Can we Help the Bots? Towards an Evaluation of their Performance and the Creation of Human Enhanced Artifact for Emotions De-escalation

Biagio Palese
Northern Illinois University
bpalese@niu.edu

Matthew D. Pickard
Northern Illinois University
mpickard@niu.edu

Marcin Bartosiak
University of Pavia
marcin.bartosiak@unipv.it

Abstract

We propose a hybrid intelligence socio-technical artifact that identifies a threshold where the chatbot requires human intervention in order to continue to perform at an appropriate level to achieve the pre-defined objective of the system. We leverage the Yield Shift Theory of Satisfaction, the Intervention Theory and the Nudge Theory to develop meta requirements and design principles for this system. We discuss the first iteration of implementation and evaluation of the artifact components.

"These mechanical slaves jump to our aid. As we step into a room, at the touch of a button, a dozen light our way... Another sits night and day at our automatic refrigerator. They start our car, run our motors, shine our shoes and curl our hair." - Jay Nash (1932)

1. Introduction

Nash made the above statement almost 90 years ago, yet these "mechanical slaves" increasingly automate tasks and decisions in every aspect of our life. AI is expected to generate up to 15.7 trillion dollars in the global economy by 2030 [1]. By 2022, 70% of all customers interactions will use machine learning, chatbots and mobile messaging [2]. Yet conversational AI will not likely replace humans in the near future. Consider complex, high-emotion conversational tasks such as technical support, suicide prevention and domestic violence hotlines. AI quickly breaks down in these contexts [3]. How do we, at least partially, scale the human conversational intelligence needed to handle these complex situations? How does an organization detect if its bot is effectively managing user emotions? If a situation requires conversational and emotional intelligence beyond the bot's capability, when and how should a human operator intervene? These questions motivate a hybrid intelligence approach. Together, human and artificial intelligence can continually learn, improve, and exceed their individual performance capacities [4, 5, 6]. In this vein, we focus on the interface between human and conversational AI.

We explore how bots can be "mechanical helpers" (not slaves) that augment humans. Our research question is: How can we leverage hybrid intelligence to scale and maximize the impact of conversational systems on social good? Specifically, we explore this question in the context of high-emotion conversational contexts (e.g., technical support). Many organizations employ chatbots to address common questions, tasks, and requests for their customers and users because it is often impractical to hire or outsource sufficient human resources. Bots tirelessly automate many repetitive human tasks. However, chatbots often create an impersonal face for the organization and have limited—but ever increasing—ability and intelligence [7]. Can chatbots self-diagnose situations where they are unable to satisfactorily help users (e.g., de-escalating customer emotions) and request that a human intervene?

To explore this quest, we propose a hybrid intelligence system that leverages emotion analysis to determine which chat conversations need human intervention. We study how to measure the performance of a chatbot in the context of a specific goal, identify a threshold for human intervention, and how to communicate the need for intervention to a human operator. We apply the Design Science Research Methodology (DSRM)— a cyclical approach where the researchers search the problem space for optimal solutions [9]. In this paper, we focus our efforts on one iteration of the DSRM. We identify and motivate the problem, implement and evaluate a solution, and communicate the results [8]. The design of the socio-technical (ST) artifact is guided by five meta-requirements (MR) and eight design principles (DP) derived from three kernel theories – Yield Shift Theory of Satisfaction [10] for the user side of the artifact and Intervention Theory [11] and Nudging Theory [12] for the operator side of the artifact.

Despite a growing literature in human-bot collaboration [13, 14], only few studies explore the real-time evaluation of chatbot performance with the objective of enhancing bot-human collaboration [15]. We propose a system that will identify and prioritize conversations at

URI: https://hdl.handle.net/10125/79375
978-0-9981331-5-7

risk and encourage human intervention using a performance score and an intervention threshold. When intervention is needed, the system will provide the human operator with contextual information to maximize the operator's ability to affectively intervene [13]. The real-time nature of the bot-to-human transfer should be unnoticeable to the end user; but, more importantly it will reduce harmful damage that an ineffective bot might cause [15]. For example, effective transitions of control can de-escalate customer frustrations that damage company reputation. They can also increase user perceptions of love, concern, and empathy that provide users with reason to remain loyal to the company in the future.

In the next section we define the problem and introduce the DSR model, the kernel theories, and the MRs and DPs. We then describe the implementation, evaluate the components, and discuss the results, future steps, and conclusions.

2. Problem Definition

While chatbots automate and scale solutions to common tasks, they have many limitations. Previous research analyzes the adoption of conversational agents [16], examines trust between human and bots [17], and develops chatbot features that make the customer interactions more pleasurable or humanized [7, 18]. While research explores human-bot collaboration [13, 14], only a few studies focus on evaluating real-time chatbots performance, identifying and measuring their limitations, and enable humans to intervene with higher intelligence and skills [15, 19]. Research still needs to find new ways to identify at-risk conversations and determine how and when it is necessary to alert humans [15]. In addition, to the best of our knowledge there is no research that evaluates bot performance from a user perspective [15] and that offers a system that can be implemented without forcing companies to create or buy new chatbots [19]. We contribute to the literature by presenting a conversational system that monitors chatbot performance using emotion analysis of the user utterances, identifies a threshold for intervention, and empowers human with information to successfully intervene.

3. Designing the Artifact: a Human Enabled Chatbots Helper

While it would be ideal to design a single ST artifact that can be generally applied to all high-emotion conversational contexts, we suspect that different high-emotion contexts will be sufficiently specific that, at least initially, we will need to tune the artifact to a specific context. Because technical support chatbot conversations are publicly available, we focus our efforts in the initial iteration on developing a hybrid intelligence conversational system for customer support. Following the DSR process [20, 21, 22], we start designing the artifact by deriving meta requirements from kernel theories and formulating the design principles that will guide its development [23]. We then show the first iteration of ST artifact implementation and its components evaluation.

3.1. Kernel Theory: Yield Shift Theory

We use Yield Shift Theory of Satisfaction (YST) to explain why we expect to be able to establish an intervention threshold by monitoring the user's emotions [10]. YST assumes that individuals subconsciously and automatically attribute a utility and likelihood to each goal they desire to achieve. The utility is the perceived benefit associated with goal achievement. The likelihood is the perceived probability of goal achievement. The product of utility and likelihood determine the perceived yield (i.e., yield = utility x likelihood). Specifically, likelihood moderates the effect that utility has on yield. Thus, the yield associated with a goal with high utility and low likelihood could be less than the yield associated with a goal with low utility and high likelihood.

YST's phenomenon of interest is the satisfaction response—an emotion. The satisfaction response is determined by shifts in perceived yield over time. YST posits three theoretical strategies to induce a yield shift: 1) change the utility individuals attribute to their active goals, 2) change the likelihood individuals attribute to their goals, and 3) change the set of active goals. A chatbot is most likely to cause a yield shift in the first two ways—a shift in perceived utility or a shift in perceived likelihood. For example, assume a customer uses a support chatbot to return a product. If the chatbot informs the customer that a return will incur a 20% restocking fee, this would create a negative utility shift (i.e., the customer changes his mind about the return, given the new fee information). However, if the chatbot is unable to locate the customer's order, this would create a negative likelihood shift (i.e., the customer loses belief that the chatbot can effectively resolve the problem). Such shifts in utility or likelihood result in yield shifts that that trigger satisfaction responses—emotions—that can be detected in the users utterances to the chatbot. In this way, YST provides a theoretical framework for establishing a human intervention threshold.

3.2. Kernel Theory: Intervention Theory

We use Intervention Theory [11] as a guide to design the intervention informing interface (i.e., the human operator side) of the ST artifact. Intervention Theory explains that "to intervene is to enter in an ongoing system of relationships […] An intervenor, in this view, assists a system to become more effective in problem-solving, decision making and decision implementation in such a way that the system can continue to be increasingly effective in these activities and have a decreasing need for the intervenor" [11:15]. In a situation where the intervenor is the human operator, we imagine an intervention interface [24] that fosters an effective and timely response from the operator.

According to Intervention Theory, there are three principles that should guide the design of interventions: leveraging valid and useful information, allowing free and informed choice, and fostering internal commitment. Valid information is information that can be verified and is known to affect the situation the intervenor is trying to influence. Useful information is information which the operator can leverage to control the situation. Free and informed choice refers to the central role of the operator in the design and implementation of the intervention. Operators should be supplied with enough information to let them decide on their own. The presence of free informed choice strengthens operator's internal commitment about performing an action - a precondition to the successful intervention. Internal commitment refers to the degree of responsibility the individual feels with respect to the intervention. The strength of internal commitment comes from operator's sense of purpose, and their belief about the control they have over their actions and the outcome of these actions. These three principles are interdependent – the presence of valid and useful information enables the operator to make decisions that are free and informed and fosters the operator's internal commitment to take an appropriate action.

3.3. Kernel Theory: (Digital) Nudge Theory

The term 'nudge' means 'any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives' [25].

Nudge theory [25] is founded on the premise that individuals often make choices based on the intuitive response to the choice environment in which the decision should be made [26]. The underlying idea, called *libertarian paternalism,* posits that the designer can alter the choice environment (thus, becoming a choice architect) so that the more beneficial choices become more salient or convenient. As a consequence, an individual facing the choice is more likely to select a more beneficial option without giving up the freedom of choice [12].

IS scholars introduced term 'digital nudging' to investigate nudges enabled by digital technology [27, 28]. Digital nudging is defined as any attempt to influence decision-making, judgment, or behavior in a predictable way by counteracting the cognitive boundaries, biases, routines, and habits that hinder individuals from acting to their own benefit in the digital sphere [29].

The concept of nudging is considered a 'digital specific phenomenon' because, even if mirroring the physical world, digital choice environments are highly visual and, thus, are better suited for influencing people [30]. Information overload is often higher in digital choice environments [31] and individuals have to manage the information flow and understand the information itself simultaneously [32]. Thus, they tend to make decisions faster, based mostly on heuristics and cognitive biases [30]. We use digital nudging to trigger a human intervention in chatbots conversation at risk.

3.4. Meta Requirements and Design Principles Discovery

We conceptualize the chatbot system as an socio technical (ST) artifact [13]. Based on the above three kernel theories, we present MRs to design a system that include both IT and social elements of the user-chatbot conversation and the operator's intervention.

First, YST posits that individuals subconsciously attribute a utility and likelihood to each goal they desire to achieve. This perceived yield triggers a satisfaction response that is manifest in the individuals' emotions. The ST artifact needs to be able to detect a change in emotion. Furthermore, according to Intervention Theory, interventions must be based on valid and useful information. The ST artifact aiming at triggering effective intervention should provide the operator a data stream of recent salient emotions that help the operator to intervene effectively.

Since a satisfaction response is an affective response, the system needs to be able to detect a valanced change in the user's emotion. To do so, the ST artifact will need to quantify both a direction (positive or negative) and magnitude of the user's emotion change. In short, if the system detects negative emotions, YST implies a decrease of the user's perception of utility or likelihood towards the chatbot. Therefore, by measuring emotion valence changes, the system can identify the optimal threshold in the negative shift and use it as a valid information to trigger human intervention.

Moreover, Intervention Theory requires that the information is not only valid but also useful. Useful

information is the information which the operator would be able to use to control the development of a situation. Thus, the ST artifact needs to analyze the revealed emotions in real-time and pass the insights to the operator when the human intervention can cause a positive change.

MR1. ST artifact collects and analyzes emotions expressed in the conversation to detect a satisfaction response

DP1.1 Use sentiment analysis to find valence change in the user's text

DP1.2 Quantify the direction of valence change

DP1.3 Quantify the magnitude of valence change

MR2. ST artifact provides information about user emotions in real-time

DP2.1 Analyze the revealed sentiment change in real-time

DP2.2 Alert the human operator immediately when the valence changes negatively

The second principle of Intervention Theory assumes that an effective intervention is based on free and informed choice. This requires that the ST artifact not only alerts the operator about the potential situations to intervene, but also provides enough data to let the operator interpret the situations and decide how to react. Thus, the operator can freely decide when to take over and how to collaborate in problematic conversations based on the interpretation of the information provided by the ST artifact. The ST artifact might, for example, visualize the valence change over time and extract the topics from the text of the conversation so that when operators are alerted, they do not have to read the whole conversation or ask questions that may further irritate or frustrate the user. Rather, they should be able to decide about the intervention and intervene immediately when they receive the alert from the ST artifact.

MR3. ST artifact contextualizes the emotions and conversation topics for the operator

DP3.1 Provide visualizations of the valence change over time

DP3.2 Provide visualizations of the topics of the conversation over time

The last principle of Intervention Theory involves fostering internal commitment in the operator. In general, when individuals feel a high degree of responsibility with respect to an intervention, the potential intervention tends to be more successful than when they perceive low degree of responsibility. To increase the internal commitment of the operator, the ST artifact can attempt to influence operator's attitudes and judgements in that they perceive the task as giving them high degree of responsibility. Designing technology for influence is the realm of the field of persuasive technology - defined as "any interactive computing system designed to change people's attitudes or behaviors" [33:1]. In particular, the ST artifact can build on digital nudging. Building on bounded rationality [34], digital nudging allows the individuals to enjoy the freedom of choice but sways them toward choices that are more beneficial for them and reinforces that behavior. Any user-interface design element can be designed to guide people's behaviors in the expected direction [28].

MR4. ST artifact triggers an intervention decision from the human operator

DP4.1 The interface nudges to the operator to intervene in conversations in which the chatbots are potentially in trouble assisting users.

4. Artifact Implementation

The artifact implementation of a DSR project occurs in multiple design iterations that lead to improvements of the ST artifact and modifications of the original MRs and/or DPs [8, 35]. In this section, we report the first iteration of the implementation of the ST artifact.

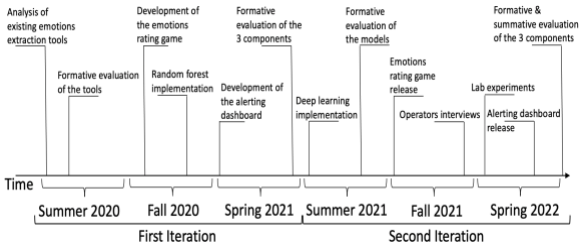

Figure 1 Artifact design and development roadmap

In Figure 1 we also provide a roadmap of future steps and iterations accompanying them with formative and summative evaluations that will help us in further developing and enhancing the ST artifact. To satisfy the meta requirements and implement the design principles derived from our kernel theories we have envisioned three major components of our ST artifact. The first component consists in an emotions extraction tool. Being able to quantify the satisfaction response directly from the conversations text is a core feature of our artifact which provides useful information for the operator. However, it is also important to make sure that the provided emotion information is valid. So, our second component is a data collection (emotions rating) game. The game is considered as an evaluation utility, namely a means to crowd-source a large set of labeled utterances to better train the emotions extraction tool with and to assess its accuracy. Finally, the third component, an alerting dashboard, has the objective to communicate the information to operators and to nudge them about problematic conversations. So, the alerting

dashboard is the front-end element of the emotions extraction tool and the emotions rating game. Together these three components constitute a single conversational system operators interact with in real time.

4.1. Emotions Extraction Tool

The purpose of the emotions extraction tool is to measure and monitor emotions during chatbot conversations. To develop our tool, we were able to acquire training data from real technical support chats. Given the context of those conversations, we focused on measuring the two emotions that are intuitively most salient to the customer's satisfaction response, namely anger and happiness. For the tool we considered and compared different solutions. From sentiment analysis packages available in software, to APIs and deep learning. We started the search by using three sentiment analyzers in R. However, they all resulted to perform poorly on our data (around 40% accuracy). The main reason of their ineffectiveness is that they are vocabulary-based sentiment analyzers. We then evaluated the performance of existing emotional analysis APIs (e.g., IBM Watson). Collectively, those APIs were able to achieve approximately 85% accuracy after enhancing them with random forest models. We are currently building deep learning models and we expect to have results ready for presentation at the conference. In the next section we provide more details of the implementation that focused on the APIs solutions.

4.1.1 Data

IS researchers use gold standard sets to assess accuracy and performance of different models and to identify the optimal one [36]. To create a gold standard set to evaluate the different APIs with, we randomly selected a subset of twelve conversations that, at face value, contained different levels of anger and happiness from technical support chats. The twelve conversations had a total of 197 utterances. Where 104 of the utterances were made by human users and the remaining 93 by bots.

4.1.2 Labeling the Data

To establish a gold standard to benchmark the performance of the emotion analysis APIs, two authors and a graduate assistant manually labeled the happiness and anger scores perceived in each of the 197 utterances on a scale of 0 (not angry/happy) to 5 (extremely angry/happy)[1]. Both the happiness (Krippendorff α = 0.72) and anger (Krippendorff α = 0.73) scores had acceptable interrater reliability. While $\alpha > 0.8$ is desirable, α values ≥ 0.667 are sufficient for tentative conclusions [37]. Given the subjective nature of perceiving emotions in text, $\alpha > 0.7$ is acceptable for our purposes. To create a composite score from the three raters, we averaged the three rater scores and rounded to the nearest whole number (0 to 5). At the end of this step, we had a human labeled gold standard set that we can use to compare the different solutions already available for emotions extraction.

4.1.3 Emotion Analysis APIs

We assessed three emotion analysis APIs. The scores provided by each API are as follows (we hypothesized the bolded scores would best measure **happiness** and the italicized scores would best measure *anger*):

- Microsoft Azure Text Analytics[2] (MA) — **positive**, neutral, *negative* sentiment
- ParallelDots Text Analysis[3] (PD)— **happiness**, *anger*, excitement, sadness, fear, and, boredom
- IBM Watson Tone Analyzer[4] (IBM)—*anger*, fear, **joy**, and sadness

We obtained sentiment or emotion scores from each API for each conversation utterance. Our unit of analysis is an utterance (and not a conversation) because it is important for the operator to be able to monitor the emotion fluctuations throughout the conversation to make an intervention decision. We report the APIs results in the evaluation section.

4.1.4 Random Forest Model of APIs

Because each API provided different scores and no single API performed well at predicting anger or happiness, we created random forest classification models using the API scores to predict anger and happiness. We used the *randomforest* package in R [38] that implements the Breiman's algorithm. By enhancing the APIs with some context specific data, we aimed to improve their accuracy. More information about the solutions discussed above and the results they provided is available in the evaluation section.

[1] More details of the procedure are available from the authors upon request

[2] https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

[3] https://komprehend.io/emotion-analysis

[4] https://www.ibm.com/cloud/watson-tone-analyzer

Given that the accurate emotion extraction is critical to our artifact, we designed a data collection game to incentivize crowd-sourced users to label the emotional content of chatbot conversations.

4.2. Emotions Rating Game

The purpose of the emotions rating game is to crowd-source emotion-labeled chat conversation data. This component is needed to constantly enhance and evaluate our emotions extraction tool. The data collection game has two main objectives. First, score the anger and happiness in chatbot conversations. Second, establish a human intervention threshold based on the anger in the user's utterances. Both tasks are important to improve our artifact. With the first one we plan to gather more human labeled scores of utterances that we can use for training or evaluation purposes. With the second one, we plan to gain valuable information to understand when a human should intervene.

Figure 2 provides a screenshot of the game interface. Raters use the anger and happiness sliders to provide emotion scores (0 to 5). Slider movement triggers dynamic emoji changes. For example, as the rater slides the scale from 0 to 5, the emoji becomes happier. Raters are also instructed to click the "Call the manager" button when they feel the customer's anger has escalated sufficient to merit human intervention. To further encourage rater engagement, after the rater scores a full conversation, the game presents them with a speed and accuracy score. Speed is simply measured by the amount of time it takes for the rater to submit a score. The countdown clock at the top of the interface further incentivizes the rater to provide a real time assessment of the emotion of each utterance that simulates a real scenario conversation. In fact, a small amount of time elapses among utterances and we need humans to provide their "gut instinct" scores. Moreover, the speed element also gives to the game a feeling of competitiveness. We measure users' accuracy score as the difference between the rater's score and the average crowd score. However, we test and exercise caution to ensure that users are not influenced to simply agree with the crowd.

Figure 2 - Emotions Rating Game

4.3. Alerting Dashboard

The purpose of the alerting dashboard is to effectively present to operators information about many user-bot conversations and enable them to identify and effectively intervene in problematic conversations. Thus, from a high-level perspective, the dashboard is the only component that operators directly interact with. The main objective of the dashboard is to monitor and inform operators about conversations that the emotions extraction tool assesses as high-risk. The dashboard also needs to present rich, contextual information about conversations in an intuitive, concise manner. The operator should be able to drill down in conversations and obtain more detailed information that will allow her to effectively intervene.

Figure 3 shows an initial prototype of the alerting dashboard created in R Shiny, using the *shinydashboard* package [39]. The top portion provides conversational-level assessments. We decided to color code the different users conversation to enhance and facilitate operators' ability to identify conversations at risk. Where red and orange indicate those conversations that might require operators' attention. For example, the operator could easily see that Jones is manifesting the largest amount of anger and decide to further select Jones from the pull-down menu to understand whether Jones's anger is rising, falling, or level. Ideally, at the conversational-level, the operator could see a list sorted based on the salient emotion—in this case, anger. By doing so operators will have all the information not only to identify problematic conversations but also to monitor overtime users' emotions changes.

Figure 3 - Alerting Dashboard Prototype

5. Artifact Evaluation

In this section we discuss the first formative evaluation of the different components of our artifact. It was necessary for us to design and implement the artifact in a sequential order. For this reason, some components are in a more advanced stage than others. For example, if we are not able to extract useful information, users emotions, with the emotions extraction tool, we have no reason to build a component (e.g., the game) that improves and evaluates such extraction process and we

can't visualize those information with a dashboard. Moreover, using the same reasoning the game comes before the dashboard because it ensures that the information is not only useful but also valid. Furthermore, the game provides us indications of the threshold level at which other humans would want someone else to jump into a chatbot conversation. That threshold is critical for us to visualize and alert operators of conversations that really need their attention. The first iteration formative evaluations of the three components are discussed below.

5.1. Emotions Extraction Tool

As discussed above we used a human labeled gold standard set to evaluate the different solutions used to extract emotions from chatbot conversations.

5.1.1 Correlation of API and Rater Scores

To confirm our hypotheses in section 4.1.3, we correlated the raters' scores for happiness and anger with the individual API scores (see Table 1). The IBM joy (r = 0.733), MA positive (r = 0.652), and PD happy (r = 0.728) scores correlated highly with the raters' happy scores. None of the API scores correlated highly with the raters' angry scores. IBM anger (r = 0.413) and PD anger (r = 0.469) had the highest correlations. So, while those APIs performance is acceptable for positive sentiment, they underperform in detecting negative sentiment.

Because the conversations in which there is a negative sentiment (e.g., anger) are those that require human attention, we built random forest models with these variables.

5.1.2 Random Forest Model of APIs

Random forest methods combine the output of multiple uncorrelated decision trees into a single classification estimation. The individual trees are created using different random subsets of the original data and features (see input features in Table 1). These subsets of data are called bags. Each bag is used to train a single decision tree. Any leftover data is consider "out-of-bag" and is used to evaluate the trained decision tree. Thus, OOB accuracy is evaluated at training time. This bootstrapping approach prevents overfitting. The number of trees was set at 500 for both the anger and happiness model.

Table 1 - Rater and API Score Pearson Correlations

	HAPPY	ANGRY
Ibm_anger	-0.072	**0.413**
ibm_fear	-0.052	0.119
ibm_joy	**0.733**	-0.162
ibm_sad	-0.178	0.393
ma_neg	-0.355	0.257
ma_neutral	-0.291	0.016
ma_pos	**0.652**	-0.273
pd_angry	-0.249	**0.469**
pd_bored	-0.344	0.155
pd_excited	**0.518**	-0.371
pd_fear	-0.367	-0.083
pd_happy	**0.728**	-0.399
pd_sad	-0.497	0.133

We split the data into train (80%, n = 158) and test (20%, n = 39) sets. The accuracy of the predictions from the test set of data is called cross-validation accuracy. It gives a measure of the performance of the ensemble of decision trees. Importantly, no part of the ensemble model has ever seen any of test set data.
We tested both full (all the API scores) and reduced (only API scores that correlated highly with the human anger and happiness ratings) models. There was not a significant performance difference, so we present the results of these reduced models (which align with our hypotheses and the correlation results):

happy = ibm_joy + pd_happy + pd_excited + ma_pos

anger = ibm_angry + pd_angry + ma_neg

The out-of-bag (OOB) accuracies for the models were: happy (92.39%) and anger (84.81%). The cross-validation accuracies were: happy (100%) and anger (89.74%).

Table 2 - Random Forest Confusion Matrix for Happiness

		Predicted						
		0	*1*	*2*	*3*	*4*	*5*	Class error
Actual	*0*	27	0	0	0	0	0	0.00
	1	0	7	0	0	0	0	0.00
	2	0	0	0	0	0	0	0.00
	3	0	0	0	3	0	0	0.00
	4	0	0	0	0	1	0	0.00
	5	0	0	0	0	0	1	0.00

Cross-validation error rate = 0.00%

As can be seen from the cross-validation confusion matrices (Table 2 and Table 3), the base rate of utterances with high emotion is low. Additionally, based

on the correlations and the random forest model, anger remains more difficult to extract, yet the random forest model represents a significant improvement to existing APIs to establishing the intervention threshold. A larger training set will likely allow the model to capture richer patterns predictive of high emotion, especially high anger, and the emotion rating game will help us in creating such set.

Table 3 - Random Forest Confusion Matrix for Anger

		Predicted						
		0	*1*	*2*	*3*	*4*	*5*	Class error
Actual	*0*	32	0	0	0	0	0	0.00
	1	0	1	1	0	0	0	0.50
	2	0	0	1	0	0	0	0.50
	3	0	0	0	0	0	0	0.00
	4	0	0	0	1	0	0	0.00
	5	0	0	0	0	2	1	0.00

Cross-validation error rate = 10.26%

5.2. Emotions Rating Game

The emotions rating game went through a different series of evaluations. During the pilots we made sure of three things. The first one is that the instructions (training) at the beginning of the game are clear, and that humans understand the purpose of the game and how to use the sliders and call the manger button. After multiple feedback from testers (5 people not involved in the project), we consider the training phase to be clear also to people outside the research team. The second one involves making sure that the users of the game stay engaged with it and rate for us as many conversations as possible so that we can create a larger labeled set. After different users feedback, we added multiple gamifications elements, from the speed timer to the ranking (e.g., leaderboard). We also decided to make the ranking available only to users that rate at least 10 conversations, so that we encourage them in playing longer and label more conversations. Our testers felt that the game is engaging and that those elements would encourage people in labeling multiple conversations. The third one is making sure the game records human raters' submissions correctly and that those data are stored correctly and ready to be used for analysis into a MySQL database. We anticipate to lunch the game at scale during Fall 2021.

5.3. Alerting Dashboard

We created three different prototypes of the dashboard. In each prototype we visualize the relevant elements (e.g., color coding of chat or over time emotions trend) with a different layout and interface. The research team assessed the different versions and suggested improvements based on the MRs and DPs. While as of right now we have not performed a formative evaluation of the dashboard with people external to the research project, we plan to evaluate the alerting dashboard by interviewing and observe operators while using them. We will then perform A/B testing with different versions. Given that the operators are our end users, such assessment will enable us to identify which version is optimal for them in performing their jobs.

6. Discussion

The proposed ST artifact represents a DSR improvement contribution [27]. The ST artifact implementation is the results of the derived MRs and advanced DPs. We contribute to the existing literature by offering a new DSR approach to enhance human-bot collaboration [13, 16]. By leveraging kernel theories we propose an artifact that takes in consideration both the users that are in conversation with bots and the operators that are monitoring such conversations [15]. Such approach enables us to not only evaluating bots performance but also to determine when and in which conversations human should intervene. More specifically the emotions extractions tool enables us to extract, analyze and contextualize users emotions while interacting with bots over time and in real time. The emotions rating game allows us to continuously improve the extraction process and to determine intervention threshold that can guide operators in deciding when and if intervene. Moreover, the alerting dashboard empowers the operators to monitor each conversation and nudges to the operators those conversations that require extra attention. Thus by leveraging the game and dashboard we respond to the open call of developing an alerting systems mechanism that take in consideration customers perspectives [15]. It also offers a new innovative way of cooperation between humans and chatbots [16]. Moreover, our artifact does not require the creation of a new chatbot like previous research [15] and it can be used to complement and enhance existing chatbots. We see possible applications in both private companies and non-profit organizations that leverage bots to scale interactions with their users beyond the customer support application used in this paper. Furthermore, we expect that our artifact can have a significant and positive impact in other complex, high-emotion conversational tasks such as students counseling and mentoring, natural disaster management, suicide prevention and domestic violence hotlines

7. Limitations and Future Work

As any work our research is not exempt from limitation. In developing the system, we assumed that customers are willing to stay in the chat for multiple utterances and show multiple emotions. Nonetheless, we acknowledge that not all customers are equal or behave equally in chatbot conversations. Especially when the customers have other ways to communicate with the company representatives, some of them will indeed quit the chat as soon as they realize that the bot is of no help. Furthermore, they will probably leave the chat before showing any negative emotions. Thus, our system will not benefit all customers but only those customers that decide to stay for multiple utterances. Moreover, we have just evaluated individual components of our artifact and not the artifact as a whole. Such evaluation is part of our future work agenda.

Furthermore, the first iteration presented in this paper represents only the basis for future development. In fact, given the evaluation results we envision the following adjustments in the second iteration. For the emotions extractions tool, while combined random forest model of the API scores performed acceptably, our next iteration will focus on two items of improvement. First, we will run custom deep learning models to predict the anger and happiness content of utterances. Second, we will use the data collected using the rating emotions game to improve the original random forest model and our deep learning models. We will use the random forest model as a baseline comparison and expect significant improvements to the results currently reported.

Additionally, we are currently searching for a large, accessible, and context appropriate chatbot data set (e.g., technical support). Human raters will label such dataset once we release the emotions rating game (Fall 2021). In addition to the perceptual measures the game provides, we plan to perform lab experiments where we will use facial recognition software and EEG to monitor the rater's face and brain waves while participants evaluate the conversations in the game. Then we will correlate the facial and brain waves data with the perceptual scores they provide to the game.

Finally, we plan to interview operators and ask them to test the three dashboard versions by the end of Fall 2021 and to improve our alerting system as soon as we get more data about when to intervene from the game.

8. Conclusions

While researchers are increasingly studying human-bot collaboration. There is still the need of conversational systems that enable humans to help bots and enhance such collaboration. Hybrid intelligence offers an opportunity to develop a system that transform chatbots from mere "mechanical slaves" to valuable teammates (aka "mechanical helpers"). The ST artifact presented in this paper offers a possible implementation of such systems.

9. References

[1] PricewaterhouseCoopers, "2019 AI Predictions: Six priorities you can't afford to ignore", *PwC*, 2019. https://www.pwc.com/us/en/services/consulting/library/artificial-intelligence-predictions-2019.html

[2] Solutions, A., "The Value of Chatbots According to Gartner", *Medium*, 2020. https://chatbotslife.com/the-value-of-chatbots-according-to-gartner-d8468688240f

[3] Modlinski, A., and M. Bartosiak, "Replaced by Machines? The Intelligent (ro)bots as the Disruptive Innovation for Human Workforce in Cross-Cultural Perspective", In *Cross Cultural Management*. Wydawnictwo Uniwersyteu Lodzkiego, 2021, 69–96.

[4] De Cremer, D., and G. Kasparov, "AI Should Augment Human Intelligence, Not Replace It", *Harvard Business Review*, 2021. https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it

[5] Dellermann, D., P. Ebel, M. Söllner, and J.M. Leimeister, "Hybrid Intelligence", *Business & Information Systems Engineering 61*(5), 2019, pp. 637–643.

[6] Wilson, H.J., and P.R. Daugherty, "Collaborative Intelligence: Humans and AI Are Joining Forces", *Harvard Business Review*, 2018.

[7] Rapp, A., L. Curti, and A. Boldi, "The human side of human-chatbot interaction: A systematic literature review of ten years of research on text-based chatbots", *International Journal of Human-Computer Studies 151*, 2021, pp. 102630.

[8] Peffers, K., T. Tuunanen, M.A. Rothenberger, and S. Chatterjee, "A design science research methodology for information systems research", *Journal of management information systems 24*(3), 2007, pp. 45–77.

[9] Simon, H.A., *The sciences of the artificial*, MIT press, 1996.

[10] Briggs, R., B. Reinig, and G.-J. de Vreede, "The Yield Shift Theory of Satisfaction and Its Application to the IS/IT Domain", *Journal of the Association for Information Systems 9*(5), 2008.

[11] Argyris, C., *Intervention Theory and Method: A Behavioral Science View*, Addison-Wesley, Reading, MA, 1970.

[12] Leonard, T.C., "Richard H. Thaler, Cass R. Sunstein, Nudge: Improving decisions about health, wealth, and

happiness", *Constitutional Political Economy 19*(4), 2008, pp. 356–360.

[13] Seeber, I., E. Bittner, R.O. Briggs, et al., "Machines as teammates: A research agenda on AI in team collaboration", *Information & Management 57*(2), 2020, pp. 103174.

[14] "Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature - Thomas O'Neill, Nathan McNeese, Amy Barron, Beau Schelble, 2020", https://journals.sagepub.com/doi/full/10.1177/0018720820960865

[15] Poser, M., S. Singh, and E. Bittner, *Hybrid Service Recovery: Design for Seamless Inquiry Handovers between Conversational Agents and Human Service Agents*, 2021.

[16] Lewandowski, T., J. Delling, C. Grotherr, and T. Böhmann, *State-of-the-Art Analysis of Adopting AI-based Conversational Agents in Organizations: A Systematic Literature Review*, 2021.

[17] Elson, J.S., D. Derrick, and G. Ligon, "Examining Trust and Reliance in Collaborations between Humans and Automated Agents", (2018).

[18] Grudin, J., and R. Jacques, "Chatbots, Humbots, and the Quest for Artificial General Intelligence", *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Association for Computing Machinery (2019), 1–11.

[19] Liu, J., Z. Gao, Y. Kang, et al., "Time to Transfer: Predicting and Evaluating Machine-Human Chatting Handoff", *arXiv:2012.07610 [cs]*, 2020.

[20] Hevner, A.R., S.T. March, J. Park, and S. Ram, "Design Science in Information Systems Research", *MIS Q. 28*(1), 2004, pp. 75–105.

[21] Sein, M., O. Henfridsson, S. Purao, M. Rossi, and R. Lindgren, "Action Design Research", *Management Information Systems Quarterly 35*(1), 2011, pp. 37–56.

[22] Walls, J.G., G.R. Widmeyer, and O.A. El Sawy, "Building an Information System Design Theory for Vigilant EIS", *Information Systems Research 3*(1), 1992, pp. 36–59.

[23] Piccoli, G., M.Ł. Bartosiak, B. Palese, and J. Rodriguez, "Designing scalability in required in-class introductory college courses", *Information and Management 57*(8), 2020.

[24] Baskerville, R.L., M.D. Myers, and Y. Yoo, "Digital first: The ontological reversal and new challenges for IS research", *MIS Quarterly 44*(2), 2020, pp. 509–523.

[25] Thaler, R.H., and C.R. Sunstein, *Nudge: Improving Decisions About Health, Wealth, and Happiness*, Penguin Group USA, New York, 2009.

[26] Kosters, M., and J. Van der Heijden, "From mechanism to virtue: Evaluating Nudge theory", *Evaluation 21*(3), 2015, pp. 276–291.

[27] Gregor, S., and A. Hevner, "Positioning and Presenting Design Science Research for Maximum Impact", *Management Information Systems Quarterly 37*(2), 2013, pp. 337–355.

[28] Weinmann, M., C. Schneider, and J. vom Brocke, "Digital Nudging", *Business & Information Systems Engineering 58*(6), 2016, pp. 433–436.

[29] Mirsch, T., C. Lehrer, and R. Jung, "Making Digital Nudging Applicable: The Digital Nudge Design Method", *ICIS 2018 Proceedings*, 2018.

[30] Lembcke, T.-B., N. Engelbrecht, A. Brendel, B. Herrenkind, and L. Kolbe, *Towards a Unified Understanding of Digital Nudging by Addressing its Analog Roots*, 2019.

[31] Benartzi, S., and J. Lehrer, *The Smarter Screen: Surprising Ways to Influence and Improve Online Behavior*, Penguin, 2017.

[32] Lee, B.-K., J.-Y. Hong, and W.-N. Lee, "How Attitude toward the Web Site Influences Consumer Brand Choice and Confidence While Shopping Online", *Journal of Computer-Mediated Communication 9*(2), 2004.

[33] Fogg, B.J., *Persuasive Technology: Using Computers to Change What We Think and Do*, 2002.

[34] Simon, H.A., "A Behavioral Model of Rational Choice", *The Quarterly Journal of Economics 69*(1), 1955, pp. 99–118.

[35] Venable, J., J. Pries-Heje, and R. Baskerville, "FEDS: a Framework for Evaluation in Design Science Research", *European Journal of Information Systems 25*(1), 2016, pp. 77–89.

[36] Palese, B., and G. Piccoli, "Evaluating Topic Modeling Interpretability Using Topic Labeled Gold-standard Sets", *Communications of the Association for Information Systems 47*(1), 2020.

[37] Krippendorff, K., "Reliability in Content Analysis", *Human Communication Research 30*(3), 2004, pp. 411–433.

[38] Breiman, L., and A. Cutler, *Breiman and Cutler's Random Forests for Classification andRegression*, 2018.

[39] Winston Chang and Barbara Borges Ribeiro, *Create Dashboards with "Shiny"*, 2018.