet and Clothes

Mentor. This retailer buys and sells women's used clothing, both online and through three brick and mortar locations in Iowa and Minnesota. We replaced the retailer's prior, manual clothing buy-back process with an AI-enabled chatbot. The process we automate was previously managed via web-form and email exchanges, or done in person at a store. We developed the chatbot using Google's Conversational AI Platform, DialogFlow, incorporating Python-based customizations. DialogFlow enables the automated processing and generation of conversational prompts and utterances in exchanges employing natural language. The Python customizations were incorporated to implement required business rules and logic, as well as to manage the conversational flow (e.g., if customer says this, do that). The chatbot was integrated with the retailer's Facebook business page, as part of the retailer's Facebook messenger profile. The retailer's Facebook page has approximately 44,000 followers.

The chatbot is designed to interact with customers who are interested in selling their used clothing to the retailer. The overall conversational interaction model has three major steps. First, the chatbot begins by requesting information on the number and types of clothing that the customer wishes to sell. Then, the chatbot provides an estimated cash offer, indicating the expected value that the retailer would be willing to pay for the clothing described. If the customer accepts the offer, the chatbot then requests additional personal details that are required to complete the transaction, including a mailing address, full legal name, and phone number. Based on this information, a shipping label is generated, which the customer can print and use to send their clothes to the retailer.

4. Methods

4.1. Experiment Design

To causally identify the impact of the aforementioned anthropomorphic features on transaction outcomes, we implement three independently randomized treatments, one associated with each of three anthropomorphic features. When a customer initiates a conversation with the chatbot for the first time, he or she is randomized into receiving zero, one, two or all three of the anthropomorphic

features, in random combinations. We describe the implementation of each treatment, below. Note that by independently randomizing each anthropomorphic feature, we ensure that there is no association between the number of features a customer receives, and which features a customer receives. Our randomization is performed on a between subjects basis. If a single customer revisits our chatbot and initiates additional conversations with our chatbot, we exclude any such subsequent observations from our analysis.

It is worth highlighting that our focus is not on any one of the anthropomorphic treatments, but rather on the number of treatments a subject receives. Our objective in delivering varied numbers of treatments is to causally shift a subject's perception of anthropomorphism in the chatbot interaction. Conceptually, this approach is analogous to the notion of *Combination Therapy* or *Polytherapy* in medicine, which refers to treating a single disease with multiple types of interventions, in concert (e.g., Möttönen et al. 1999). We opt for this approach, rather than attempting to manipulate the intensity of a given anthropomorphic feature by shifting its level, for two reasons. First, it is not altogether clear how dosage manipulations could be achieved with each of the treatments, e.g., it is not altogether clear what would constitute more versus less humor. Second, the perception that one is certainly interfacing with a human actor is unlikely to be achieved through a single manipulation, even in a text-based setting. A chatbot that responds instantaneously, yet also drops a joke into the conversation, may be perceived as having some human traits. However, it is unlikely that simply adding more jokes into the exchange will achieve further improvements. Thus, it is reasonable to assume that anthropomorphism depends a great deal on delivering a sufficient constellation of anthropomorphic features as part of the exchange.[7]

Additionally, for all customers, we introduce random variation into the cash offer. In the original buy-back process, the retailer would calculate an initial cash offer based on a fixed amount of $3.50 per clothing item. We randomly perturbed the offer around the fixed baseline offer for each

[7] We offer later analyses, namely manipulation checks, which indicate that perceived anthropomorphism is increasing in the number of treatments received, providing support for our argued mechanism.

customer, drawing from a random normal distribution with mean 0 and variance 0.5. That is, our offer perturbations were implemented by taking the $3.50 baseline offer previously employed by the retailer, and adding a random value drawn from this normal distribution. Drawing from a normal distribution allowed us to accommodate concerns on the part of the retail partner that cash offers would be 'too extreme' in either direction, creating customer experience issues on the one hand and economic losses for the retailer on the other hand.

Social Presence. To operationalize anthropomorphic social presence, we do so through a combination of a name, linguistic features and social cues related to reading and authoring messages. We thus adopt a methodology similar to that of Araujo (2018). More specifically, in this treatment, we first give the chatbot a randomly drawn human name from the 1990 census, which the chatbot uses to introduce itself at the outset of the conversation. Second, like Araujo (2018), the chatbot employs relatively informal, casual language (as opposed to more formal, professional language). An example of the initial greeting manipulation can be found in the table below.

Table 1 Social Presence Manipulation

Condition	Message
0	*"Hello I am an automated service bot here to assist with shipping previously used maternity clothing for money."*
1	*Hi I'm Teddy here to help you with shipping previously loved maternity clothes for* $

In the human-like condition, users will also see the cues typically associated with messages exchanged between humans. On the Facebook Messenger platform, these cues include both read receipts when a message is sent to the chatbot, as well as the display of a cue indicating that the chatbot is typing a message. An example of the typing feature can be seen in Figure 2 and read receipts in Figure 3.

In conditions where these cues are not present, the user sees simply the white messenger background without the read receipts or typing features.

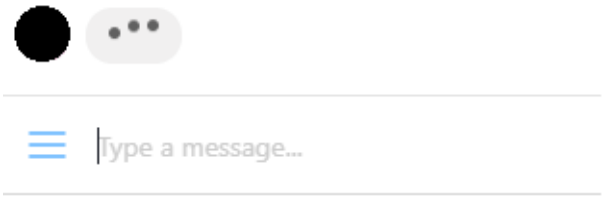

Figure 2 Typing Feature

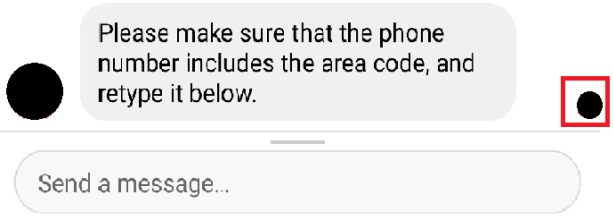

Figure 3 Read Receipt is Shown as Small Profile Image on Right

Communication Delays. Similar to Moon (1999) and Holtgraves and Han (2007), we implement a dynamic delay of 70 words per minute. This is within the range of those that type professionally [8]. In the non-human-like condition, users will experience instant responses.

Humor. To operationalize the humor construct we insert a random joke drawn from an approved list of 4 jokes. These jokes were deemed to be inoffensive, and suitable for any age. The random jokes are added into the dialogue, right before the customer receives the estimate for the clothes they will be selling to the retailer. In conditions that do not have humor present, the customer is asked if they will wait a moment while the chatbot totals up their estimate, and a 5 second long pause ensues. This interaction is depicted in Figure 4. A brief summary of all manipulations can be found in Table 2.

4.2. Empirical Specification, Variables & Data

In our analyses, we are interested in understanding the effect of increasing 'humanization' of the chatbot on i) the probability of conversion, and ii) the moderating effect on the relationship

[8] https://www.livechatinc.com/typing-speed-test/#/

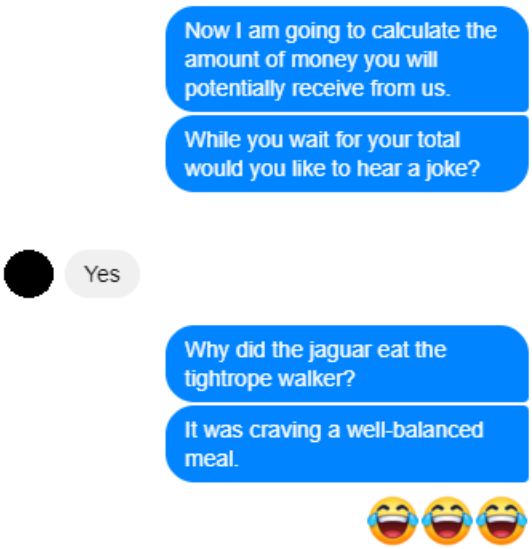

Figure 4 **Joke Example**

Table 2 **Chatbot Features**

Feature	**Description**
Social Presence	*Human Name, Informal Language, Typing Cues*
Delay	*Dynamically typed 70 WPM delay*
Humor	*Randomly selected joke before estimate*

between randomly varied offer amount, and conversion. Accordingly, our primary outcome variable of interest is a binary indicator of conversion. Our independent variables include a series of dummy variables reflecting different levels of the number of anthropomorphic treatments a subject received, *Treatment_Count*, as well as a measure reflecting our offer perturbation, *Cash Offer*, which we mean-center for the sake of simplicity.

We first estimate a series of Linear Probability Models (LPMs), regressing conversion on our treatment count dummies and our offer deviation measure, to understand their direct effects. Subsequently, we interact the dummies and the offer measure, to understand the moderating effects of interest, i.e., how increasing anthropomorphism moderates offer sensitivity. Our final cash offer sensitivity model is reflected below in Equation 1, where subjects are indexed by i.

$$Convert_i = \alpha + \beta_1 \cdot 1\ Treatment_i + \beta_2 \cdot 2\ Treatments_i + \beta_3 \cdot 3\ Treatments_i + \delta \cdot Cash\ Offer_i + \gamma_1 \cdot 1\ Treatment_i \cdot Cash\ Offer_i + \gamma_2 \cdot 2\ Treatments_i \cdot Cash\ Offer_i + \gamma_3 \cdot 3\ Treatments_i \cdot Cash\ Offer_i + \epsilon_i \quad (1)$$

Our experiment includes 323 subjects who initiated a conversation with our chatbot between November 16^{th} and December 31^{st} of 2018. We present the descriptive statistics for our variables in Table 3. As can be seen, approximately 8.36% converted, meaning they completed the buy-back procedure and obtained a shipping label to send their clothes to the retailer. We also observe that the average user received 1.5 anthropomorphism treatments. Figure 5 depicts the distribution of randomized per item offers that were assigned to subjects. As explained earlier, the distribution of offer deviations is normal.

Table 3 Descriptive Statistics

Variable	Mean	St. Dev.	Min	Max
Social Presence	0.56	0.50	0.00	1.00
Delay	0.48	0.51	0.00	1.00
Humor	0.46	0.50	0.00	1.00
Treatment Count	1.50	0.89	0.00	3.00
Cash Offer	-0.02	0.68	-1.82	1.43
Conversion	.0836	.2772	0	1

5. Results

We begin by estimating a linear probability model, incorporating only the main effects of each variable. We then progress to incorporating interactions, to recover any effect of cash offer increases on conversion outcomes under alternative levels of anthropomorphism.

Considering the results in Table 4, in Column 1, the constant term indicates that the baseline rate of conversion in the control condition (no anthropomorphic treatments) is approximately

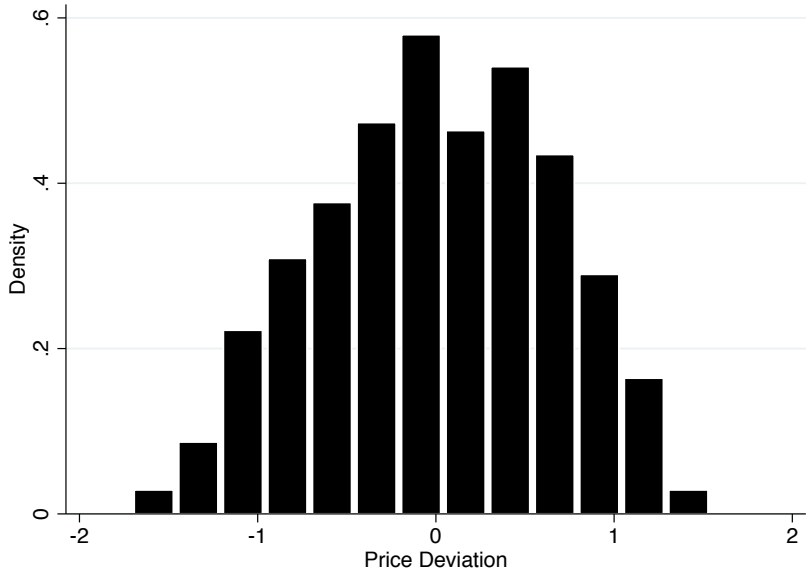

Figure 5 Distribution of Per Item Offer Deviation

2.6%. We observe positive coefficients associated with all other variables in the model. Specifically, we observe that a single anthropomorphic treatment is associated with a 6.7% increase in the probability of conversion ($p < 0.10$), relative to control; a pair of treatments is associated with a 5.0% increase in the probability of conversion (though the result is not statistically significant relative to a null hypothesis of 0); and the receipt of all three treatments in tandem is associated with a 10.8% increase in the probability of conversion ($p < 0.05$). Although the coefficient on cash offer is positive as we expect (given this is a cash offer made to the customer, not a cash offer charged), the coefficient is not statistically significant. That said, the estimate indicates that a \$1.00 increase in the cash offer is associated, on average, with a 2.7% increase in the probability of conversion.

Next, considering the interaction model in Column 2, the main effects associated with the intensity of anthropomorphism remain quite consistent, except that all three estimates are now statistically significant at commonly accepted thresholds (when our cash offer manipulation is 0). Additionally, considering the cash offer interactions, we see that all coefficients are positive and increasing in the number of treatments. Of particular note, we observe that the cash offer

Table 4 Treatment Count Model (LPM)

Variable	DV = Convert	DV = Convert
1 *Treatment*	0.067* (0.036)	0.076** (0.032)
2 *Treatments*	0.050 (0.034)	0.060** (0.030)
3 *Treatments*	0.108** (0.055)	0.109** (0.049)
1 *Treatment* · *Cash Offer*	—	0.052 (0.064)
2 *Treatments* · *Cash Offer*	—	0.086 (0.069)
3 *Treatments* · *Cash Offer*	—	0.211** (0.087)
Cash Offer	0.027 (0.022)	−0.058 (0.057)
Intercept	0.026 (0.023)	0.017 (0.017)
Observations	323	323
R^2	0.016	0.037
F	1.60 $(4, 319)$	3.85*** $(7, 316)$

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

manipulation has a statistically significant interaction with the delivery of three anthropomorphic treatments, relative to the delivery of none ($p < 0.05$). This finding indicates that, in the presence of sufficient anthropomorphism, consumers become significantly more offer sensitive.

6. Robustness

6.1. Estimator Choice & Regression Specification

We begin by considering the robustness of our results to possible concerns of multicollinearity, as well as to our choice of estimator. We report analyses addressing possible concerns of multicollinearity in Appendix A, where we provide evidence that this is not a serious concern in our analysis. Subsequently, in Appendix B, we explore the robustness of our results to our choice of estimator, namely the Linear Probability Model. There, we demonstrate that our results remain stable under alternative estimator choices.

6.2. Replication

We next assessed the replicability of our main finding, that anthropomorphism increases transaction rates, conducting a second, simpler experiment in the same field setting. With this replication,

we sought to again address possible concerns that our results somehow derive from aggregating across multiple treatments. With that concern in mind, we sought to evaluate the treatment effect of just a single anthropomorphism treatment, relative to a control condition. This replication thus allowed us to assess whether, given sufficient power, a single anthropomorphism intervention would yield statistically significant estimates of increased conversion. We focused on the social presence treatment in this replication, because it is the intervention that aligns most intuitively with anthropomorphism (Araujo 2018).

The replication was conducted in the same field context. The only distinction in this case is that our experiment was limited to just two conditions: the control condition, in which no anthropomorphism treatment was delivered, and the social presence condition. As before, we assessed the relationship between the treatment and the probability of successful conversion. This experiment was carried out over a 1-month period, from late June to late July of 2019. Recruitment for the replication study was conducted in the same manner, employing Facebook messenger advertisements.

This experiment involved 546 subjects, who were approximately balanced in their assignment to treatment and control; the mean value of our treatment indicator, *Social Presence*, was 0.46. As before, we regressed a binary indicator of transaction conversion onto a treatment dummy, employing a Linear Probability Model. As before, we observe a positive, statistically significant effect on conversion rates with this single, individual treatment. Specifically, social presence features led to an approximate 5% increase in the transaction conversion rate ($p = 0.046$). Thus, we successfully replicate the main result. Moreover, we conclude that, given sufficient statistical power, we can detect that a single anthropomorphism treatment can translate to tangible benefits for transaction conversion.

6.3. Manipulation & Randomization Checks

We performed a manipulation check with 19 volunteers, to ensure that the various treatments were properly experienced by users, and that they had the expected effects on both anthropomorphism

level and perceptions of manipulations. To determine if end users indeed experienced the delay and humor treatments, we asked participants to rate their agreement with certain statements, on a scale 1 (Strongly Agree) to 6 (Strongly Disagree). For the humor treatment, the statement was: *The customer service agent was humorous.* For the delay treatment, the statement was: *The customer service agent took a long time to respond.* To analyze the survey responses we used the Mann-Whitney U test (Mann and Whitney 1947). We find that the there is a significant difference between responses that were in the humor and non-humor conditions and the delay and non-delay condition. This is significant at the p $\leq$.01 level.

Table 5 Results Mann-Whitney Rank Sum Test for Manipulation's Perceptions of Delay & Humor

Condition Comparison	$p-value$	**z**
Humor vs Non-Humor	0.0045	2.842
Delay vs Non-Delay	0.0006	3.417

In addition to running the tests for both the humor and delay manipulations, we also tested whether the delivery of these features in tandem with linguistic features led to a higher perception of anthropomorphism. To test this, we used a semantic differential scale, including survey items first introduced by Powers and Kiesler (2006). These survey items are also a component of the Godspeed Questionnaire (Bartneck et al. 2008), a widely used survey in the HCI and Human Robot Interaction literature to measure anthropomorphism (Weiss and Bartneck 2015). The semantic scale ranges from 1 to 6, for five binary word associations: (Fake, Natural), (Machine-like, Human-like), (Unconscious, Conscious), (Artificial, Life-like), (Moving Rigidly, Moving Elegantly). The lower the score, the less anthropomorphic the artifact is perceived to be. Note that we adapted the final word-pair to our textual context, replacing it with (Messages Rigidly, Moving Elegantly). The original scale was developed for use with physical artifacts, i.e., robots, to capture perceptions of movement in physical space; however, because our artifact only exists on the Facebook messenger platform, slight modification was necessary. We averaged the values across the 5 semantic differential scale items to arrive at our final measure.

To determine if the addition of these features leads to higher perceptions of anthropomorphism, we sum the treatment dummies associated with the features: Social Presence, Communication Delays, and Humor, such that we construct a measure capturing the number of treatments a subject receives (which we expect to associate with increasing levels of perceived anthropomorphism). We then perform an Ordinary Least Squares regression of the mean anthropomorphism differential scale response against the count of treatments received. Doing so, we find a statistically significant, positive association ($\beta = 0.619;\ p < .10$). This manipulation check parallels our main analyses, described earlier, in which we explore the relationship between the number of treatments a subject receives, and their conversion response. Conceptually, our approach is analogous to the notion of *Combination Therapy* or *Polytherapy* in medicine, which refers to efforts to tackle a single disease with multiple treatments, in tandem (e.g., Möttönen et al. 1999). Measures similar to that we employ here have been advanced in the medical literature, i.e., based on a summation over treatment interventions received by a patient or subject (Frei et al. 1998). Thus, rather than attempt to manipulate the intensity of anthropomorphism by shifting the levels of any given treatment (it is not altogether clear what would constitute more versus less humor, or greater versus less social presence), we opt for the delivery of more versus fewer treatment options, in combination, to achieve our manipulations.

In addition to these manipulation checks, we also conducted a number of randomization checks, to assess the efficacy of our randomization procedure. Because we randomize in real-time, as subjects arrive, and only have a small set of information describing our subjects available from Facebook, we are limited in the types of randomization checks we are able to perform. As such, one check we can perform is to assess the significance of the association between the number of treatments a subject was assigned and the day on which they entered our sample. To assess this, we perform a Multinomial Logistic Regression of the number of treatments assigned on a vector of day of week indicators. We report the results of this regression in Table 6, where all coefficients are statistically insignificant. A similar analysis performed as a ordinal logistic regression also yields null results. This provides some assurance that our randomization procedure was effective.

Table 6 Randomization Check (MLOGIT; DV=Treatment Count)

Variable	Treatments = 1	Treatments = 2	Treatments = 3
Tuesday	0.872 (0.696)	0.280 (0.722)	0.118 (0.859)
Wednesday	1.034 (0.689)	1.069 (0.683)	0.929 (0.774)
Thursday	0.178 (0.599)	0.118 (0.596)	-0.352 (0.750)
Friday	0.588 (0.661)	0.057 (0.687)	0.300 (0.778)
Saturday	0.523 (0.630)	0.463 (0.627)	0.405 (0.728)
Sunday	0.187 (0.620)	0.554 (0.598)	-0.442 (0.794)
Constant	0.575 (0.417)	0.636 (0.413)	-0.118 (0.487)
Observations		324	
Pseudo R^2		0.014	
Wald Chi^2		10.94 (18)	

Note: The baseline outcome is 0 Treatments; Robust SEs.

Beyond this assessment of inter-temporal randomization, we also assessed randomization efficacy in two other ways. Specifically, we assessed possible systematic associations between the per-unit cash offer and the treatments a subject was assigned, as well as possible systematic associations between the per-unit cash offer and the number of clothes a subject wished to sell. Each evaluation was conducted via a series of pairwise t-tests, testing for significant differences in pairwise group means. This was done both in terms of treatment count assignments, as well as specific treatment assignments. In all cases, we observe statistically insignificant differences across groups. These results are presented in Appendix C.

7. Mechanism Exploration

Although we have demonstrated a robust, positive, causal relationship between anthropomorphism features and transaction conversion, it is important to also assess the boundary conditions for our findings, as well as to assess the extent to which anthropomorphism is the primary mechanism behind this relationship. Accordingly, we undertook a variety of secondary analyses and controlled experiments. We first sought to better understand the extent of perceived anthropomorphism associated with our most anthropomorphic chatbot, and how it compared with an obvious benchmark,

namely a true human agent. This exercise is important, because it speaks to the potential for further gains, above and beyond the anthropomorphism levels we implemented in this study.

To assess this question, we recruited 54 turkers from Amazon Mechanical Turk and assigned them to either interface with i) our most anthropomorphic chatbot, or ii) a human agent, drawn at random from a pool of four graduate research assistants.[9] These human customer service agents were given a high-level verbal instruction about the information they needed to supply and collect from visitors to complete the buy-back process, including examples of past chatbot interactions.

Each research assistant received a brief training session with one of the authors, and each was observed in a customer service interaction before the experiment was begun to ensure proper understanding of the script. Subsequent to interacting with a customer service agent (either the chatbot or a human), the turkers were asked to respond to a pair of survey items, rating their perceptions of the respective agent's anthropomorphism. To gauge anthropomorphism, we utilized a semantic differential scale, including survey items first introduced by Powers and Kiesler (2006), which ask the subject to rate their interaction on a 1 to 6 scale for five binary word associations: (Fake, Natural), (Machinelike, Humanlike), (Unconscious, Conscious), (Artificial, Lifelike), (Messages Rigidly, Messages Elegantly).

The results of this comparison are presented below in Figure 6, which depicts group means and 95% confidence intervals. A Mann-Whitney U test indicates that a randomly drawn human agent was perceived to be more anthropomorphic than the fully anthropomorphic chatbot, to a statistically significant degree ($p < 0.05$). The difference on a 6-point scale is 2.97 vs. 3.93, this finding does suggest that there is room to further increase perceived anthropomorphism of our chatbot, and perhaps garner greater benefits for transaction outcomes.

[9] The use of multiple human agents is particularly important for this analysis, if we wish our results to be plausibly generalizable. If we were to compare our chatbot against a single human agent, it would be quite difficult to draw conclusions about how the bot might compare to human agents, broadly, versus the particular human agent participating in the study.

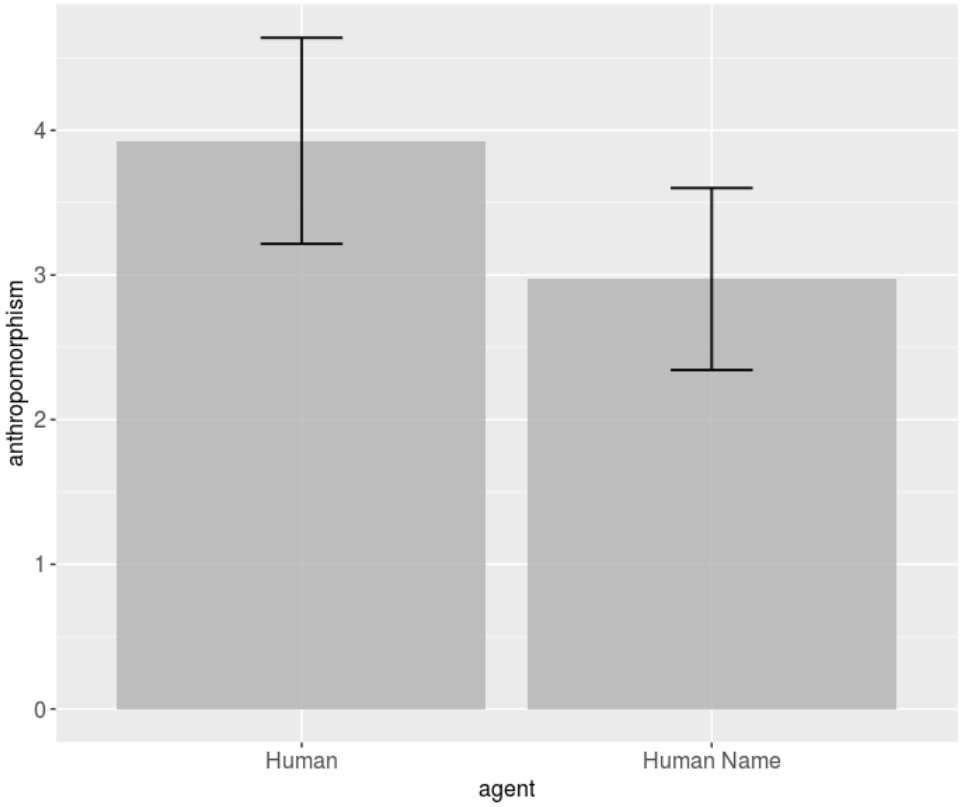

Figure 6 Perceived Anthropomorphism - (L) True Human vs. (R) Anthropomorphic Chatbot

Next, we sought to understand the extent to which our results might derive from our anthropomorphic treatments causing subjects to believe they were truly interfacing with a human agent, versus whether subjects were aware the agent was autonomous and were merely personifying its behavior. Understanding this aspect is important for two reasons. First, there has recently been a push from government regulators to require the disclosure of agents' autonomous nature at the outset of any customer interactions.[10] Accordingly, from a practical perspective, if our results are somehow dependent on the absence of formal disclosure, this would be undesirable, as the value of these findings would be undercut by ongoing regulatory changes in the market. Second, recent work involving voice-based chatbots has reported that a failure to disclose a bot's autonomous nature at the outset of interactions can have detrimental effects on transaction outcomes, if a customer initially believes the agent to be a human, and discovers its autonomous nature only later (Luo et al. 2019).

[10] https://www.natlawreview.com/article/get-all-your-bots-row-2018-california-bot-disclosure-law-comes-online-soon

Our analysis was conducted in a manner similar to the above anthropomorphism bench-marking exercise. Specifically, we recruited 52 turkers to interface with one of two chatbots: i) our fully anthropomorphic chatbot (which lacks explicit disclosure that it is autonomous) and ii) our fully anthropomorphic chatbot, incorporating disclosure. Up-front disclosure was achieved in the latter case by removing the human name and replacing it with the title 'Customer Service Chatbot'. Again, subsequent to these turkers' interactions with their assigned agent, we asked them to respond to survey items. Because we lack objective transaction outcomes in this context, we instead relied upon a proxy response, namely an indication of likeability. For this purpose, we employed adaptations of the survey questions from Mathur and Reichling (2016), obtaining responses to the following prompt: "rate how enjoyable/unpleasant it was interacting with your customer service agent," responding using a sliding scale from -100 to 100. The results are presented below in Figure 7, which again depicts group means and 95% confidence intervals.

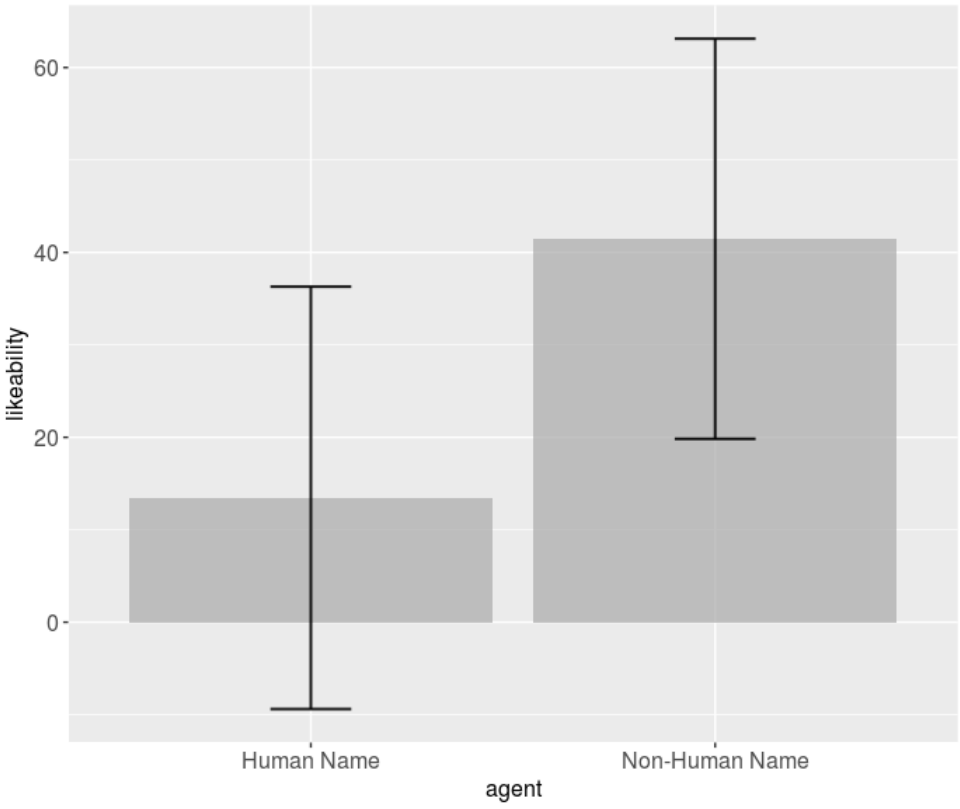

Figure 7 Perceived Likeability - (L) Undisclosed Chatbot vs. (R) Disclosed Chatbot

Interestingly, in this case, we find that, counter to expectation, the fully anthropomorphic chatbot without disclosure was perceived to be significantly *less* likeable than the same chatbot incorporating disclosure ($p < 0.10$). Importantly, this finding indicates that the increases in transaction rates are not dependent upon a lack of disclosure that the agent is autonomous. To the contrary, explicit disclosure appears to *improve* customer perceptions. It is plausible that this occurs because, in our context, users can very quickly deduce that the agent is not human, based on its conversational behavior (even without disclosure). Thus, when the chatbot initially presents a human name, this may create an expectation of human interaction, only to be let down shortly thereafter when the customer perceives that responses are automated. What is more, such rapid realization of the chatbot's autonomous nature may lead customers to perceive some attempt at deception. Under this logic, our findings are in fact consistent with those recently reported by Luo et al. (2019), who found that individuals reacted negatively to delayed disclosure of a chatbot's autonomous nature, versus earlier disclosure.

Having evaluated the anthropomorphism of our chatbots relative to human agents, and having considered whether our results are somehow dependent upon a lack of disclosure, we next turned our attention to an exploration of the underlying mechanisms by which anthropomorphism may benefit transaction outcomes. Our earlier offer elasticity result speaks to this somewhat, in that it suggests that subjects think differently when engaging with an anthropomorphic chatbot. However, we wished to identify concrete evidence of how this differential mindset may benefit transaction outcomes.

One particularly plausible mechanism pertains to humans' trust and willingness to engage in information sharing with autonomous agents. Prior work has observed that a socializing technology can lead to increased persuasion of users (Holzwarth et al. 2006, Wang et al. 2007b) and can lead to more intimate self-disclosure (Moon 2000). In a customer service interaction, social cues may thus lead to greater comfort with the automated customer service agent, on the human customer's part, which then leads to increased levels of information sharing (Sproull et al. 1996). It is therefore

possible that the positive relationship between anthropomorphism and transaction conversion is driven, at least in part, by customers' increased willingness to share sensitive data with the customer service agent that is necessary to complete the transaction.

To explore this possibility, we revisited our original experimental results, considering the treatments' relationship with different information disclosure milestones within the clothing buy-back process. After the offer is seen by a subject, the chatbot proceeds to ask a series of questions to collect contact information that is necessary to complete the transaction. Some of that information is innocuous (i.e., the required dimensions for a shipping box), whereas other information is relatively sensitive (i.e., mailing address, legal name, telephone number). In Table 7, we present the results of repeating our main regression using these different milestones as alternative dependent variables.

As we can see from the results, the anthropomorphic treatments begin to have a statistically significant effect as the customer moves further into the process, as the information becomes more sensitive. Although exploratory in nature, these initial results suggest a partial explanation for the effects we see. Certainly, they point to a potentially fruitful area for further inquiry and policy debate around the incorporation of features aimed to achieve anthropomorphism in autonomous, customer-facing agents.

Table 7 Information Disclosure Milestones (LPM)

Variable	DV = Box Size	DV = Mailing Address	DV = Legal Name	DV = Phone Number
1 *Treatment*	0.042 (0.0546)	0.063 (0.043)	0.078** (0.036)	0.078** (0.036)
2 *Treatments*	0.061 (0.0558)	0.056 (0.043)	0.071** (0.036)	0.071** (0.036)
3 *Treatments*	0.066 (0.0712)	0.113* (0.064)	0.136** (0.060)	0.136** (0.060)
Intercept	0.093 (0.045)	0.047 (0.032)	0.023 (0.0231)	0.023 (0.0231)
Observations	323	323	323	323
R^2	0.004	0.009	0.015	0.015
F	0.46 (3, 319)	1.30 (3, 319)	2.88** (3, 319)	2.88** (3, 319)

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

8. Discussion & Conclusion

Our study offers a novel glimpse into how chatbot anthropomorphism, in a real-world customer service setting, influences business outcomes. We explore prior design theory from HCI, which speaks to the consequences of incorporating anthropomorphic features into an autonomous agent, and the implication for various social outcomes, e.g trust. Although there is reason to believe that trust will lead to customer satisfaction, thereby translating to economic benefits for the firm, it is important to recognize that customer trust and satisfaction with a service provider are only two mediating factors that determine transaction outcomes. For example, although a customer may be more trusting of a 'human-like' autonomous agent, they may simultaneously perceive operational inefficiency, and then opt to transact with an alternative provider. Nonetheless, our results are consistent with the notion that anthropomorphic features have a direct, beneficial relationship with transaction outcomes. Our findings are also consistent with prior studies of anthropomorphism's impact upon trust.

Interestingly, we also find that while anthropomorphism influences transaction conversion positively, it also impacts a customer's offer sensitivity. While our context is somewhat unique to retailers, our findings do give reason to believe that high levels of anthropomorphism is not to be incorporated in all customer service chatbots, and its benefits may be dependent on contextual factors. We also find that anthropomorphism, in our context, plays the most important role in sensitive information disclosure. More specifically, we analysed how anthropomorphism influenced conversion of intermediate variables within the buyback process, and found that it plays a bigger role as customers input more personal information. Though preliminary, this highlights that in certain contexts in which firms require information from their customer, high levels of anthropomorphism could be advantageous. In further experiments discussed in the Appendix on Mechanical Turk, we also find that the individual treatment drives likeability of the agent, and this in turn could be driving much of these conversion outcomes.

Another notable finding comes from our follow-up studies involving crowd-workers. We sought to evaluate whether the practice of disclosing the chatbot's autonomous nature would influence

user perceptions of likeability (our proxy for customer satisfaction). Ultimately, we found that disclosure (i.e., a chatbot that uses a name like 'Customer Service Chatbot') was *more* likeable than the undisclosed chatbot (employing a human name). As we noted earlier, we believe this occurs because customers quickly come to realize that they are not interacting with a human, even in the absence of explicit disclosure. Whereas disclosure makes this clear immediately, a failure to disclose may thus translate to delayed (and unplanned) disclosure, which customers could interpret as an attempt at deception, or falling short of their expectations (Oliver 1977). This finding once again points to the importance of context, and customer expectations. If customers are operating in an environment where they anticipate engaging with automated customer service agents, their expectations for the exchange may be quite different than alternative settings in which a human agent is expected. Recent research has observed that many consumers have grown more comfortable with the notion of algorithms in their daily lives, going so far as to exhibit 'algorithm appreciation' (Logg et al. 2019). This aspect is important for firms considering the design and implementation of autonomous customer service agents.

Additionally, chatbots represent a means by which firms can ensure consistent performance in their human facing customer service roles. In many customer service jobs, individuals are expected to perform routinized tasks with nearly mechanistic efficiency and perfection. This is difficult because individual workers behave differently from each other, as well as the same individual varies their behavior throughout the day. This standardization of service delivery is both a chief concern among most retailers today,[11] as well as a key reason many firms are considering implementing autonomous customer service agents.[12] As such, a potentially effective compromise, that simultaneously leverages the social intelligence of humans, in tandem with the standardized delivery enabled by autonomous agents, is to imbue chatbots with social intelligence (Wang et al. 2007a). Although

[11] eMarketer - Leading Challenges Facing Retailers: https://www.emarketer.com/chart/229895/leading-business-challenges-facing-in-store-retail-according-us-retailers-may-2019-of-respondents

[12] Drift - 2018 State of Chatbots Report: https://www.drift.com/wp-content/uploads/2018/01/2018-state-of-chatbots-report.pdf

current conversational technologies are unlikely to replace the best human customer service agents in the short term, it is plausible that socially intelligent chatbots could lead to improvements in the customer experience if employees exhibit issues with consistency of service delivery and service experience. This observation resonates with the findings of Luo et al. (2019) that autonomous agents may perform better than inexperienced workers in a sales context.

Our research also points to possible opportunities for intelligence augmentation (Jain et al. 2018). First, our work demonstrates that augmenting AI-enabled autonomous agents with human-like social intelligence can increase their performance in customer service settings (Wang et al. 2007a). What is more, our research design suggests a procedure by which firms might leverage autonomous chatbot implementations to experimentally evaluate the most effective patterns of customer interaction, with an eye toward informing the training of human customer service agents. For instance, our experimental results demonstrated that some degree of humor (discussed in Appendix D) can lead to increased conversion rates in this clothing buy-back process. Accordingly, companies might leverage this approach to deduce what works in their context, with their customer base.

Also important to note, our findings are particular to this retailing cash offer scenario. Whether these results will translate to a purchasing, frequently asked questions or healthcare implementation of a chatbot, requires more research. Where anthropomorphism could keep users more engaged in some scenarios, it could also lead to further user frustrations. For example, in a medical diagnosis context, incorporating these anthropomorphic features could inadvertently trigger patients to try and portray themselves in a more positive light (Sproull et al. 1996), and give less accurate depictions of their symptoms. Although anthropomorphism is one aspect that AI designers can use to impact user experience, we also believe that there is fruitful future work evaluating many other aspects like chatbot personality and user based customization.

In summary, our work provides a unique first step toward understanding social and behavioral factors that are worth considering in firms' deployment of autonomous, AI-enabled systems in

customer-facing roles. We show that while overall transaction conversion positively increases with anthropomorphism, anthropomorphizing agents can come with several unintended consequences, like greater offer sensitivity. Given that the deployment of chatbots is already quite common, it behooves researchers to further our understanding of best practices for design and implementation of these systems, and what collateral consequences such design decisions may have on the human-agent interaction. It is our hope that this study will spur a new stream of literature in that direction.

Acknowledgments

References

Aaker JL (1997) Dimensions of brand personality. *Journal of Marketing Research* 34(3):347356, URL http://dx.doi.org/10.1177/002224379703400304.

Ai C, Norton EC (2003) Interaction terms in logit and probit models. *Economics letters* 80(1):123–129.

Araujo T (2018) Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior* 85:183189, URL http://dx.doi.org/10.1016/j.chb.2018.03.051.

Barrett JL, Keil FC (1996) Conceptualizing a nonnatural entity: Anthropomorphism in god concepts. *Cognitive Psychology* 31(3):219247, URL http://dx.doi.org/10.1006/cogp.1996.0017.

Bartneck C, Kuli D, Croft E, Zoghbi S (2008) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics* 1(1):7181, URL http://dx.doi.org/10.1007/s12369-008-0001-3.

Brown P, Levinson SC (1987) *Politeness: some universals in language use* (Cambridge Univ. Pr.).

Caporael L (1986) Anthropomorphism and mechanomorphism: Two faces of the human machine. *Computers in Human Behavior* 2(3):215234, URL http://dx.doi.org/10.1016/0747-5632(86)90004-x.

Cassell J, Bickmore T (2000) External manifestations of trustworthiness in the interface. *Communications of the ACM* 43(12):5056, URL http://dx.doi.org/10.1145/355112.355123.

Crozier (2017) Lufthansa delays chatbot's responses to make it more 'human'. URL https://www.itnews.com.au/news/lufthansa-delays-chatbots-responses-to-make-it-more-human-462643.

Dale R (2016) The return of the chatbots. *Natural Language Engineering* 22(05):811817, URL http://dx.doi.org/10.1017/s1351324916000243.

Deke J, et al. (2014) Using the linear probability model to estimate impacts on binary outcomes in randomized controlled trials. Technical report, Department of Health and Human Services, URL https://www.hhs.gov/ash/oah/sites/default/files/ash/oah/oah-initiatives/assets/lpm-tabrief.pdf.

Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1):114126, URL http://dx.doi.org/10.1037/xge0000033.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):11551170, URL http://dx.doi.org/10.1287/mnsc.2016.2643.

Dolen WMV, Ruyter KD, Streukens S (2008) The effect of humor in electronic service encounters. *Journal of Economic Psychology* 29(2):160179, URL http://dx.doi.org/10.1016/j.joep.2007.05.001.

Don A, Brennan S, Laurel B, Shneiderman B (1992) Anthropomorphism: from eliza to terminator 2. *Proceedings of the SIGCHI conference on Human factors in computing systems*, 67–70 (ACM).

Duffy BR (2003) Anthropomorphism and the social robot. *Robotics and Autonomous Systems* 42(3-4):177190, URL http://dx.doi.org/10.1016/s0921-8890(02)00374-3.

Francis L, Monahan K, Berger C (1999) A laughing matter? the uses of humor in medical interactions. *Motivation and Emotion* 23:155174.

Frei E, Elias A, Wheeler C, Richardson P, Hryniuk W (1998) The relationship between high-dose treatment and combination chemotherapy: the concept of summation dose intensity. *Clinical cancer research* 4(9):2027–2037.

Fussell SR, Kiesler S, Setlock LD, Yew V (2008) How people anthropomorphize robots. *Proceedings of the 3rd international conference on Human robot interaction - HRI 08* URL http://dx.doi.org/10.1145/1349822.1349842.

Gefen D, Straub DW (2003) Managing user trust in b2c e-services. *e-Service Journal* 2(2):724, URL http://dx.doi.org/10.1353/esj.2003.0011.

Gnewuch U, Morana S, Adam M, Maedche A (2018) Faster is not always better: Understanding the effect of dynamic response delays in human-chatbot interaction. *Twenty-Sixth European Conference on Information Systems* .

Goetz J, Kiesler S, Powers A (2003) Matching robot appearance and behavior to tasks to improve human-robot cooperation. *The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003.* URL http://dx.doi.org/10.1109/roman.2003.1251796.

Gurth W, Schmittberger R, Schwarze B (1982) An experimental analysis of ultimatum bargaining. *An experimental analysis of ultimatum bargaining* 3(4):367388.

Guthrie SE (1995) *Faces in the clouds a new theory of religion* (Oxford University Press).

Heider F, Simmel M (1944) An experimental study of apparent behavior. *The American Journal of Psychology* 57(2):243, URL http://dx.doi.org/10.2307/1416950.

Holtgraves T (2011) *Language as social action social psychology and language use* (Routledge).

Holtgraves T, Han TL (2007) A procedure for studying online conversational processing using a chat bot. *Behavior Research Methods* 39(1):156163, URL http://dx.doi.org/10.3758/bf03192855.

Holtgraves T, Ross S, Weywadt C, Han T (2007) Perceiving artificial social agents. *Computers in Human Behavior* 23(5):21632174, URL http://dx.doi.org/10.1016/j.chb.2006.02.017.

Holzwarth M, Janiszewski C, Neumann MM (2006) The influence of avatars on online consumer shopping behavior. *Journal of Marketing* 70(4):1936, URL http://dx.doi.org/10.1509/jmkg.70.4.019.

Horrace WC, Oaxaca RL (2006) Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters* 90(3):321327, URL http://dx.doi.org/10.1016/j.econlet.2005.08.024.

Jain H, Padmanabhan B, Pavlou PA, Santanam RT (2018) Call for papersspecial issue of information systems researchhumans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research* 29(1):250–251.

Kennedy JS (2003) *The new anthropomorphism* (Lightning Source UK).

Kiesler S, Powers A, Fussell SR, Torrey C (2008) Anthropomorphic interactions with a robot and robotlike agent. *Social Cognition* 26(2):169181, URL http://dx.doi.org/10.1521/soco.2008.26.2.169.

Kiesler S, Siegel J, Mcguire TW (1984) Social psychological aspects of computer-mediated communication. *American Psychologist* 39(10):11231134, URL http://dx.doi.org/10.1037/0003-066x.39.10.1123.

Kiesler S, Sproull L, Waters K (1996) A prisoners dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology* 70(1):4765, URL http://dx.doi.org/10.1037//0022-3514.70.1.47.

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S (2017) Human decisions and machine predictions*. *The Quarterly Journal of Economics* URL http://dx.doi.org/10.1093/qje/qjx032.

Lance CE (1988) Residual centering, exploratory and confirmatory moderator analysis, and decomposition of effects in path models containing interactions. *Applied Psychological Measurement* 12(2):163175, URL http://dx.doi.org/10.1177/014662168801200205.

Liebman N, Gergle D (2016) It-s (not) simply a matter of time: The relationship between cmc cues and interpersonal affinity. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing - CSCW 16* URL http://dx.doi.org/10.1145/2818048.2819945.

Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151:90103, URL http://dx.doi.org/10.1016/j.obhdp.2018.12.005.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* URL http://dx.doi.org/10.1287/mksc.2019.1192.

Malle BF, Pearce GE (2001) Attention to behavioral events during interaction: Two actor-observer gaps and three attempts to close them. *Journal of Personality and Social Psychology* 81(2):278294, URL http://dx.doi.org/10.1037//0022-3514.81.2.278.

Malone PB (1980) Humor: A double-edged tool for todays managers? *The Academy of Management Review* 5(3):357, URL http://dx.doi.org/10.2307/257110.

Mann HB, Whitney DR (1947) On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18(1):5060, URL http://dx.doi.org/10.1214/aoms/1177730491.

Mathur MB, Reichling DB (2016) Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition* 146:22–32.

Meuter ML, Ostrom AL, Roundtree RI, Bitner MJ (2000) Self-service technologies: Understanding customer satisfaction with technology-based service encounters. *Journal of Marketing* 64(3):5064, URL http://dx.doi.org/10.1509/jmkg.64.3.50.18024.

Mone G (2016) The edge of the uncanny. *Communications of the ACM* 59(9):1719, URL http://dx.doi.org/10.1145/2967977.

Moon Y (1999) The effects of physical distance and response latency on persuasion in computer-mediated communication and humancomputer communication. *Journal of Experimental Psychology: Applied* 5(4):379392, URL http://dx.doi.org/10.1037/1076-898x.5.4.379.

Moon Y (2000) Intimate exchanges: Using computers to elicit selfdisclosure from consumers. *Journal of Consumer Research* 26(4):323339, URL http://dx.doi.org/10.1086/209566.

Moretti L, Pellegrino GD (2010) Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion* 10(2):169180, URL http://dx.doi.org/10.1037/a0017826.

Morkes J, Kernal HK, Nass C (1999) Effects of humor in task-oriented human-computer interaction and computer-mediated communication: A direct test of srct theory. *HumanComputer Interaction* 14(4):395435, URL http://dx.doi.org/10.1207/s15327051hci1404_2.

Möttönen T, Hannonen P, Leirisalo-Repo M, Nissilä M, Kautiainen H, Korpela M, Laasonen L, Julkunen H, Luukkainen R, Vuori K, et al. (1999) Comparison of combination therapy with single-drug therapy in early rheumatoid arthritis: a randomised trial. *The Lancet* 353(9164):1568–1573.

Nass C, Lee KM (2001) Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied* 7(3):171181, URL http://dx.doi.org/10.1037//1076-898x.7.3.171.

Nass C, Moon Y (2000) Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 56(1):81103, URL http://dx.doi.org/10.1111/0022-4537.00153.

Nass C, Steuer J, Tauber ER (1994) Computers are social actors. *Conference companion on Human factors in computing systems - CHI 94* URL http://dx.doi.org/10.1145/259963.260288.

Niculescu A, Dijk BV, Nijholt A, Li H, See SL (2013) Making social robots more attractive: The effects of voice pitch, humor and empathy. *International Journal of Social Robotics* 5(2):171191, URL http://dx.doi.org/10.1007/s12369-012-0171-x.

Nowak KL, Biocca F (2003) The effect of the agency and anthropomorphism on users sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators and Virtual Environments* 12(5):481494, URL http://dx.doi.org/10.1162/105474603322761289.

Oliver RL (1977) Effect of expectation and disconfirmation on postexposure product evaluations: An alternative interpretation. *Journal of Applied Psychology* 62(4):480486, URL http://dx.doi.org/10.1037/0021-9010.62.4.480.

Powers A, Kiesler S (2006) The advisor robot. *Proceeding of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction - HRI 06* URL http://dx.doi.org/10.1145/1121241.1121280.

Sah YJ, Peng W (2015) Effects of visual and linguistic anthropomorphic cues on social perception, self-awareness, and information disclosure in a health website. *Computers in Human Behavior* 45:392401, URL http://dx.doi.org/10.1016/j.chb.2014.12.055.

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD (2003) The neural basis of economic decision-making in the ultimatum game. *Science* 300(5626):1755–1758.

Sproull L, Subramani M, Kiesler S, Walker J, Waters K (1996) When the interface is a face. *Human-Computer Interaction* 11(2):97124, URL http://dx.doi.org/10.1207/s15327051hci1102_1.

Tambe P, Cappelli P, Yakubovich V (2019) Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61(4):1542, URL http://dx.doi.org/10.1177/0008125619867910.

Taylor S (1994) Waiting for service: The relationship between delays and evaluations of service. *Journal of Marketing* 58(2):56, URL http://dx.doi.org/10.2307/1252269.

Torta E, Dijk EV, Ruijten PAM, Cuijpers RH (2013) The ultimatum game as measurement tool for anthropomorphism in humanrobot interaction. *Social Robotics Lecture Notes in Computer Science* 209217, URL http://dx.doi.org/10.1007/978-3-319-02675-6_21.

Verhagen T, Nes JV, Feldberg F, Dolen WV (2014) Virtual customer service agents: Using social presence and personalization to shape online service encounters. *Journal of Computer-Mediated Communication* 19(3):529545, URL http://dx.doi.org/10.1111/jcc4.12066.

Walther JB (1992) Interpersonal effects in computer-mediated interaction. *Communication Research* 19(1):5290, URL http://dx.doi.org/10.1177/009365092019001003.

Walther JB, Tidwell LC (1995) Nonverbal cues in computermediated communication, and the effect of chronemics on relational communication. *Journal of Organizational Computing* 5(4):355378, URL http://dx.doi.org/10.1080/10919399509540258.

Wang FY, Carley KM, Zeng D, Mao W (2007a) Social computing: From social informatics to social intelligence. *IEEE Intelligent systems* 22(2):79–83.

Wang LC, Baker J, Wagner JA, Wakefield K (2007b) Can a retail web site be social? *Journal of Marketing* 71(3):143157, URL http://dx.doi.org/10.1509/jmkg.71.3.143.

Weiss A, Bartneck C (2015) Meta analysis of the usage of the godspeed questionnaire series. *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* URL http://dx.doi.org/10.1109/roman.2015.7333568.

Wilson H, Daugherty P, Bianzino N (2017) When ai becomes the new face of your brand. *Harvard Business Review* 27.

Wirtz J, Patterson PG, Kunz WH, Gruber T, Lu VN, Paluch S, Martins A (2018) Brave new world: service robots in the frontline. *Journal of Service Management* 29(5):907931, URL http://dx.doi.org/10.1108/josm-04-2018-0119.

Xu K, Lombard M (2017) Persuasive computing: Feeling peer pressure from multiple computer agents. *Computers in Human Behavior* 74:152162, URL http://dx.doi.org/10.1016/j.chb.2017.04.043.

SUPPLEMENTAL APPENDICES

Appendix A: Multicollinearity

We also assessed whether our regression results are subject to multicollinearity issues. As one might expect, the interaction between the *Treatment Count* and *Cash Offer* exhibits a relatively high correlation with the constituent terms, and produces a relatively high variance inflation factor (VIF) in turn. However, it should be kept in mind that high VIFs are not typically problematic when they result from correlations between interaction terms and their constituent variables. To demonstrate this in our setting, we apply the residual-centering approach of (Lance 1988). The resulting regression yields very similar results to the baseline model, and the VIF values (reported along side the centered model values) are well within normal thresholds (see Table A.1). Ultimately, we conclude that collinearity is not influencing the results.

Table A.1 Treatment Count Model (LPM; DV = Convert)

Variable	Original Model	Residual Centering	VIF	1/VIF
1 *Treatment*	0.076** (0.032)	.0756** (.0317)	2.46	0.407
2 *Treatments*	0.060** (0.030)	.0603** (.0297)	2.46	0.4062
3 *Treatments*	0.109** (0.049)	.1213** (.0508)	1.85	0.5408
1 *Treatment* · *Cash Offer*	0.052 (0.064)	.0115 (.0162)	3.42	0.2926
2 *Treatments* · *Cash Offer*	0.086 (0.069)	.0216 (.0175)	3.43	0.2916
3 *Treatments* · *Cash Offer*	0.211** (0.087)	.0509** (.0121)	2.46	0.4072
Cash Offer	−0.058 (0.057)	.0062 (.0057)	1.01	0.9936
Intercept	0.017 (0.017)	.0174 (.0169)	Mean VIF	2.44
Observations	323	323		
R^2	0.037	0.036		
F	3.85*** (7, 316)	3.82*** (7, 315)		

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Appendix B: Estimator Choice

One possible concern with our results is that they are somehow dependent upon bias or inconsistency of the LPM (Horrace and Oaxaca 2006). It is important to note, however, first, that the typical concerns with bias and inconsistency of OLS and binary outcomes are not applicable to experimental treatment impact evaluations (Deke et al. 2014). Second, even in observational data, Horrace and Oaxaca (2006) have shown that the bias underlying LPMs is unlikely to be severe when the vast majority of predicted values a resulting model yields fall entirely within the 0-1 range. In the event that any predicted values do lie outside the feasible range, those authors propose the application of a trimming estimator. This estimator is a standard LPM that simply omits those observations holding infeasible predicted values. Employing this procedure notably only results in our excluding 5 observations from the original sample and, as can be seen in Table B.1, our coefficients remain essentially unchanged.

Table B.1 Trimmed OLS (LPM; DV = Convert)

Coefficient	Model (1)
1 *Treatment*	0.0870^{***} (0.0279)
2 *Treatments*	0.0711^{***} (0.0259)
3 *Treatments*	0.1171^{**} (0.0461)
1 *Treatment* · *Cash Offer*	0.01938 (0.0217)
2 *Treatments* · *Cash Offer*	0.0295 (0.0240)
3 *Treatments* · *Cash Offer*	0.0295^{**} (0.0230)
Cash Offer	−0.0223 (0.0217)
Intercept	0.0054 (0.0075)
Observations	318
R^2	0.035
F	3.82^{***} (7, 310)

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Although a Logistic regression is often viewed as preferable when dealing with binary outcomes, because it has the desirable property of constraining predicted values to lie within the 0-1 interval, it is important to keep in mind that this estimator also has the *undesirable* property of yielding coefficients that are difficult to understand or interpret. This is true for two reasons. First, logistic regression deals with odds ratios, which

often lack straightforward intuition, given their multiplicative nature. Second, the coefficients and standard errors associated with interaction terms in Logistic Regression are not directly interpretable (Ai and Norton 2003). That said, we also estimated a logistic regression model, the results of which are presented below in Table B.2. As can be seen, these results are qualitatively similar to those reported elsewhere, in terms of sign and statistical significance of the estimated coefficients.

Table B.2 Logit (DV = Convert)

Coefficient	Model (1)
1 $Treatment$	3.879^{***} (1.083)
2 $Treatments$	3.641^{***} (1.099)
3 $Treatments$	3.813^{***} (1.168)
1 $Treatment \cdot Cash\,Offer$	1.127^{***} (0.337)
2 $Treatments \cdot Cash\,Offer$	1.268^{***} (0.361)
3 $Treatments \cdot Cash\,Offer$	1.533^{***} (0.351)
$Cash\,Offer$	-1.161^{***} (0.324)
Intercept	-6.168 (1.036)
$Observations$	323
$Wald\ Chi^2$	29.14^{***}
$Pseudo\ R^2$	0.0659

Note: Robust SEs; ** $p < 0.05$, * $p < 0.10$.

Appendix C: Randomization Checks

In this section, we report additional randomization checks, evaluating the orthogonality of cash offer and treatment assignment to one another, as well as between cash offer assignment and the clothing items that a subject brought to the buy-back procedure. Table C.1 shows the pairwise comparisons of the average cash offer assigned between alternative conditions, defined in terms of the number of treatments assigned. In Table C.2, we also report pairwise comparisons between each of the eight individual conditions, defined in terms of the unique combination of treatments assigned. All mean differences are statistically insignificant at the p<.05 level. These null results indicate that cash offer assignment was orthogonal to anthropomorphism treatment assignment.

Table C.1 Pairwise Comparisons cash offer and Number of Treatments

Test	Condition Comparison	*t*-stat	*p*-value
1	1 Treatment vs 2 Treatments	-0.228	0.820
2	1 Treatment vs 3 Treatments	-0.700	0.485
3	2 Treatments vs 3 Treatments	-0.527	0.600
4	1 Treatment vs 0 Treatments	-0.957	0.340
5	2 Treatments vs 0 Treatments	-1.117	0.266
6	3 Treatments vs 0 Treatments	-1.405	0.164

Finally, evaluating the correlation between the cash offer a subject was assigned and the number of clothes he or she was was seeking to sell (conditional on their progressing beyond the cash offer offer stage of the conversation), we again observed a statistically insignificant relationship ($p > 0.05$), implying that cash offer assignment was orthogonal to customer characteristics.

Table C.2 Pairwise Comparisons Cash Offer and Combination of Treatments

Test	Condition Comparison	*t*-stat	*p*-value
1	SP vs D	.093	0.93
2	SP vs H	0.172	0.86
3	SP vs SP & D	0.107	0.92
4	SP vs D & H	0.02	0.98
5	SP vs SP & H	-0.759	0.45
6	SP vs SP & D & H	0.554	0.58
7	SP vs Control	0.808	0.42
6	D vs H	0.052	0.96
7	D vs SP & D	0.00	1.00
8	D vs D & H	0.062	0.95
9	D vs SP & H	.771	0.44
10	D vs SP & D & H	-0.582	0.56
11	D vs Control	.093	0.54
12	H vs SP & D	-0.059	0.95
13	H vs D & H	-0.1370	0.89
14	H vs SP & H	-1.070	0.29
15	H vs SP & D & H	-0.777	0.44
16	H vs Control	0.445	0.77
17	SP & D vs D & H	-0.071	0.94
18	SP & D vs SP & H	-0.883	0.38
19	SP & D vs SP & D & H	-0.668	0.51
20	SP & D vs Control	0.710	0.48
21	D & H vs SP & H	-0.707	0.48
22	D & H vs SP & D & H	-0.505	0.61
23	D & H vs Control	0.717	0.48
24	SP & H vs SP & D & H	0.178	0.86
25	SP & H vs Control	1.717	0.09
26	SP & D & H vs Control	1.398	0.17

SP= Social Presence, D=Delay, H=Humor

Appendix D: Individual Treatments

The focus of the study is on the affects of anthropomorphism. However, from a practical standpoint, it is likely useful to also understand which of our treatments are most effective, individually. We therefore conducted additional analyses and another experiment on Amazon Mechanical Turk, aimed at addressing this question.

First, we report on our Turk experiment. In this experiment, we evaluated the desirability of individual treatment interventions in terms of subjects' reported perception of chatbot likeability. We limit this analysis to an Appendix, because it is not altogether clear whether responses from this artificial setting, in which subjects are paid to participate, would necessarily mirror results obtained in a field setting, wherein individuals organically opt into chatbot interactions. That said, results of this analysis may provide a useful indication of which anthropomorphic interventions may be particularly useful in practice.

We recruited 426 subjects on Mechanical Turk to interact with four versions of our chatbot, assigned at random: i. control, ii. social presence, iii. delay and iv. humor. We limited participation such that a given Turker could complete the HIT exactly once, to avoid concerns about interference and cross-over across conditions. After Turkers interacted with the their assigned chatbot, they were asked to rate the chatbot on the Mathur and Reichling (2016) enjoyable/unpleasant scale, which ranges from -100 to 100. We find that the humor treatment yields a significantly larger, positive effect than either the control or the delay treatment - Mann-Whitney U tests indicate statistical significance at conventional levels (p ¡ 0.05). We provide a visual depiction of the average likeability report by experimental condition in Figure D.1.

We also note here that the delay treatment yields significantly lower likeability than the control condition in this setting. As noted earlier, it is not clear whether this finding would also apply to our field setting. It should be kept in mind that crowd-workers are paid for their time. As such, our delay treatment in this setting not only manipulates anthropomorphism; it also implies that turkers are earning a lower effective wage. Moreover, we would note that we also do not account for interactions between different anthropomorphic treatments here. As such, it remains possible that delay can have a strictly positive, amplifying effect, as long as it is implemented in tandem with other anthropomorphic treatments.

Next, we revisited the data from our initial field experiment. A natural approach to consider is to simply remove our *Treatment Count* dummies and replace them with individual treatment dummies, along with all possible interactions. Unfortunately, such a model is under-powered, and yields a statistically insignificant

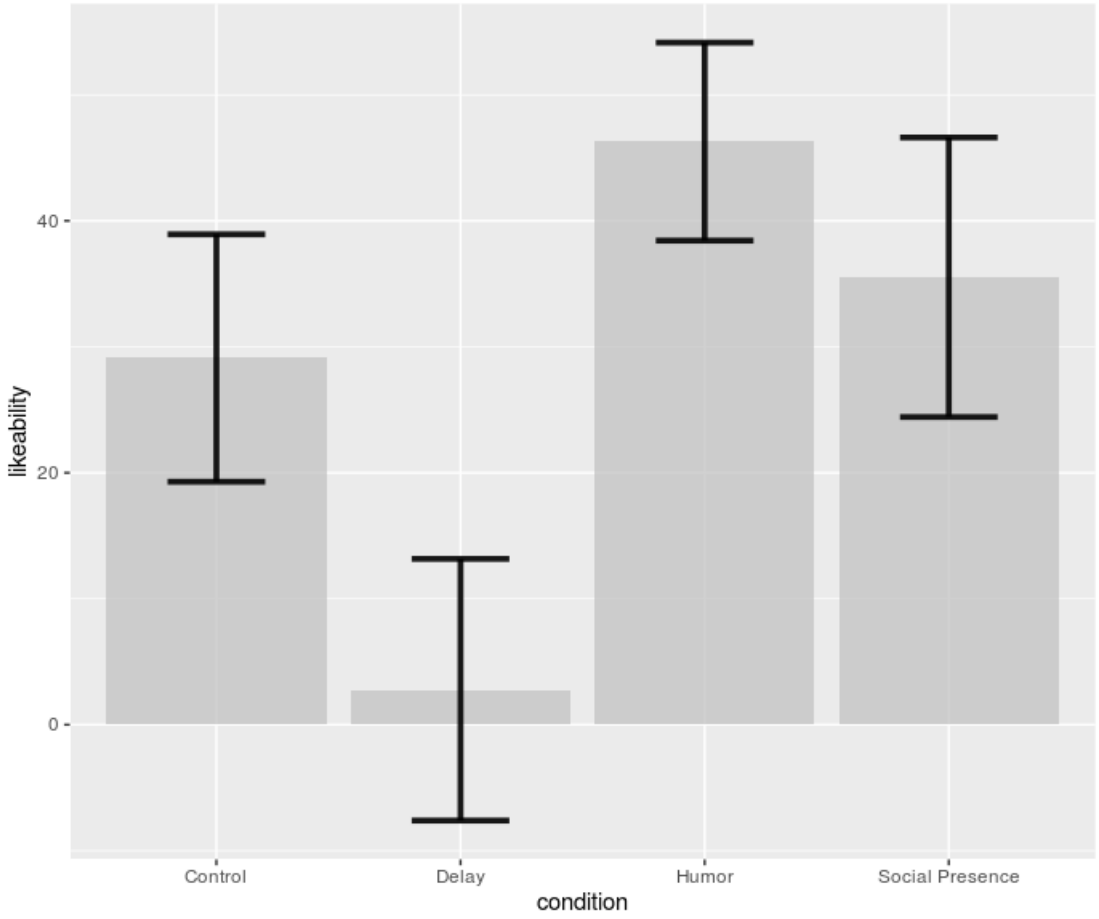

Figure D.1 Perceived Likeability - Individual Anthropomorphic Treatments

overall model fit. Accordingly, we considered a simpler regression specification, which ignores interactions between anthropomorphism treatments, and merely seeks to assess average main effects of each individual treatment, as well as their offer interactions. The model remains valid, of course, because all treatments and offer manipulations were varied exogenously.

Notably, this new model, estimated using the Horrace and Oaxaca (2006) trimming estimator, is statistically significant overall. The models yields an F-stat of 2.70 (7, 301), implying a p-value of 0.01 for overall model fit. The model results appear below in Table D.1. We observe results consistent with those seen in our Amazon Mechanical Turk study, above. That is, *Humor* has a significant, positive effect on conversion, whereas the coefficients on our two other interventions are null. Further, the effect of *Humor* is significantly larger than the effect of *Delay* ($p = 0.07$). Additionally, we see that *Delay* has a statistically significant interaction with *Cash Offer*, suggesting it has a particular influence on offer sensitivity. Of course, these

results are far from conclusive; additional work should be pursued to identify the ideal anthropomorphic interventions for retail settings.

Table D.1 LPM (DV = Convert)

Coefficient	Trimmed LPM (1)
$Delay$	-0.012 (0.030)
$Humor$	0.067^{**} (0.032)
$SocialPresence$	0.018 (0.031)
$Delay \cdot Cash\,Offer$	0.156^{**} (0.054)
$Humor \cdot Cash\,Offer$	-0.002 (0.051)
$SocialPresence \cdot Cash\,Offer$	0.091 (0.053)
$Cash\,Offer$	-0.103 (0.057)
$Observations$	309
$F-stat$	$2.70^{*}(7, 301)$
R^2	0.052

Note: Robust SEs; ** $p<0.05$, * $p<0.10$.