# CHAPTER XY

# SEEKING ATTENTION: TESTING A MODEL OF INITIATING SERVICE INTERACTIONS

SEBASTIAN LOTH, KERSTIN HUTH, JAN P. DE RUITER

http://dx.doi.org/10.1163/9789004260160_012

## 1. Introduction

Initiating a customer-staff interaction is a highly significant part of any service encounter. Failing to recognise the customer's interest to initiate an interaction may be fatal for the service encounter as a whole. Such mistakes are preventable and thus, constitute one of the worst outcomes of a service encounter (Smith, Bolton, and Wagner

1999). Even if the customer is eventually noticed and offered a high compensation, the initially failed and repaired service was perceived as worse than an error-free service (McCollough, Berry, and Yadav 2000). Consequently, accurately detecting a customer's intention to interact is essential in service encounters.

An increasing number of robots is used in the service domain (International Federation of Robotics 2013). These robots have to interact with customers who do not have prior knowledge about the technology being used. Thus, a robot has to understand the customers' natural intuitive behaviour, specifically when initiating an interaction. This is one of the most difficult challenges in a bar scenario with several customers. The cues that are used for signalling the intention to order a drink might be very subtle, e.g. if a customer sits at the bar and decides to order another drink, s/he might not get up or move to another location. On the other hand, the system should not respond to everybody because inviting customers to place an order if they had no intention to do so is annoying for them. Additionally, recognising the scene through the robot's sensors is complicated by the fact that bars are often dimly lit and noisy environments. Thus, an explicit and robust account of detecting customers who wish to place an order is crucial.

## 1.1. Intention recognition

Identifying customers who would like to order a drink requires that the robot recognises the actions that the customers are currently performing and, most importantly, it has to understand the customers' intention. Goals and intentions have to be distinguished at various levels (Jeannerod 2006; Searle 1983; Van Overwalle and Baetens 2009). We use the term *intention* to refer to the meaning that an agent aims to communicate by performing an action. In contrast, *goal* refers to the immediate effect that an action or a sequence of actions may have. For example, cracking an egg into a bowl could be part of a sequence of actions with the goal of preparing scrambled eggs. The agent's intention could be to communicate that s/he takes care of the meal whereas the interlocutor should continue with another task. Levinson (1995, 227) referred to this kind of communicative actions as *signals* and argued that the agent's intention is the premise of the observable actions. In terms of logic, inferring the intention means identifying the premise from a given conclusion (observable actions) which is logically intractable (Levinson 1995, 231). This is due to the fact that there is an infinite set of premises that would warrant the same conclusion, e.g. conclusion $p$ can be drawn given $q\&p$ or $q\&(q\rightarrow p)$ or $(p|q)\&(p\&\neg q)$ and so on. However, humans can understand social signals by relying on heuristics and their knowledge of normally expected behaviour (Levinson 1995). Thus, our approach was to make use of the social skills of customers, bartenders and participants in lab experiments in order to identify the relevant signals normally used in the bar scenario.

The first step in understanding an agent's intention is action recognition. This was defined as matching the percept of an action to a corresponding action in memory (e.g., Jeannerod 2006). In humans, so-called mirror-neurons contribute to recognising actions and to identifying their goal (Iacoboni et al. 2005; Johnson-Frey et al. 2003; Kilner, Friston, and Frith 2007; Wurm and Schubotz 2012; Van Overwalle and Baetens 2009). The bartending robot has to rely on computer vision for recognising non-verbal actions. Research in this area has focused on correctly classifying actions such as waving, walking and running (Poppe 2010). Also, the agent's pose (Shotton et al. 2013), hands and face can be identified and tracked (Baltzakis, Pateraki, and Trahanias 2012). That means the robotic sensors are able to extract the posture, movements and actions performed by the customers in (close to) real-time. As a second step, these data have to be interpreted in a social context.

Within the clearly defined scenario of a bartending robot, deriving the customers' intentions can be simplified. Before a service interaction has started, the robot has to distinguish between customers who would like to initiate an interaction and those who do not. This distinction has to be precise and robust. In order to achieve this, a set of explicit rules is required which specifies what is necessary and what is sufficient for recognising that a customer is bidding for attention. The necessary signals are present in all interactions. Thus, the absence of a necessary signal allows the system to conclude that it should not respond to a customer. This prevents that the robot is overly responsive and in turn, annoying customers who do not wish to place an order. The set of sufficient signals includes all necessary signals and possibly some additional signals. If the robot detects the sufficient signals, it should invite a customer to place an order; otherwise the robot would appear unresponsive. For example, in order to make tea it is necessary to boil water but this could be used for making coffee as well. Thus, if there is no hot water, there will be no tea (absence of a necessary signal). But the presence of hot water (presence of a necessary signal) is not sufficient to conclude that there will be tea. Finally, apart from defining the necessary and sufficient signals, a general preference to either invite or not to invite a customer to place an order has to be specified, e.g. if the robot's sensor data are inconclusive. We review related work in the next section and introduce our natural data collection and the experiment in the following sections.

## 1.2. Related work

Orkin and Roy (2007; 2009) used an online restaurant game for collecting the behaviour of several thousand players. The recordings were used for generating action maps for a virtual waiter. However, the results showed that relying on observable behaviour alone was not sufficient for deriving a meaningful structure of the interactions.

Using handcrafted models and/or deriving models from lab data did not work as intended in the real world (e.g., Bohus and Horvitz 2009; Michalowski, Sabanovic, and Simmons 2006). For example, Michalowski et al. (2006) presented human-robot data collected with a robotic receptionist. Relying on proxemics (Hall 1969), their model triggered a greeting whenever a potential interlocutor was close to the robot. But people found it disturbing when they passed by the reception desk and the robot greeted them (cf. Goffman 1963; Michalowski, Sabanovic, and Simmons 2006, 766). The false alarms were due to defining the set of sufficient signals for initiating an interaction too loosely, i.e. triggering a greeting too easily.

Sidner and her colleagues relied on gaze direction, mutual face gaze, adjacency pairs, and backchannels (Holroyd, Ponsler, and Koakiettaveechai 2009; Holroyd et al. 2011; Rich et al. 2010; Sidner et al. 2005; Sidner and Lee 2003). This model was inspired by research on human behaviour in lab sessions and research on social behaviour (e.g., Schegloff and Sacks 1973). In starting an interaction, the model relied on eye gaze. However, measuring a user’s eye gaze requires that an eye tracking system is calibrated in advance of each interaction, which is not suitable in a real-world application. Thus, we derived a set of rules from real world data for ensuring its applicability in the real world.

## 2. Natural data collection

A video corpus of real-life customer-staff interactions at a bar was recorded in several club locations in Germany. In total, 108 service interactions were recorded including 105 bids for attention. The customers' actions in the time span just before the bartender invited them to place an order were annotated using the ELAN annotation software (Wittenburg et al. 2006). The data are summarised in Table *1* and show how many bids for attention included at least one occurrence of each signal. That means the exact timing of the actions was ignored as well as how often each signal occurred within a single interaction.

| Behaviour | Number of interactions |
|---|---|
| *Customer body and posture* | |
| Body to bar | 95 |
| Head to bar | 93 |
| Being at the bar | 92 |
| Approaching bar | 44 |
| Leaning on bar | 12 |
| Turning to bar | 11 |
| Further away from bar | 4 |
| *Customer head and looking direction* | |
| Looking at bartender | 86 |
| Head gesture | 11 |
| Looking at money | 7 |
| Looking at assortment | 3 |
| Looking at menu | 1 |
| *Mimic* | |
| Raising eyebrows | 5 |
| Smiling | 1 |
| *Customer attention focus* | |
| Attention to bartender | 91 |
| Attention to (another) human | 32 |
| Attention to object | 49 |
| *Customer hand movements* | |
| Holding object/bottle | 17 |
| Hand gesture to bartender | 4 |
| *Customer speech* | |
| Speaking to bartender | 10 |
| Speaking to others | 21 |

Table 1: Summary of customer behaviour per interaction in 105 bids for attention

The frequency data in Table *1* reflects the observed behaviour of the customers. But, as mentioned above, relying on observable behaviour alone is not sufficient for determining what exactly was meaningful to the bartenders (cf. Levinson 1995). However, the distinction between behaviour that coincided with a response and behaviour that triggered a response is crucial. For example, if customers scratched their heads frequently; this behaviour would occur with a high frequency. Yet, this is not necessarily informative, i.e. head scratching and bidding for attention coincide but this does not imply a causal relationship. However, the real-life data provided a solid base for deriving hypotheses about informative signals which have to be then validated in experiments.

Candidates for necessary and sufficient behaviours were identified using the data in Table *1*. The analysis was limited to distinguishing between highly frequent behaviours occurring in almost all interactions (e.g., *looking at bartender* in 82% or in 86 out of 105 interactions) and rare behaviours (e.g., *looking at money* in 7% or in 7 out of 105 interactions). All customers were right at the bar or approached the bar. Accordingly, *Being at the bar* was identified as a candidate for a necessary signal. The remaining high frequency behaviours *attention to bartender*, *looking at bartender, head to bar* and *body to bar* are similar as they indicate that the customers looked at the bar. We summarised all the contributing behaviours in a single signal and referred to it as *Looking at the bar*. As outlined above, using an eye tracker and calibrating it to the customers is not feasible in a real world application, thus the attention focus and gaze cannot be estimated reliably. However, head and body orientation provides a reliable indication of where a person is looking. Thus, *Looking at the bar* (approximated by head and body orientation) is another candidate for necessary signals. The data in Table *1* suggest that customers successfully attracted the attention of the bartender only by *Being at the bar* and *Looking at the bar,* whereas other behaviours were optional. Thus, we hypothesised that this set of two signals is sufficient. In sum, the analysis of the natural data collected suggests that the set of signals formed by *Being at the bar* and *Looking at the bar* (approximated by head and body orientation) is necessary and sufficient.

## 3. Experiment

The aim of this experiment was to test whether the hypothesised necessary and sufficient signals were exhaustive and minimal.

Additionally, we investigated whether participants checked the signals in a particular order and what kind of errors they committed. This information is important, e.g. in order to define a response if the sensor data are inconclusive. To avoid ambiguity, participants in the lab experiments are referred to as *participants* and people who participated in the natural data collection are referred to as *customers*.

Participants performed a classification task of snapshots taken from the real-life corpus which preserved as much of the social context as possible. This avoided problems associated with staging stimuli and contrasts with placing a robotic system in the wild for collecting data (Bohus and Horvitz 2009). In particular, if participants have to interact with an existing system, they might adapt to this specific implementation and deviate from their natural behaviour. Thus, we relied on real-life stimuli for investigating natural and unbiased interactions. A potential downside of a lab-setting is the time flow of events. When participants in the lab are asked to respond to a snapshot, they do not experience the time constraints of a real social interaction where the response delays are typically very short. For example, research on turn-taking has shown that interlocutors try to anticipate the end of a turn for a seamless conversation (De Ruiter, Mitterer, and

Enfield 2006). In order to hinder participants from extensive introspection, a time limit was enforced. The time limit was set such that the accuracy of the response did not unduly suffer.

## 3.1. Methods

*Participants.* Thirty-one participants from the university population volunteered for the experiment and received €3 in exchange for their time.

*Materials and design.* Participants were asked to imagine they were in the role of a bartender and to indicate through the buttons of a gamepad whether the snapshot showed a customer who was bidding for their attention (*yes*-response) or whether there was no customer who required their attention (*no*-response).

In order to test whether each of the two identified signals was necessary, snapshots were selected such that only one of the signals was present. Thirty-nine snapshots were selected from the natural data collection such that customers stood or sat at the bar, but did not look at the bar/bartender (e.g., customers searching their bag or engaging in another conversation). This condition is referred to as *Being at the bar*. Accordingly, 39 snapshots of people *Looking at the bar*, but not being at the bar were selected. These snapshots depicted customers who had turned towards the bar from some distance. The snapshots in these conditions showed customers who were not placing an order. If these signals were both necessary, *no*-responses were expected in both conditions.

The experiment included two types of *yes*-trials. First, snapshots of actual orders were used and are referred to as *Ordering*. These snapshots were expected to trigger *yes*-responses. This condition formed the baseline and tested whether the participants were able to perform the task successfully. The second *yes*-condition used snapshots of customers who were not actually bidding for attention, but accidentally produced both signals. These snapshots showed customers who *were at the bar* and *looked at the bar*, but did not bid for attention. If the two candidate signals formed the sufficient set, participants should be deceived into giving a *yes*-response. If some other signal was required for identifying an order, a *no*-response was expected. This condition is referred to as *Not ordering*. Only 37 of these stimuli could be identified. In order to balance the number of expected *yes*- and *no*-responses, 41 snapshots of real orders were included. Furthermore, the number of expected *yes*- and *no*-responses was matched for each club location. Examples of the snapshots are presented in Figure *1*.

About 11 hours of recorded materials were scanned for selecting the snapshots according to the conditions of the experiment. The snapshots were double checked to ensure that all visible customers were to be classified in the same condition, e.g. all customers in the snapshot were bidding for attention. This allowed us to attribute the participants' response to a specific condition.

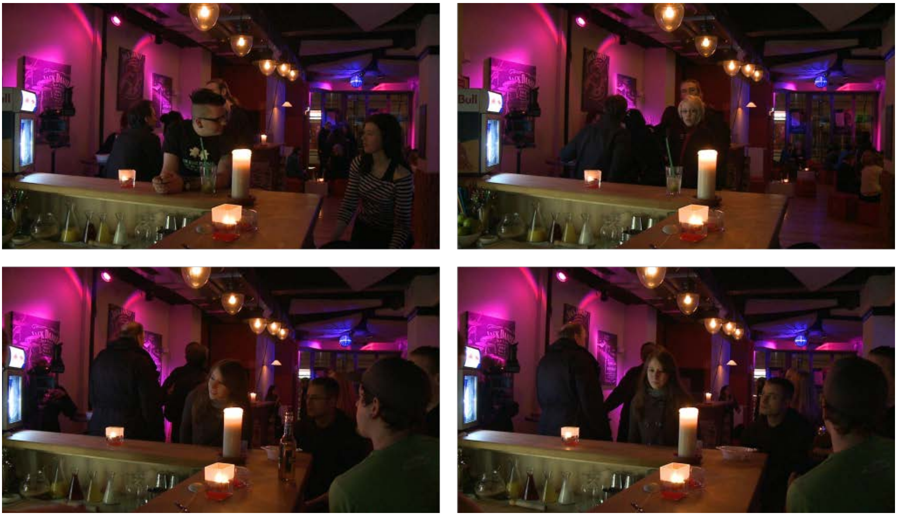

Figure 1: A grid of example snapshots recorded in the "Movie", Bielefeld. Top left: Being at the bar (*no*-response expected); Top right: Looking at the bar (*no*-response expected); Bottom left: Ordering (*yes*-response expected); Bottom right: Not ordering (*yes*-response expected)

*Procedure*. Participants were informed that snapshots taken from recordings in bars would be presented on screen. They were seated in front of a computer screen and their written consent was collected. A gamepad was handed to the participants and its red (*no*-response) and green (*yes*-response) marked buttons were explained. The gamepads were prepared so that participants used their dominant hand for giving a *yes*-response and the other hand for *no*-responses. All presentations on screen and the measurement of response times were controlled by

DmDX (version 4.0.4.9, Forster and Forster 2003). The task instructions were presented on screen and asked the participants to indicate by pressing the respective button whether the snapshot showed a customer who was bidding for their attention. Each trial started with a 500 ms presentation of a fixation cross which informed participants about the upcoming snapshot. Following it, each snapshot was presented for a maximum of 3000 ms. The image disappeared as soon as participants responded and the screen remained blank for 500 ms. If participants failed to respond within 3000 ms, an on-screen message informed them that their response was too slow. This message was the only information about time limit. No other feedback was provided during and after the experiment. The experimental sessions commenced with four practice trials resembling each of the conditions in the experiment. These items were not repeated. After a self-paced break, the 156 experimental items were presented in random order. The session was interrupted by self-paced breaks every 39 trials. The experimental session took about 15 min. A general debriefing was provided after the experimental session.

## 3.2. Results

The practice items were excluded from the analysis. Out of 4836 trials, 67 (1.40%) did not receive a response (see Table *2*), i.e. each participant exceeded the time limit without giving a response in about two trials on average. The number of missed responses did not differ significantly by condition [$\chi^2$(3, *N*=4836) = 2.307, *p*=.511]. All missed responses were excluded from further analyses.

| Condition | Expected response | Number of | | | | Response score |
|---|---|---|---|---|---|---|
| | | Missed responses | Valid responses | *Yes*-responses | *No*-responses | |
| *Being at the bar* | *No* | 14<br>(1.2%) | 1195<br>(98.8%) | 292<br>(24%) | 903<br>(76%) | -0.51 |
| *Looking at the bar* | *No* | 16<br>(1.3%) | 1193<br>(98.7%) | 319<br>(27%) | 874<br>(73%) | -0.47 |
| *Ordering* | *Yes* | 16<br>(1.3%) | 1255<br>(98.7%) | 1034<br>(82%) | 221<br>(18%) | +0.65 |
| *Not ordering* | *Yes* | 21<br>(1.8%) | 1126<br>(98.2%) | 947<br>(84%) | 179<br>(16%) | +0.68 |

Table 2: Categorial results of the experiment

All responses were scored as +1 if the participant pressed the *yes*-button and -1 in case of a *no*-response. Thus, a perfect agreement amongst all participants that a snapshot showed a customer bidding for attention would result in a mean response score of +1.00 and that no customer bid for attention in a score of -1.00. Random responses would result in a mean response score close to 0.00. The mean values for each condition are presented in Table 2. It should be noted, that the expected response and the correct response were not always equal. Specifically, the majority of participants produced *yes*-responses in the *Not ordering* condition. This was compatible with our expectation, but actually a *no*-response would have been correct.

The response scores were analysed using a binomial test. For each of the four conditions, this showed that the response scores were significantly different from 0.0: *Being at the bar* [$Z$ = 17.646, $p$<.001], *Looking at the bar* [$Z$ = 16.039, $p$<.001], *Ordering* [$Z$ = 22.291, $p$<.001] and *Not ordering* [$Z$ = 22.857, $p$<.001]. In order to evaluate whether the location of the recordings and the handedness of the participants had any effect on the results, a binary logistic regression was performed using condition (coding whether a *yes*- or a *no*-response was expected), handedness and a dummy recoding of the three bar locations as independent variables. The analysis showed that the condition was the only statistically significant predictor of the responses [$Z$ = 1367.248, $p$<.001]. There was no statistically significant effect of handedness [$Z$ = 1.882, $p$=.170] or the variables coding location [$Z$ = 1.863, $p$=.172] and [$Z$ = 1.724, $p$=.189]. The difference in explained variance of the full model [Cox and Snell $R^2$=.302] and the model using condition as the only predictor variable [Cox and Snell $R^2$=.300] was negligible, thus the location and handedness were not considered in further analyses of this dataset.

The categorial responses in each condition were compared using Chi-square tests. The small numerical difference between the conditions receiving predominantly *no*-responses *Being at the bar* and *Looking at the bar* [$\chi^2$(1, $N$=2389) = 1.754, $p$=.185] was not statistically significant. Similarly, there was no statistically significant difference between the conditions that were predominantly associated with *yes*-responses: *Ordering* and *Not ordering* [$\chi^2$(1, $N$=2381) = 1.245, $p$=.264]. A Chi-square test was also performed for comparing the level of agreement in participants' judgement, i.e. comparing whether the proportion of expected and unexpected responses differed across conditions. The expected *no*-responses in the *Being at the bar* and *Looking at the bar* conditions were compared to the expected *yes*-responses in the *Ordering* and *Not ordering* conditions. The test revealed a statistically significant difference [$\chi^2(1, N=4769) = 55.100, p<.001, \phi=0.11$] indicating a greater agreement when participants were expected to give *yes*-responses compared to the *no*-responses.

The categorial responses were also analysed using signal detection theory. The *Being at the bar* and *Looking at the bar* trials reflected snapshots where the signal was absent and a *no*-response was expected, i.e. no customer was bidding for attention. These two conditions were combined. Similarly, the *Ordering* and *Not ordering* trials were combined (see Table *3*). The results showed that *d'* was 1.62, which indicated that participants performed well above chance. The bias was 0.31 which indicated that the participants preferred *yes*-over *no*-responses.

| | *Yes*-response | *No*-response |
|---|---|---|
| Signals present<br>(*yes*-response expected) | Hit<br>0.832 (1981) | Miss<br>0.168 (400) |
| Signals absent<br>(*no*-response expected) | False alarm<br>0.256 (611) | Correct rejection<br>0.744 (1777) |

Table 3: Proportions of yes- and no-responses as a function of the presence of the two signals *Being at the bar* and *Looking at the bar*. The numbers in brackets show the absolute number of responses.

To analyse the response times (RTs, see Table *4*), a mixed model analysis was performed using R (R development core team 2007) and lmer in the lme4 package (Baayen, Davidson, and Bates 2008; Bates and Sarkar 2007; Bates 2005). The results are reported as *F*-test. If the effect was significant at conventional levels ($\alpha$ = .05) the effect size according to Cohen (1969, 348) computed using G*Power (Faul et al. 2007) is reported. The difference in mean RT was tested using a Markov chain Monte Carlo (MCMC) simulation with 10,000 steps (Baayen, Davidson, and Bates 2008; for examples see Brysbaert 2007)[1]. The MCMC probability and the corresponding effect size of the equivalent *t*-test (Cohen 1969, 38) are reported. The analyses included participants, items and location as sources of random variance.

| Condition | Expected response | *Yes*-responses | | *No*-responses | |
|---|---|---|---|---|---|
| | | Mean RT | SD | Mean RT | SD |
| *Being directly at the bar* | *No* | 1558 ms | 483 ms | 1459 ms | 493 ms |

[1] Baayen, Davidson and Bates (2008, 396–397) suggested that Markov chain Monte Carlo (MCMC) simulations for directly sampling from the posterior distribution of the parameters offers one option for avoiding some fallacies of using the *t*- and *F*-distributions. This specific implementation uses non-informative priors. For a general introduction to MCMC see, e.g. Andrieu, de Freitas, Doucet and Jordan (2003).

| | | | | | |
|---|---|---|---|---|---|
| *Looking at the bar* | *No* | 1550 ms | 512 ms | 1352 ms | 493 ms |
| *Ordering* | *Yes* | 1327 ms | 461 ms | 1543 ms | 534 ms |
| *Not ordering* | *Yes* | 1313 ms | 494 ms | 1567 ms | 524 ms |

Table 4: Results of the experiment. The response times and their standard deviations were computed for valid responses.

The mixed model analysis tested whether the expected responses were performed faster or slower than unexpected responses. This analysis is comparable to the analysis of correct and false responses in decision experiments. There was a significant difference [$F(1, 4678) = 90.324$, $f$=0.14] indicating that expected responses were performed faster than unexpected responses [$M_{diff}$ = 191 ms, *pMCMC*<.001, *d*=0.38].

As with the nominal data, we were interested in whether there was a difference between the two conditions associated with the same response. The mixed model included a term for testing these contrasts within the expected and unexpected responses (condition was a nested factor under expectation). The analysis showed a small, but significant effect of this term on RT [$F(6, 4673) = 4.506$, $f$=0.08]. The comparison of the expected *no*-responses to *Being at the bar* and *Looking at the bar* revealed a statistically significant difference [$M_{diff}$ = 107 ms, *pMCMC*=.003, *d*=0.22]. This indicated that *no*-responses were produced faster if the customers looked at the bar from a distance compared to sitting or standing at the bar. There was no such difference in the unexpected *yes*-responses [$M_{diff}$ = 13 ms, *pMCMC*=.276]. Contrasting the *Ordering* and *Not ordering* conditions revealed no statistically significant difference in expected *yes*-responses [$M_{diff}$ = 14 ms, *pMCMC*=.706] and unexpected *no*-responses [$M_{diff}$ = 24.0 ms, *pMCMC*=.901]. Finally, we were interested in whether participants were faster to recognise an ordering customer compared to recognising that nobody was about to order. For this purpose, the *yes*-responses to the *Ordering* and *Not ordering* stimuli were combined and compared to the combination of the *no*-responses to the *Being at bar* and *Looking at bar* conditions. This analysis showed a significant difference [$M_{diff}$ = 86 ms, *pMCMC*<.001, *d*=0.18] indicating that spotting a customer was performed faster than establishing that no customer was about to order. The analysis of the unexpected responses across these conditions revealed no such difference [$M_{diff}$ = 3 ms, *pMCMC*=.630].

## 3.3. Discussion

The experiment used real-life stimuli and asked participants to indicate whether the customers depicted in a snapshot had the intention to place an order. This does not only require recognising an action but, most importantly, to interpret this action in a specific context. Thus, the social context presented in natural stimuli was crucial in this experiment. However, natural stimuli are less homogeneous than those generated in the lab. Each snapshot of our stimuli showed customers in different poses, people in the background and objects in various configurations. In contrast, lab generated stimuli typically use a constant background and control for body posture and facial expression of people appearing in the picture. Thus, responding to the more complex natural stimuli results in longer RTs than responding to lab generated stimuli. The RTs obtained in this experiment are comparable to other studies using natural stimuli, e.g. classification of grey-scale portrait photographs in female or male faces (O’Toole et al. 1998). In contrast, RTs in classification tasks using lab generated stimuli were much shorter (e.g., “Is this object human-made or natural?”, Gollan et al. 2005; “Is this a fruit or an animal?”, Snodgrass and McCullough 1986). Thus, the time limit had to be set such that participants could inspect and understand the snapshots but also effectively hinder them from an extensive introspection of their intuition. That means using natural stimuli required adapting the experimental methods, but most importantly the natural stimuli reflect the real life and increase the ecological validity of our findings.

The analysis of the baseline condition using snapshots of real orders showed that participants recognised that customers were bidding for attention with a high level of agreement (response score was 0.65, i.e. 82.5% of the responses were *yes*-responses). The signal detection analysis provided converging evidence (*d’* of 1.62). This indicates that participants were able to perform the task successfully and that the results are credible and interpretable.

The analysis of the natural data collection suggested that the two signals *Being at the bar* and *Looking at the bar* were necessary for getting the attention of bar staff. If one of these signals was absent, the participants judged the snapshots as customers not bidding for attention. This provides a clear indication that both signals are necessary. The *Not ordering* condition tested whether these signals were also sufficient for signalling the intention to place an order. The results showed that the presence of these signals was sufficiently strong to fool participants into misperceiving customers as bidding for attention who actually were not doing so. Comparing the baseline and this misleading condition showed no statistically significant difference in the categorial responses and the RTs. The similarity of the results suggests that the information processed by the participants was very similar in both conditions. Thus, we concluded that *Being at the bar* and *Looking at the bar* together form the sufficient set of signals for recognising that a customer is bidding for attention.

The analysis of the RTs suggests that participants checked the position and body posture of the customers sequentially. Participants responded faster if the customer was located further away from the bar (*Looking at the bar* condition) and they took longer if customers were right at the bar (*Being at the bar* condition). This suggests that participants checked whether there was somebody at the bar in a first step. If no customer was at the bar, one of the necessary signals was absent. This was sufficient for concluding that a *no*-response was appropriate. But if there was a customer at the bar, a second analysis of the customer's body posture, head direction, engagement in other conversations and so on was required. Only this additional analysis provided the required information for evaluating whether a *no*-response was appropriate. This explains that the *Being at the bar* condition received slower responses than the *Looking at the bar* condition. These findings suggest that the first process (checking the area at the bar) filtered the data for the second process (checking customers body and head orientation), i.e. the processes operated sequentially. However, these results do not allow excluding a parallel processing of the signals. In a parallel model, evaluating the head and body direction would always take more time than checking whether there are customers at the bar. Thus, the results of both processes would be available to the participants in sequence. The experimental data are compatible with both models. However, the sequential processing has advantages for the implementation in a robotic system. In a sequential account, the body posture is only relevant for customers who are right at the bar. In contrast, a parallel analysis requires that the head and body orientation is computed for all customers irrespectively of their distance to the bar. Thus, the computational load is lower with sequential than with parallel processing. Consequently, the sequential account is preferable for our purposes.

The results of the classification experiment enabled us to analyse the reaction times of the participants which provided valuable insight into how the snapshots were processed from a bartender’s perspective. However, this kind of experiment is limited to investigating which signals were used and does not allow investigating the relative time course of the customers’ actions and the participants’ responses. This is important for setting the response speed of a robot. If the system is too fast, the number of false alarms could be unduly high. On the other hand, if the system is too slow, the robot would appear as unresponsive. This needs to be addressed in future research.

The analysis of the unexpected responses showed that participants were careful not to miss a potential order, i.e. they tried to avoid ignoring a customer. This experiment provided three sources of evidence for this conclusion. First, there was a bias of 0.31 indicating that participants had a general preference to identify snapshots as an order (giving a *yes*-response). Secondly, participants were more accurate when a *yes*-response than when a *no*-response was expected. That means, if they made a mistake this was more likely to be a false alarm (mistaking a customer) than a miss (ignoring a customer). Thirdly, the RTs in expected *no*-responses were slower than in expected *yes*-responses. This can be attributed to an exhaustive (or at least more thorough) inspection of the snapshot when no ordering customer was identified. In turn, there was an additional effort before producing a *no*-response. These data suggest that there was a trade-off between committing false alarms (mistaking a customer) and misses (ignoring a customer). In this trade-off, participants subconsciously avoided misses (ignoring customers) by accepting an increased rate of false alarms (mistaking customers). This could be attributed to greater social cost associated to misses than to false alarms. Thus, if the sensor data of a robotic bartender are inconclusive, the robot should invite customers to place an order. In turn, the robot's behaviour would reflect that participants preferred false alarms (mistaking a customer) over misses (ignoring a customer).

In sum, two signals are necessary and together form the sufficient set of signals for identifying the intention to place an order. First, the customers position themselves right at the bar and, secondly, look at the bar/bartender. Participants checked the presence of these signals sequentially, i.e. they applied a two-step procedure. If participants misjudged a snapshot, results showed that it was preferable to invite customers to order by mistake than to ignore a customer.

## 4. Conclusions

In order to enable a bartending robot to recognise if a customer bids for attention, we developed a simple decision policy by conducting a study with natural data. First, we recorded real customer-staff interactions in bars for identifying the customers' natural behaviour when bidding for attention. Secondly, we tested which of their behaviours were interpreted as a signal in an experiment. This experiment relied on natural stimuli because they provided the social context of a bar scene for recognising social intentions and, importantly, ensured the applicability of our findings for a robotic bartender employed in the real world but also for human-human service encounters.

The results of the experiment showed that it is necessary for customers to be right at the bar and to look at the bar/bartender if they want to attract the attention of bar staff. If both signals were present at the same time, they were also sufficient. Thus, if customers who are close to the bar also look at the bar, the bartending robot should invite them to place an order. More specifically, the participants checked the customer's distance to the bar first and whether they were looking at the bar/bartender in a second step. That means that the robotic sensors have to measure the customers’ distance to the bar. Only customers in close proximity to the bar have to be analysed in more detail with regards to their body posture and head orientation, but customers who are further away can be ignored. Sequentially analysing the cues reduces the computational demand compared to processing all cues at the same time. Even though our experiment focussed on the bar scenario, this policy is relevant to settings where the service staff operates behind a counter and their customers move to an area dedicated for providing the service, e.g. service desks, ticket counters, and corner shops. In all these settings, human and robotic service staff should specifically attend this dedicated area in order to avoid missing a customer.

This relatively simple policy commits the same mistakes as humans. If both signals are present, this policy has to assume that a customer would like to place an order. The results of the experiment showed that if both signals were present, the customers in the snapshots were judged as bidding for attention regardless of whether they were actually bidding for attention or accidentally produced these signals. That means that customers expect to be invited for placing an order if they produce the relevant signals. For example, customers who are already being served quickly respond to an invitation to place an order by another member of staff without being surprised or taking offence. Thus, committing these mistakes is socially appropriate and observable in human staff rather than a fault in the robotic policy. Furthermore, this policy scales to multiple customers. If several customers approach the bartending robot, the two-step procedure applies to each customer. However, they have to be served in appropriate order (Foster et al. 2012; Petrick and Foster 2012).

Inconclusive sensor data are a source of mistakes that is specific to the robotic bartender. In the experiment, participants showed a general preference to identify a customer as bidding for attention. That means it is preferable to invite a customer to place an order rather than to ignore a customer. Thus, the robot's decision policy should express the same preference and invite customers to place an order. In sum, this policy is very robust, scalable and even its mistakes reflect natural human behaviour.

## References

Andrieu, Christophe, Nando de Freitas, Arnaud Doucet, and Micheal I. Jordan. 2003. "An Introduction to MCMC for Machine Learning." *Machine Learning* 50, 1:5–43. doi:10.1023/A:1020281327116.

Baayen, R. Harald, Douglas J. Davidson, and Douglas M. Bates. 2008. "Mixed-Effects Modeling with Crossed Random Effects for Subjects and Items." *Journal of Memory and Language* 59, 4:390–412. doi:10.1016/j.jml.2007.12.005.

Baltzakis, Haris, Maria Pateraki, and Panos Trahanias. 2012. "Visual Tracking of Hands, Faces and Facial Features of Multiple Persons." *Machine Vision and Applications* 23, 6:1141–1157. doi:10.1007/s00138-012-0409-5.

Bates, Douglas M. 2005. "Fitting Linear Mixed Models in R." *R News*, May.

Bates, Douglas M., and Deepayan Sarkar. 2007. *lme4: Linear Mixed-Effects Models Using S4 Classes* (version 0.99875-6). R package.

Bohus, Dan, and Eric Horvitz. 2009. "Models for Multiparty Engagement in Open-World Dialog." In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 225–234. London, UK: Association for Computational Linguistics. http://www.aclweb.org/anthology/W/W09/W09-3933.

Brysbaert, Marc. 2007. "'The Language-as-Fixed-Effect Fallacy': Some Simple SPSS Solutions to a Complex Problem (Version 2.0)". Royal Holloway. http://crr.ugent.be/papers/The%20language%20as%20fixed%20effect%20fallacy%20Version%202%200.pdf.

Cohen, Jacob. 1969. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

De Ruiter, Jan P., Holger Mitterer, and Nick J. Enfield. 2006. "Projecting the End of a Speaker's Turn: A Cognitive Cornerstone of Conversation." *Language* 82, 3:515–535.

Faul, Franz, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. "G*Power 3: A Flexible Statistical Power Analysis Program for Social, Behavioral, and Biomedical Sciences.” *Behavior Research Methods* 39, 2:175–191. doi:10.3758/BF03193146.

Forster, Kenneth I., and Jonathan C. Forster. 2003. “DMDX: A Windows Display Program with Millisecond Accuracy.” *Behavior Research Methods, Instruments, & Computers* 35, 1:116–124. doi:10.3758/BF03195503.

Foster, Mary Ellen, Andre Gaschler, Manuel Giuliani, Amy Isard, Maria Pateraki, and Ronald P. A. Petrick. 2012. “Two People Walk into a Bar: Dynamic Multi-Party Social Interaction with a Robot Agent.” In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*. Santa Monica, USA: ACM Press. doi:10.1145/2388676.2388680.

Goffman, Erving. 1963. *Behaviour in Public Places*. Galt, Ontario: Collier-Macmillan Canada Ltd. http://solomon.soth.alexanderstreet.com/cgi-bin/asp/philo/soth/getdoc.pl?S10019969-D000001.

Gollan, Tanar H., Rosa I. Montoya, Christine Fennema-Notestine, and Shaunna K. Morris. 2005. “Bilingualism Affects Picture Naming but Not Picture Classification.” *Memory & Cognition* 33, 7:1220–1234. doi:10.3758/BF03193224.

Hall, Edward T. 1969. *The Hidden Dimension: An Anthropologist Examines Humans' Use of Space in Public and Private.* Garden City, New York: Anchor Books, Doubleday & Company Inc.

Holroyd, Aaron, Brett Ponsler, and Punsak Koakiettaveechai. 2009. "Hand-Eye Coordination in a Humanoid Robot". Major Qualifying Project Report CR1-0802. Worcester, UK: Worcester Polytechnic Institute. http://web.cs.wpi.edu/~rich/hri/HolroydEtAl09.pdf.

Holroyd, Aaron, Charles Rich, Candace L. Sidner, and Brett Ponsler. 2011. "Generating Connection Events for Human-Robot Collaboration." In *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication*, 241–246. Atlanta, GA: IEEE. doi:10.1109/ROMAN.2011.6005245.

Iacoboni, Marco, Istvan Molnar-Szakacs, Vittorio Gallese, Giovanni Buccino, John C. Mazziotta, and Giacomo Rizzolatti. 2005. "Grasping the Intentions of Others with One's Own Mirror Neuron System." *PLoS Biology* 3, 3:e79. doi:10.1371/journal.pbio.0030079.

International Federation of Robotics. 2013. “World Robotics 2013 Service Robots”. World Robotics. http://www.ifr.org/service-robots/statistics/.

Jeannerod, Marc. 2006. *Motor Cognition: What Actions Tell the Self*. Oxford Psychology Series no. 42. Oxford; New York: Oxford University Press.

Johnson-Frey, Scott H., Farah R. Maloof, Roger Newman-Norlund, Chloe Farrer, Souheil Inati, and Scott T. Grafton. 2003. “Actions or Hand-Object Interactions? Human Inferior Frontal Cortex and Action Observation.” *Neuron* 39, 6:1053–1058. doi:10.1016/S0896-6273(03)00524-5.

Kilner, James M., Karl J. Friston, and Chris D. Frith. 2007. “Predictive Coding: An Account of the Mirror Neuron System.” *Cognitive Processing* 8, 3:159–166. doi:10.1007/s10339-007-0170-2.

Levinson, Stephen C. 1995. “Interactional Biases in Human Thinking.” In *Social Intelligence and Interaction: Expressions and Implications of the Social Bias in Human Intelligence,* edited by Esther N. Goody, 221–260. Cambridge [England] ; New York: Cambridge University Press.

McCollough, Micheal A., Leonard L. Berry, and Manjit S. Yadav. 2000. “An Empirical Investigation of Customer Satisfaction after Service Failure and Recovery.” *Journal of Service Research* 3, 2:121–137. doi:10.1177/109467050032002.

Michalowski, Marek P., Selma Sabanovic, and Reid Simmons. 2006. “A Spatial Model of Engagement for a Social Robot.” In *Proceedings of the 9th IEEE International Workshop on Advanced Motion Control,* 762–767. Istanbul: IEEE. doi:10.1109/AMC.2006.1631755.

O’Toole, Alice J., Kenneth A. Deffenbacher, Dominique Valentin, Karen McKee, David Huff, and Hervé Abdi. 1998. “The Perception of Face Gender: The Role of Stimulus Structure in Recognition and Classification.” *Memory & Cognition* 26, 1:146–160. doi:10.3758/BF03211378.

Orkin, Jeff, and Deb Roy. 2007. “The Restaurant Game: Learning Social Behavior and Language from Thousands of Players Online.” *Journal of Game Development* 3, 1:39–60.

Orkin, Jeff, and Deb Roy. 2009. “Automatic Learning and Generation of Social Behaviour from Collective Human Gameplay.” In *Proceedings of the 8th International Conference on Autonomous*

*Agents and Multimagent Systems : May 10-15, 2009, Budapest, Hungary*. International Foundation for Autonomous Agent and Multiagent Systems.

Petrick, Ronald P. A., and Mary Ellen Foster. 2012. "What Would You Like to Drink? Recognising and Planning with Social States in a Robot Bartender Domain." In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*. http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5211/5575.

Poppe, Ronald. 2010. "A Survey on Vision-Based Human Action Recognition." *Image and Vision Computing* 28, 6:976–990. doi:10.1016/j.imavis.2009.11.014.

R development core team. 2007. *R: A Language and Environment for Statistical Computing* (version 2.12.0). Wien, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Rich, Charles, Brett Ponsler, Aaron Holroyd, and Candace L. Sidner. 2010. "Recognizing Engagement in Human-Robot Interaction." In *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, 375–382. Osaka: ACM Press. doi:10.1145/1734454.1734580.

Schegloff, Emanuel A., and Harvey Sacks. 1973. "Opening up Closings." *Semiotica* 8, 4:289–327. doi:10.1515/semi.1973.8.4.289.

Searle, John R. 1983. *Intentionality, an Essay in the Philosophy of Mind*. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Shotton, Jamie, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. 2013. "Real-Time Human Pose Recognition in Parts from Single Depth Images." *Communications of the ACM* 56, 1:116–124. doi:10.1145/2398356.2398381.

Sidner, Candace L., and Christopher Lee. 2003. "Engagement Rules for Human-Robot Collaborative Interactions." In IEEE International Conference on Systems, Man and Cybernetics, 4:3957–3962. Washington, DC: IEEE. doi:10.1109/ICSMC.2003.1244506.

Sidner, Candace L., Christopher Lee, Cory D. Kidd, Neal Lesh, and Charles Rich. 2005. "Explorations in Engagement for Humans and Robots." *Artificial Intelligence* 166, 1-2:140–164. doi:10.1016/j.artint.2005.03.005.

Smith, Amy K., Ruth N. Bolton, and Janet Wagner. 1999. “A Model of Customer Satisfaction with Service Encounters Involving Failure and Recovery.” *Journal of Marketing Research* 36, 3:356–372.

Snodgrass, Joan Gay, and Brian McCullough. 1986. “The Role of Visual Similarity in Picture Categorization.” *Journal of Experimental Psychology: Learning, Memory, and Cognition* 12, 1:147–154. doi:10.1037/0278-7393.12.1.147.

Van Overwalle, Frank, and Kris Baetens. 2009. “Understanding Others’ Actions and Goals by Mirror and Mentalizing Systems: A Meta-Analysis.” *NeuroImage* 48, 3:564–584. doi:10.1016/j.neuroimage.2009.06.009.

Wittenburg, Peter, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. “ELAN: A Professional Framework for Multimodality Research.” In *Proceedings of LREC 2006*. http://tla.mpi.nl/tools/tla-tools/elan/.

Wurm, Moritz F., and Ricarda I. Schubotz. 2012. “Squeezing Lemons in the Bathroom: Contextual Information Modulates Action Recognition.” *NeuroImage* 59, 2:1551–1559. doi:10.1016/j.neuroimage.2011.08.038.