Perception of Rhythmic Agency for Conversational Labeling

Manuscript accepted for publication in Human-Computer Interaction

Christine Guo Yu, University of Cambridge
(now at Accenture Interactive UK, Email: gy238@cantab.ac.uk)

Alan F. Blackwell, University of Cambridge

Ian Cross, University of Cambridge

Abstract

There are well-understood benefits of designing intelligent interactive systems to support mixed-initiative interaction, in which there is a back-and-forth dynamic between user and system. We explore the effects of introducing rhythmic elements into this interaction, building on the phenomenon of conversational entrainment in human-to-human conversation. We demonstrate that well-chosen rhythms reduce user stress, increase confidence, and improve the sense of agency through which a person feels in control of their own actions. We apply these findings in the context of a simple intelligent labeling system (based on a deployed product), showing that a more conversational dynamic in labeling retains all of the above benefits, while also improving accuracy of user contributions by comparison to the currently universal design approach in which the user is solely responsible for maintaining any rhythm in interaction events.

Keywords: mixed initiative interaction; labeling; agency; interaction timing; rhythm

Introduction

Machine learning algorithms often need to be trained with sample datasets, prepared by humans who label the content manually. However, data labeling is tedious and repetitive, with humans often replicating trivial decisions that could have been automated, and not getting sufficient time to devote to cases where their judgment would have added more value. Much recent research has therefore focused on approaches to labeling that optimize information gain by assisting the user with trivial cases, or directing them toward more complex ones, so that the user can devote time and attention to subtle cues and ambiguous cases, potential bias (e.g. Binns, Veale, Van Kleek & Shadbolt, 2017), or inventing new label categories (e.g. Kulesza, Amershi, Caruana, Fisher & Charles, 2014).

The consequence of this increasingly common strategy is that labeling interfaces become more *conversational*. The system is providing implicit feedback to the user about the model under development, while the user's decisions on more difficult cases are implicitly challenging the system to update the model it holds. Such a perspective shifts from the relatively predictable view of machine learning, in which facts are simply repeated until the machine retains them, to one that more closely imitates teaching and learning situations between two humans, where the knowledge being acquired is dialogically shared, probed and questioned as in a conversation.

We report investigations into a particular characteristic of human conversation that has not previously been explored for its relevance to interactive machine learning - rhythmic timing in conversation. This can be contrasted on one hand to earlier models of "dialog" in HCI where the user takes the initiative, issuing commands while the system responds with information as soon as it can in order to be seen as "smooth" (Miller, 1968; Nielsen, 1993), or on the other hand to task optimization models where the system takes the initiative, prompting for information that is supplied by the user as quickly and efficiently as possible, with minimum latency (Bernstein, Brandt, Miller & Karger, 2011). In mixed initiative interaction (Horvitz, 1999), neither of these existing design models is appropriate, and we suggest that attention to rhythm and timing becomes far more important.

In design approaches where the system has the potential to complete the user's actions, or even to take the initiative and make decisions, there will be a back-and-forth flow of initiative between the user and the system, resembling participatory turn-taking in human conversation. In human conversation[1], poorly timed participation is associated with negative or inappropriate social effects (see, e.g., Richardson, Dale & Kirkham, 2006; Benus, Gravano & Hirschberg, 2011). We ask: *during mixed-initiative interaction*

[1] While the HCI field has explored conversational interaction design using Grice's Maxims for human-human conversation (Grice, 1975), existing research places more focus on the *logic* or *content* of the interaction (i.e. quality, quantity, relevance, and manner) (Brennan & Hulteen, 1995; Kehler, 2000). Our research is concerned with whether or not the *temporal coordination* (or the "turn-taking") aspect in interpersonal conversation can also inform HCI design.

such as interactive labeling, how do the rhythmic timing characteristics of the interaction influence users' experience, and how should these be designed?

After a review of relevant literature, the remainder of this paper presents two experiments. The first lays the ground for understanding user perceptions of rhythm when interacting with a system during controlled experimental tasks. The second applies those results to an intelligent labeling application, showing that the user perceptions resulting from changes in rhythm do persist in realistic task contexts. Finally, we show that rhythm which mimics human conversational patterns results in improved accuracy for labeling, by comparison to the currently standard design approach in which the user is free to set their own rhythm when making labeling decisions.

Literature review

Conversational labeling

Labeling lays the foundation for the supervised training of machine-learning based artificial intelligence (AI) algorithms (Brodley, Rebbapragada, Small & Wallace, 2012). The primary purpose of labeling is to construct a training dataset that exemplifies human subjective interpretation - considered to be the "ground truth" of human intelligence for AI tasks such as language interpretation, social judgements, creative expression or emotion classification. Based on these labeled datasets, AI classifiers emulate human intelligence and replicate human judgements (Ware, Frank, Holmes, Hall & Witten, 2001; Blackwell, 2015). Well-established research resources have been constructed this way. For instance, the ImageNet database offers "millions of cleanly sorted images" to train computer vision and pattern recognition algorithms (Deng, Dong, Socher, Li, Li & Li, 2009), while in more subjective tasks such as emotion recognition, human experts are recruited to label corpuses of naturalistic expressions in order to train affective computing systems that reflect human responses (Afzal & Robinson, 2014).

In business applications of machine learning classifiers, it is often the case that a high-quality labeled dataset is not available - for example, because the cost of labeling would be prohibitively expensive, or because a newly established business has not yet collected sufficient customer data for reliable training. Furthermore, while research into machine learning primarily aims to improve the proportion of correct classifications, commercial users of machine learning classifiers must often pay more attention to incorrect classifications, since these are the points at which software automation fails, and business processes must rely on manual correction to complete transactions. However, manual data labeling is tedious and repetitive, with humans often replicating trivial decisions that could have been automated, and not getting sufficient time to devote to cases where their judgment would have added more value.

These pragmatic considerations have led to the development of interactive machine learning systems, such as Microsoft Excel FlashFill, BrainCel, Gneiss, etc. (Gulwani, 2011; Sarkar, Jamnik, Blackwell & Spott, 2015; Chang & Myers, 2014), in which labeling is carried out "online", with feedback from the partially-trained classifier being used to inform the user about current performance and potential weaknesses of the statistical model so far, and users being provided with tools to correct the model (for an early paradigm of interactive model construction in the HCI literature, see Fails & Olsen Jr, 2003). This process of interactive labeling can be considered as a variety of programming by example, where the user causes the system to work as desired by demonstrating how it ought to behave (Lieberman, 2000; Menon, Tamuz, Gulwani, Lampson & Kalai, 2013). From this perspective, the process of observing system behavior and providing new labels to correct erroneous behavior is a kind of debugging (Kulesza, Burnett, Wong & Stumpf, 2015). Through direct interaction with data, it is likely that future semi-automated classification and inference systems will routinely demonstrate mixed-initiative interaction characteristics (Sarkar et al, 2015).

Throughout the development of interactive information systems, encompassing many kinds of user interface, the conventionally expected design goal has been for systems to respond as quickly as possible to user actions. This is ideal when the user is taking the initiative and issuing commands. A few researchers have also experimented with increasing productivity in labeling applications, by forcing the user to respond at a rhythm set by the system, faster than the user would work by preference. This typically leads to an increase in errors, due to speed-accuracy trade-offs, but potentially compensated by the overall improvement in efficiency that can be obtained by comparing a greater number of user judgments (Krishna, Hata, Chen, Kravitz, Shamma, Li & Bernstein, 2016).

However, in mixed initiative situations - including conversational approaches to intelligent labeling - the system has the potential to automatically complete the user's actions, take the initiative or make independent decisions. As a result, there will be a back-and-forth flow of initiative between the user and the system, resembling participatory turn taking in human conversation. In order to design such systems, it is necessary to have greater understanding of the temporal dynamics in such "conversation".

Rhythmic agency in mixed-initiative interaction

Traditionally, human-computer interaction design was based on the assumption that the user was the decision maker and the *initiator* of an action, and the computer system was the *responder*, executing the user's commands. Within this dynamic, the user typically expects that the *faster* the system responds, the better control he/she has, and indeed much engineering of traditional user interfaces has been dedicated to minimizing system response times, with consequent improvements in subjective user experience. However, recent developments in intelligent user interfaces, including interactive labeling systems

of the kind discussed in the previous section, aim to sometimes make decisions and take actions on the user's behalf, meaning that the nature of the interaction has changed to *mixed-initiative*. Here, both the user and the system can take the initiative, and the control is passed back and forth between the user and the system (Horvitz, 1999).

This new dynamic poses new challenges to interaction design. A key usability issue is how to "support the user's internal locus of control" (Shneiderman, 2010) during the handover of initiatives, especially when the system increasingly appears to have autonomy. Recent studies have described users' sense of control as the experience of agency (Haggard & Tsakiris, 2009), a concept derived from earlier philosophical and psychological theories (Bratman, 1999; McCann, 1998). By definition, the experience of agency arises when a person feels that he/she is in control of his/her actions, and is responsible for, or has the ownership of, the consequences those actions have caused in the external world (Coyle, Moore, Kristensson, Fletcher & Blackwell, 2012). HCI researchers have approached this topic from several angles, for example looking at the effect of adopting different input modalities (Coyle et al. 2012; Limerick, Moore & Coyle, 2015), and of improving output quality and feedback channels (Chafe, 1993; Farrer, Bouchereau, Jeannerod & Franck, 2008; Berthaut, Coyle, Moore & Limerick, 2015).

Timing of action and response is a property inherent in all kinds of interaction, with several implications for usability beyond the mundane question of minimizing system response time for optimal task duration (i.e. the system-rt factor in GOMS-style analysis (Gray, John & Atwood, 1993)).

Firstly, timing is a crucial cognitive factor when the user is building a "causal link" between an action and its consequence(s). Sense of agency relies implicitly on this causal link. However, this sense can easily be impaired if the consequence is presented with poor timing, for instance if it is too *early*, to an extent that it may appear to have causes that precede the action, or if it is too *late* to be linked and recognized as caused by that action (Wegner & Wheatley, 1999; Moore & Obhi, 2012).

Secondly, the timing of external stimuli can affect cognitive function. Neuropsychological studies provide evidence that rhythmic stimuli help the brain form temporal expectations, based on which target events can be selectively attended to. This can enhance signal detection whilst optimizing the allocation of cognitive resources. Previous research also suggests that a pre-occupied mind is less likely to feel "in control" because establishing causal links requires cognitive resources (Hon, Poh & Soon, 2013), meaning that predictable timing in the interaction can potentially allow us to feel more control. Thirdly, timing can influence interaction experience. Studies in social psychology and music performance have shown the beneficial aspects of rhythmic entrainment. Entrainment is when people adapt to each other's rhythm and eventually align their behaviors in time, just as in physical systems such as coupled pendulums that will gradually adjust themselves to fall into the same period or phase (Clayton, Sager & Will, 2005). Entrainment can improve the temporal predictability of an interaction, which allows people to build mutual trust and empathy and induces pro-

social behaviors (Knight, Spiro & Cross, 2017). Despite this previous research demonstrating the influence of rhythmic interaction in human cognition, rhythmic timing has not yet been considered as a design element when building mixed-initiative applications such as interactive labeling systems.

The following sections investigate the potential of timing as a design resource for interactive labeling through two empirical studies: one investigating the effect of alternative interaction rhythms on subjective factors in user experience, and the second investigating the consequence of applying the same rhythms in a practical design context for interactive labeling.

Experiment 1

The first phase in this research used established experimental paradigms to investigate whether interaction rhythm does have a reliably measurable effect on the user experience of agency as hypothesized. We summarize two experiments that were previously reported at the British HCI conference. The present paper summarizes those experiments, with new statistical analysis of the results. Readers interested in more details of the experimental procedure can refer to the previous publication (Yu & Blackwell, 2017).

This experiment is reported in two parts: the first exploring the effect of rhythm during on-screen interaction with *visual* prompts, and the second exploring interaction in response to *audio* stimuli. The same participants were used in both parts, which were administered consecutively in a single session for each participant. The use of audio stimuli in the second part, although less representative of user interface design, allows us to apply an experimental paradigm that was originally developed to investigate perception of agency in neuroscience research. As described below, this "Libet clock method" allows us to measure the phenomenon of "intentional binding," which creates an illusion resulting in time errors when reporting the time of events while attending to a moving clock.

In each part of the experiment, and throughout this paper, we report the effect on the user of four different experimental conditions, as follows: In condition Computer-sets-Rhythm (CR) the Computer takes the initiative using rhythmic - or "*periodic*[2]", to be precise - timing. In condition Computer-Arrhythmic (CA) the Computer takes the initiative using aperiodic time intervals. In condition User-sets-Rhythm (UR) the User takes the initiative, setting their own rhythm, typically approximating to a regular time period between clicks. In condition User-followed-by-Computer (UC) the User again

[2] Though rhythm is often used interchangeably with the term periodicity (Patel, 2010), "being periodic" is a stricter criterion than "being rhythmic": while any series of events in time can comprise a rhythm, only the series that has an underlying beat structure is periodic. In other words, all periodic processes are rhythmic, but not all rhythmic processes are periodic.

takes the initiative, and the Computer follows the period set by the user, imitating conversational entrainment between humans.

After each condition, participants completed the NASA-TLX scale (Hart & Staveland, 1988) to assess mental demand, physical demand, temporal demand, performance, effort and frustration on 6 separate 21-gradation sub-scales (5-point steps within 100 points) presented on the screen. We also asked participants to rate the following 5 items: a) "The software adapted to me" vs. "I adapted to the software"; b) "I was controlling the pace" vs. "The software was controlling the pace"; c) "The software intended to help me" vs. "The software intended to challenge me"; d) "I felt relaxed during this task" vs. "I felt stressed during this task"; and e) "I felt confident in my answers" vs. "I felt unconfident in my answers".

We recruited 22 participants from informal networks among staff and students of the University of Cambridge. A small gift was given in appreciation of their time. The experiment was reviewed by the ethics committee of the Cambridge University Computer Laboratory.

Experiment 1a: Visuo-spatial presentation task

Participants were told that this experiment would study "how people follow various sequences of events on a screen", not mentioning timing or rhythm. The procedure had three stages. In each stage, participants had to attend in a clockwise sequence to targets at four screen locations (upper left, upper right, lower right, lower left). In the *target stage*, a cross was shown in sequence at each of the four locations. In CA and CR conditions, the user simply observed the crosses appearing. In UR and UC conditions, the user clicked on each cross as it appeared, with the system monitoring the timing used. In the *presentation stage*, a different shape was shown in sequence at each of the four locations. In the CA, CR and UC conditions, the user simply observed the shapes. In the UR condition, the user clicked each shape as it appeared. In the *recall stage*, the user was asked to recall which shape had appeared in each location in turn: the system displayed the four possible shapes in an on-screen menu at each location, and the user had to click on the one they remembered being there. Before the three stages, the user carried out a simple practice task, clicking around the four locations for 30 rounds. During this practice task, the user was told to click at a rate they found comfortable, while the system recorded the average rate of clicking.

The rhythmic behavior of the system in the four conditions was varied as follows: In the CA condition, the time intervals between stimulus presentations were randomized. In the CR condition, all stimuli were presented periodically, appearing at regular intervals (using the "comfortable" rate observed during the practice task). In the UR condition, the user sets the rhythm (period) in all stages, by choosing the speed at which they click. In the UC condition, the system observes the period of the user clicks in the presentation stage, and then imitates the same period in the recall stage. The order of these four conditions was randomized across participants. In conditions CA, UC and

UR[3], the intervals between events are compared, testing for cross-correlation coefficients between the series of user click intervals and the series of visual stimuli presentation intervals. Cross-correlation is a measure for the "similarity of two interacting series as a function of the displacement of one relative to the other" (Boker, Rotondo, Xu & King, 2002), with its value ranging between 0 and 1. Increased value of the cross-correlation coefficients between user intervals and system intervals provides a measure of rhythmic entrainment between the two.

Experiment 1b: Auditory presentation task

Experiment 1b used the same set of experimental conditions (CA, CR, UR, UC), but with auditory rather than visual stimuli - a simple series of beeps. Participants were told that the purpose was to explore "how people follow various sequences of sounds from a computer". In place of the practice task, the user explicitly chose their preferred period by adjusting a slider, then clicking to confirm after hearing 16 beeps at that steady speed.

In all four conditions, the participant watched the rotating sweep hand on a "Libet clock" while listening to a series of beeps. When the beeps stopped, they reported the perceived clock time at the last beep. The series ended after (randomly) 7, 8, 9 or 10 beeps, so that the participant could not anticipate when the end would come. In the CA condition, the interval between the beeps was irregular. In the CR condition the beeps occurred at the preferred constant speed. In the UC condition, participants clicked a button to make the computer beep the first 4 times, after which the computer system continued to beep at the speed of their clicks. In the UR condition, participants continued clicking the button at their own pace throughout. The sequence of these four conditions was randomized across participants.

The Libet clock method relies on the "intentional binding" illusion that occurs naturally when people feel in control of their actions (also described as "sense of agency"), first applied in HCI research by Coyle et al. (2012). If a person believes they are in control of an action, this illusion causes them to perceive the outcome (of that action) as happening *earlier* than its actual time. Conversely, if a person does not assume control of an action, they perceive the outcome as happening *later* than its actual time. In other words, the perceived time interval between an action and the outcome is *shortened* when a person feels in control, and is *prolonged* when they do not. To measure this effect, we followed the standard approach in Libet clock experiments: the participant reported the position where they saw the rotating clock hand at the time of the last beep, and we calculated the error in this report that results

[3] In CR condition, the system intervals were strictly periodic, of which the standard deviation was 0, hence the cross-correlation formula (see Boker, Rotondo, Xu & King, 2002) was not applicable to CR condition.

from the illusion. The direction (e.g. positive / negative) and size of the error can implicitly measure participants' experience of being in (or out of) control and to what extent they feel so (Libet, Gleason, Wright & Pearl, 1983). Participants practiced each condition three times, then repeated the task 30 times. After each block, participants also provided the same subjective ratings as before.

It is worth noting that explicit measures (such as self-report) and implicit measures (such as intentional binding) of sense of agency (SoA) are often, though not always, congruent (Moore, Wegner & Haggard, 2009; Ebert & Wegner, 2010). One theory is that the sense of agency consists of heterogenous dimensions, and explicit and implicit measures to date offer different accounts of the sense of agency (Synofzik, Vosgerau & Newen, 2008): while implicit measures such as intentional binding may represent the feeling of agency (FoA), which is passive, primary and perceptual (Synofzik et al., 2008), explicit measures such as self-report represents the judgement of agency (JoA), which is an active and reflective attribution process on a conceptual level (Ebert & Wegner, 2010; Gallagher, 2007). Nevertheless, both FoA and JoA contribute to the overall SoA, and one may be more prominent than the other depending on the "context and task requirements" (Synofzik et al., 2008; Ebert & Wegner, 2010).

The sequence of the four experimental conditions (CA, CR, UR, UC) in both parts of the experiment were counterbalanced, though the two parts (1a, 1b) were not, which might have introduced a learning bias. However, this risk was partially mitigated given that the task design for each part of the experiment was sufficiently different and challenging (e.g. memorizing randomized visual prompts, attending to randomized beep sounds while observing a Libet clock), and each participant was given a short break and a new task brief before part two, asked to choose a preferred baseline rhythm (which would be used in CR, and to generate randomized CA intervals around it) to listen to using a slider, as well as to try 3 rounds in each condition as warm-up before proceeding to the formal tasks.

Results

Experiments 1a and 1b both compare measures in the four conditions {CA, CR, UR, UC}. In the following analyses, we first report an omnibus test to determine whether there is any statistically significant difference between the four conditions. If a significant difference is found, we then carry out contrast analysis (Rosenthal, Robert & Rosnow, 1985; Haans, 2018) to test whether the ordering of the measured values in the four conditions supports hypotheses in relation to sense of agency. Bonferroni adjustment is applied to these significance tests (i.e., where k tests are carried out, $\alpha = 0.05/k$). Not all measures follow a normal distribution - we report this for each measure (based on Shapiro-Wilk test for normality), and use a non-parametric test whenever the data is not normally distributed.

Results for Experiment 1a

Sense of control: Participants' subjective ratings of sense of control were not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant effect on participants' perceived sense of control ($\chi2 = 43.340$, df=3, $p < 0.001$). Contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable, approximately periodic, user-led rhythm did increase the user's sense of agency ($F(1, 21) = MS/MSe = 715322.227/14079.751 = 50.805$, $p < 0.001$).

Perceived task stress: Participants' responses on TLX sub-scales were not normally distributed. An omnibus Friedman test found that experimental condition had significant effects on perception of physical demand ($\chi2 = 12.277$, df=3, p=0.006), task success ($\chi2 = 13.206$, df=3, p=0.004) and effort ($\chi2 = 9.332$, df=3, p=0.025). Contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable and user-led rhythm was associated with perceived task success ($F(1,21) = MS/MSe = 721.636/109.160 = 6.611$, $p = 0.018$). In conditions UC and UR, the user is required to click more often, and contrast analysis confirmed that participants noticed this increased physical activity ($F(1,21) = MS/MSe = 1298.227/98.703 = 13.153$, $p = 0.002$). To explore perceived effort, we considered whether the UC condition would be perceived by the user as involving greater responsibility for establishing the dyadic rhythm. We therefore tested two contrasts: linearly reducing effort (CA > CR > UC > UR), and a cubic model reflecting greater perceived effort in UC. Contrast analysis (after Bonferroni correction) confirmed this second model ($F(1,21) = MS/MSe = 744.727 / 79.775 = 9.335$, p=0.006).

Rhythmic entrainment: The intervals in the CR condition do not vary over time, so cross-correlation coefficient is calculated only for conditions {CA, UC, UR}. These coefficients were normally distributed. An omnibus ANOVA confirmed that the experimental condition had a significant main effect ($F = MS/MSe = 0.733/0.023 = 31.630$, df=2, $p<0.001$). Entrainment results from two agents accommodating their rhythm to each other, so we expect to observe an entrainment effect in condition UC, where the computer follows the rhythm set by the user. To test whether cross-correlation is greater in UC than the other two conditions, we use a quadratic model (CA < UC > UR). Contrast analysis with a quadratic model confirmed that rhythmic entrainment does result in greater cross-correlation in the UC condition ($F(1,21) = MS/MSe = 15.989/0.203 = 78.717$, $p<0.001$).

Recall accuracy: The numbers of stimuli correctly recalled were not normally distributed. An omnibus Friedman test indicated a marginally significant main effect ($\chi2 = 8.497$, df=3, p=0.037) across the four conditions. However, despite participants' subjective impression that they had been more successful in the more rhythmic conditions, contrast analysis with a linear model (CA < CR < UC < UR) showed that influence of rhythm on recall performance, although having a tendency in the expected

direction, was not statistically significant (F(1, 21) =MS/MSe = 192.045/132.807 =1.446, p=0.243).

Results for Experiment 1b

Sense of control: Participants' subjective ratings of sense of control were not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant effect on participants' perceived sense of control ($\chi 2$ = 43.248, df=3, p<0.001). As in experiment 1a, contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable, approximately periodic and user-led rhythm did increase the user's sense of agency (F(1,21) =MS/MSe =1093146.183/18314.087 =59.689, p < 0.001).

Intentional binding: The estimation errors between outcome event and reported Libet clock readings were normally distributed. An omnibus ANOVA confirmed that the experimental condition had a significant main effect on intentional binding (F= MS/MSe = 46384.201/5062.114 =9.163, df=2, p<0.001). Our hypothesis was that the intentional binding effect would increase with more predictable and user-led rhythm. Contrast analysis using a linear model (CA < CR < UC < UR) confirms this effect (F(1,21) =MS/MSe =2311110.291/102964.616 = 22.446, p < 0.001).

Judgment of agency: It is not always the case that subjective judgement of agency (JoA), corresponds directly to the feeling of agency (FoA) as measured by the intentional binding effect (Synofzik et al., 2008). We tested this association using the Spearman Rank-order Correlation Test (noting that subjective reports were not normally distributed), finding a strong positive correlation (r=0.249, p=0.019). These findings indicate that temporal structure of an interaction can affect both FoA and JoA when users are forming experiences of agency (Wegner & Sparrow, 2004; Moore, Wegner & Haggard, 2009).

Perceived system adaptation: Participants' ratings of how adaptive and helpful the system was were not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant effect on participants' rating of both perceived adaptivity and perceived helpfulness of the system ($\chi 2$ =18.192, df=3, p<0.001; $\chi 2$ =18.269, df=3, p<0.001). In the UR condition, the computer does not need to adapt to the user, so we expect a quadratic trend (CA<CR<>UC>UR) in these two ratings. This was confirmed for both adaptivity (F(1,21) =MS/MSe = 11546.182/2311.229 = 4.996, p = 0.036) and helpfulness (F(1,21) =MS/MSe =15290.909/1321.385 =11.572, p = 0.003).

Perceived stress: As in experiment 1a, ratings on TLX sub-scales were not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant effect on perceived mental demand ($\chi 2$ = 9.690, df=3, p=0.021) and

perceived effort ($\chi 2$ =15.426, df=3, p=0.001). Contrast analysis with a linear model (CA > CR > UC > UR) confirmed that more predictable and user-led rhythm had resulted in less perceived mental demand (F(1,21) =MS/MSe =1891.636/135.827 = 13.927, p = 0.001) and task effort (F(1,21) =MS/MSe = 1106.182/80.087 = 13.812, p = 0.001).

This was further corroborated by participants' reported ratings of relaxation/stress. An omnibus Friedman test confirmed that experimental condition did have a significant effect on perceived relaxation/stress ($\chi 2$ =11.816, df=3, p=0.008). As in experiment 1a, contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable and user-led rhythm did reduce perceived stress (F(1,21) =MS/MSe = 67765.500/6682.071= 10.141, p = 0.004).

Confidence: participants' ratings on TLX perceived task successfulness subscale were not normally distributed. An omnibus Friedman test confirmed a significant effect ($\chi 2$ =12.672, df=3, p=0.005), and as in experiment 1a, contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable and user-led rhythm did make participants perceive the tasks as more successful (F(1,21) =MS/MSe = 1222.545/90.450 = 13.516, p = 0.001).

This was further corroborated by participants' reported ratings of confidence in their own answers, which were also not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant effect on this confidence ($\chi 2$ =18.722, df=3, p<0.001). Contrast analysis with a linear model (CA < CR < UC < UR) confirmed that more predictable and user-led rhythm did increase participants' confidence (F(1,21) =MS/MSe =67101.136/5652.374 = 11.871, p = 0.002).

Experiment 1 Summary

These preliminary experiments have demonstrated that manipulating rhythmic elements of interaction does influence the user's perception of that interaction in a controlled experimental context, including: the feelings of agency that support a sense of control; temporal entrainment that is typical of natural human conversation; and perceptions of effort, stress and task performance. We have demonstrated those effects using controlled experimental tasks designed specifically to measure them, but the next question is whether the same effects will be observed in a more realistic interactive application, especially an application of the type where machine learning or other AI methods might result in mixed initiative experiences such as conversational interaction. We therefore designed a second experiment, to test whether similar effects are observed during realistic interaction with an intelligent system.

Experiment 2

In this experiment we wished to investigate how the user's interaction behavior and sense of control would be affected by the rhythmic aspects of interaction in an AI-assisted labeling system. The design of the system was inspired by CODA (Blackwell et al, 2018), an open source web-based tool that was originally created to support researchers from Africa's Voices Foundation (AVF) in efficiently analyzing large numbers of short texts (> 250,000 text messages). Current applications of CODA include collating public understanding of the COVID-19 pandemic in Kenya, or understanding maternal health issues among displaced populations in Somalian refugee camps. AVF staff are typically translator/researchers who read messages in a local language (often informal or using hybrid slang), in order to categorize and review them thematically. CODA was designed to make this process more efficient by enabling a conversational interaction style.

When an AVF researcher uses CODA, text messages are initially presented in a black and white table. As the researcher labels each message with a thematic code, the color of the row changes to show which theme it has been labeled with. The table is progressively colored in as the researcher labels more messages (Blackwell et al., 2018). At the same time as CODA is populated with manual labels, natural language processing algorithms are able to offer semi-automated decision support: based on the labeling decisions already made by the researcher, CODA infers the potential label for unlabeled messages and pre-colors those rows. This system initiative is communicated by using different shades of the theme colors, where the level of color saturation corresponds to the level of the statistical confidence of the inferred labels. Researchers are able to structure their work for efficient construction of thematic categories, re-ordering the rows by theme (to either confirm or correct thematic groups) or by confidence, to address the most ambiguous cases or to provide training cases that will contribute to low-confidence regions of the language model.

The main purpose of AI-assisted labeling systems like CODA is to allow human experts to make the most efficient use of their valuable time, to "get the greatest benefit from their analytic decisions" (Blackwell, 2015), but the temporal aspects of such interaction have not been manipulated in previous design work, nor has the impact on users' sense of control been assessed. We therefore created an experimental labeling interface with presentation and visual style similar to CODA. Our motivation in using a design similar to an existing product was to provide improved external validity for the experimental study, while making a customized version of this tool meant that the labeling tasks could be constructed in a controlled manner in order to investigate the effects of timing on interaction. This approach is comparable to previous user-centered investigations of intelligent labeling systems, such as the "AutoCoder" UI that was created by Holliday et al. (2016) to investigate user trust in automated labels.

An imaginary task scenario (simplified by comparison to the work of AVF, and related to situations that would be familiar to experimental participants recruited in the UK rather than public health in Africa) was developed as follows:

> *An online shopping mall has a data center. Recently they developed a few machine learning algorithms, which can process customers' enquiry messages, and automatically label messages into several categories, such as 'delivery', 'exchange and return', 'membership' and so on.*
>
> *However the performance of those algorithms are quite poor at the moment, and the system often makes wrong judgements. Therefore they are now recruiting people to manually train the algorithms, to make them better.*
>
> *As one of the first steps, the data center wants to let the algorithms judge whether an enquiry message is about 'product delivery' or not.*

In order to minimize participant bias caused by experimental expectation, we introduced the experiment by telling participants that our goal was to *"study the efficiency and performance of different database algorithms developed for an online shopping mall data center, which will be trained during their interaction with users in order to achieve better sentence processing and automatic labeling"*. We informed participants that their task was to *"check the system's judgment on each message by clicking the 'Correct' or 'Wrong' button"*. We did not mention timing or rhythm in the briefing. We further explained that in this study they were only expected to distinguish whether a message was about *"product delivery"* or not, and we gave them a definition of product delivery and a list of relevant keywords.

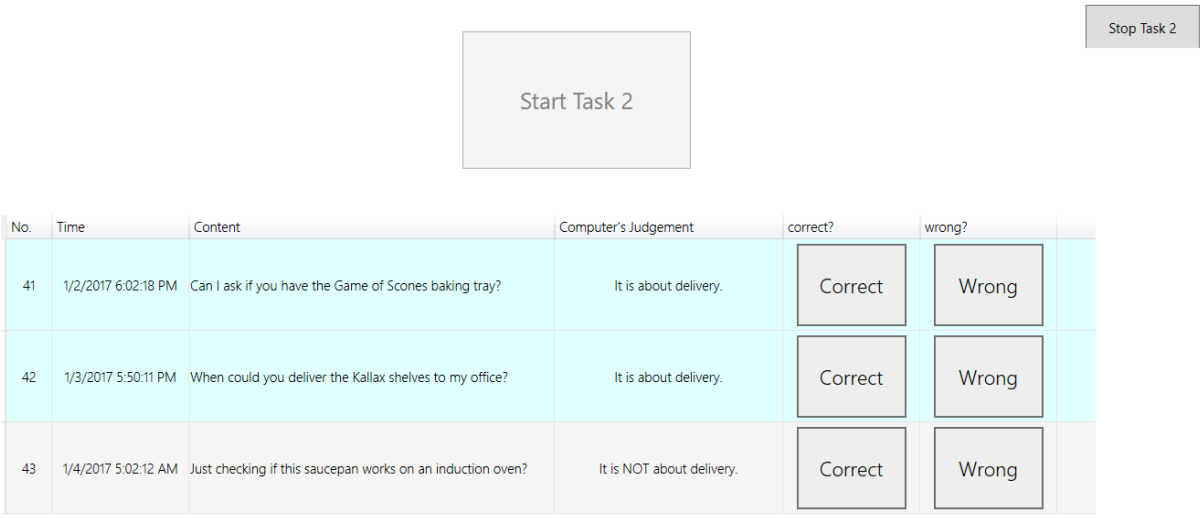

Figure 1 - the experimental labeling interface, in the style of the CODA system

As seen in Figure 1, each message is presented in a single row (together with a sequence number and time stamp – these are not used in the experiment, but are included to closely reflect the interface of the real CODA system whose appearance we are emulating). The row is colored either blue, if the system-proposed category says that the message is about delivery, or gray, if the system-proposed category says that the

message is not about delivery. The participant must read the text of the message, and decide whether the system-proposed category for this message is correct or wrong. The participant must click either the "Correct" button or the "Wrong" button, after which the cursor advances to the next message. New messages pushed by the system continually appear at the bottom of the table, at intervals as discussed below.

We recruited 15 participants from informal networks among staff and students of the University of Cambridge, all native English speakers (age $M = 26.4$, $\sigma = 5.81$). Each participant completed a practice stage, which gave them an overview of all the procedures, including practicing the basic operation by labeling 40 messages. After the practice tasks, they were asked to complete four blocks, each of which involved making decisions about 30 messages. After each block, participants reported sense of control and stress level using the same scales applied in Experiment 1.

For the 30 messages in each block, 20 were *not* about product delivery, while 10 *were* about product delivery. We then randomly selected 10 of the 20 non-delivery messages and deliberately labeled them incorrectly as "product delivery". Similarly, 5 of the 10 delivery messages were randomly labeled incorrectly as "not product delivery". The result is that in each block of 30 messages, 15 were labeled correctly, and 15 were labeled incorrectly. The composition of the test data items is summarized as a confusion matrix in Table 1. The participant's task is to correctly adjust these (initially balanced) proportions by fixing the labels. To reduce learning effects, we randomized the sequence in which messages were presented within each block, and also randomized the order of the three blocks for each participant. All 120 test messages were reviewed by a native English speaker in a pilot study, to confirm that our ground truth intention (delivery or non-delivery) had been correctly expressed in the text.

		Label presented by the system		
		Delivery	Not Delivery	Totals
Actual content of the text	Delivery	5	5	10
	Not Delivery	10	10	20
	Totals	15	15	30

Table 1 - confusion matrix of labels presented for 30 test data items in each block, of which 50% are incorrect

The experiment was reviewed and approved by the ethics committee of the Cambridge University Computer Laboratory. Participants were compensated with a gift voucher. After the session, participants were debriefed, telling them that in addition to their labeling results, we had also been interested in how the timing pattern of the system's actions had influenced their interaction behaviors and their subjective experience of

agency.

Independent variable and manipulation

There was one independent variable in Experiment 2: the rhythm of interaction between the user and computer when the messages are presented to be coded. In order to directly compare this more realistic mixed-initiative application to our earlier experimental tasks, we created versions of the application based on the same four timing variants used in experiment 1: an arrhythmic (aperiodic) condition (equivalent to the previous CA) where the computer takes the initiative, and messages presented for coding appear at irregular intervals; a condition in which the computer takes the initiative in presenting messages for coding, but following a predictable rhythm (equivalent to CR); a condition in which the user takes the initiative, with the computer following in alignment with that rhythm to present new messages (equivalent to UC); and a condition in which the user sets the rhythm at which messages are presented (equivalent to UR).

In the computer-initiated CA and CR conditions, participants clicked a "Start Task" button to begin, after which the system began pushing new messages onto the bottom of the table at intervals determined by the condition. The base interval length in the CR condition was a fixed value of 4.4s. This value was determined based on previous literature, indicating that an optimal line length for screen reading was 50-60 characters per line (cpl) (Dyson & Haselgrove, 2001), and that the effective reading rate on screen was around 150 words per minute (Muter & Maurutto, 1991). All of our test messages fell into the 50-60 cpl range and the average length was around 11 words, which should take a native English speaker roughly 4.4s to read and comprehend. We wished to place participants under a small degree of time pressure, so we reduced this estimated reading time by 10% to 4s. Previous research was also used to make an estimate of mouse selection time for large on-screen targets, calculated to be about 0.4s for the visual layout we used (Akamatsu & MacKenzie, 1996). Therefore altogether the rhythmic interval in Task 2 was set as 4.4s (4s to read + 0.4s to click). The random interval series in the CA condition were generated in MATLAB, with mean set as 4.4s and range linearly distributed with equal probability between 2.2s and 6.6s. Each adjacent pair of two intervals had at least a 500ms difference in duration, in order to ensure that participants would notice the random variations.

In the UC condition, the system would first push a message, then wait for the user to judge and label it. This first interval between the system's push and the user's response plus 500ms would be the interval between the 2nd and the 3rd system's push, and then the 2nd interval between the system's push and the user's response plus 500ms would be the interval between the 3rd and the 4th system's push, and so on. If the user sped up when labeling this message, their next message would be pushed earlier, and if the user slowed down, the next push would be delayed. Therefore the timing would be implicitly set by the user, but initiated by the system. The UR condition was considered as a control condition, corresponding to the conventional design of labeling systems

(including CODA), in which the user is able to dictate the pace and assume full control of all actions.

Dependent variables and measurements

The experimental system recorded timestamps of all participant mouse clicks, and all labeling decisions. In the results section we discuss the speed of labeling in terms of interval between the decision clicks and length of the accumulated queue. We also discuss accuracy of the participant decisions, in judging whether the presented label was correct or incorrect. We pay particular attention to the number of user decisions that incorrectly reject accurate labels offered by the system (i.e. "false positive" in error diagnosis) and the number of user decisions that incorrectly accept wrong labels offered by the system (i.e. "false negative" in error diagnosis).

As with Experiments 1a and 1b, participants were also asked for subjective ratings of their sense of control and stress level. As before, we used the NASA-TLX scales to assess participants' mental demand, physical demand, temporal demand, confidence in performance, perceived effort and frustration – each dimension as a 21-gradation sub-scale (5-point steps within 100 points) presented on the screen. We also asked participants to rate the same 5 items validated in Experiment 1: a) "The software adapted to me" vs. "I adapted to the software"; b) "I was controlling the pace" vs. "The software was controlling the pace"; c) "The software intended to help me" vs. "The software intended to challenge me"; d) "I felt relaxed during this task" vs. "I felt stressed during this task"; and e) "I felt confident in my answers" vs. "I felt unconfident in my answers".

Results analysis

The measures described above were analyzed using the same overall procedure as in Experiment 1. We performed an omnibus test, either repeated-measure one-way ANOVA or non-parametric Friedman Test (depending on whether the values of that measure were normally distributed), to test whether each measure did show significant variation across the four different conditions. If significant variation was observed, we carried out a planned contrast analysis (using SPSS, and following the procedures introduced in (Rosenthal et al., 1985) and (Haans, 2018)). We report for each hypothesis the contrast model, of which the assigned contrast weights and test procedure are described in (Rosenthal et al., 1985; Furr & Rosenthal, 2003). When multiple contrasts were tested against the same data, the alpha level for each F test was corrected using the Bonferroni method ($\alpha = 0.05/k$).

Test of workload equivalence: Before testing the effects of manipulating rhythm, we checked that the base workload (the rate at which the user was making coding decisions) was equivalent between conditions after order randomization, message

sequencing, and system adaptations to the user's speed. We therefore compared the average work rate (interval between coding actions) in the four conditions. After removing three outliers (based on quantile-quantile plot), the average interval between coding decisions was found to be between 4.102s and 4.455s, confirming our design estimate that it would take a native English speaker 4.4s to read and label one message. The intervals were normally distributed, but failed the Mauchly Test of Sphericity ($W=0.041$, $\chi2=27.942$, $DoF=5$, $p<0.001$), so the non-parametric Friedman Test was used to confirm that there was no significant overall effect of condition on task load ($\chi2 = 3.109$, $p = 0.375$).

Sense of control: In mixed-initiative tasks such as AI-assisted labeling, it is important for the user to feel that they are in control of the process, while still maintaining an effective rate at which the user provides labeling decisions to the system. In experiment 1b, we had been able to evaluate the sense of agency directly, by measuring the intentional binding illusion. In this more realistic mixed-initiative application, it is not possible to use the intentional binding paradigm, so we employed self-report to measure sense of control. We removed two invalid responses (one null response, and one selecting "1" or "100" on all items). The remaining data was not normally distributed. An omnibus Friedman test confirmed that experimental condition did have a significant main effect on reported sense of control ($\chi2 = 15.259$, $p=0.002$), with means as shown in Figure 2. Our expectation based on Experiment 1 was that users would feel a greater sense of control when the rhythm of the labeling activity followed their initiative - that having the system control the rhythm during AI-assisted labeling would impair the user's perceived control, while a user-led rhythm would preserve their perceived control. We performed contrast analysis with a linear model (CA < CR < UC < UR), which confirmed that user-led rhythm does promote greater sense of control ($F(1,13) = MS/MSe = 191178.286/5763.824 = 33.169$, $p < 0.001$).

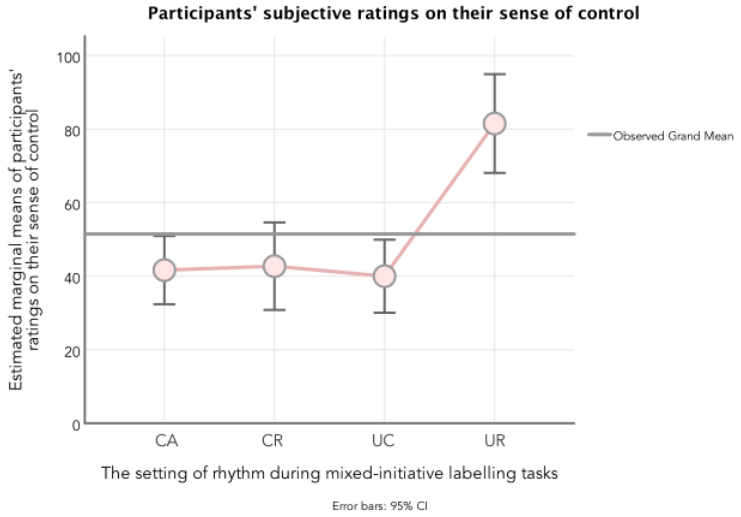

Figure 2 - sense of control (self-report) - marginal means for contrast analysis between four rhythm conditions

Perceived effort and stress: As described, the AI-assisted labeling interface was

designed to include an element of time pressure, in order to evaluate the effect of interaction rhythm during tasks that have appreciable user performance demands. We collected participant responses using six of the NASA TLX 21-gradation sub-scales. Ratings on the "mental demand", "temporal demand", and "effort" sub-scales were normally distributed, while "physical demand", "success", and "frustration" were not. Appropriate omnibus tests of these subscales found that the experimental condition had significant effects on sub-scales for "mental demand" ($F = MS/MSe = 11.929/3.262 = 3.657$, $df = 3$, $p = 0.020$), "temporal demand" ($F = MS/MSe = 79.732/7.771 = 3.657$, $df = 3$, $p < 0.001$) and "effort" ($F = MS/MSe = 12.810/3.438 = 3.657$, $df = 3$, $p = 0.019$), with means as shown in Figure 3. Our expectation was that users would perceive the task as less demanding when the system followed the user's rhythm. Contrast analysis with a linear model (CA > CR > UC > UR) confirmed that user-led rhythm is associated with both reduction in perceived mental demand ($F(1,13) = MS/MSe = 604.571/63.341 = 9.545$, $p = 0.009$) and reduction in perceived temporal demand ($F(1,13) = MS/MSe = 3363.500/122.423 = 27.474$, $p < 0.001$). In experiment 1, we had considered two models for perceived effort - a linear model that increased with more user-led rhythms (CA > CR > UC > UR), and also a cubic model in which UC would be perceived as more effortful because the user is taking responsibility for maintaining interaction rhythm (CA > CR > UC < UR). With Bonferroni correction for the two comparisons, contrast analysis does not support the first of these models ($F(1,13) = MS/MSe = 240.286/106.901 = 2.248$, $p = 0.158$), but does confirm the second model, in which the user setting a rhythm that is followed by the computer is perceived as more effortful than simply allowing the user to follow their own pace ($F(1,13) = MS/MSe = 528.286/80.132 = 6.593$, $p = 0.023$).

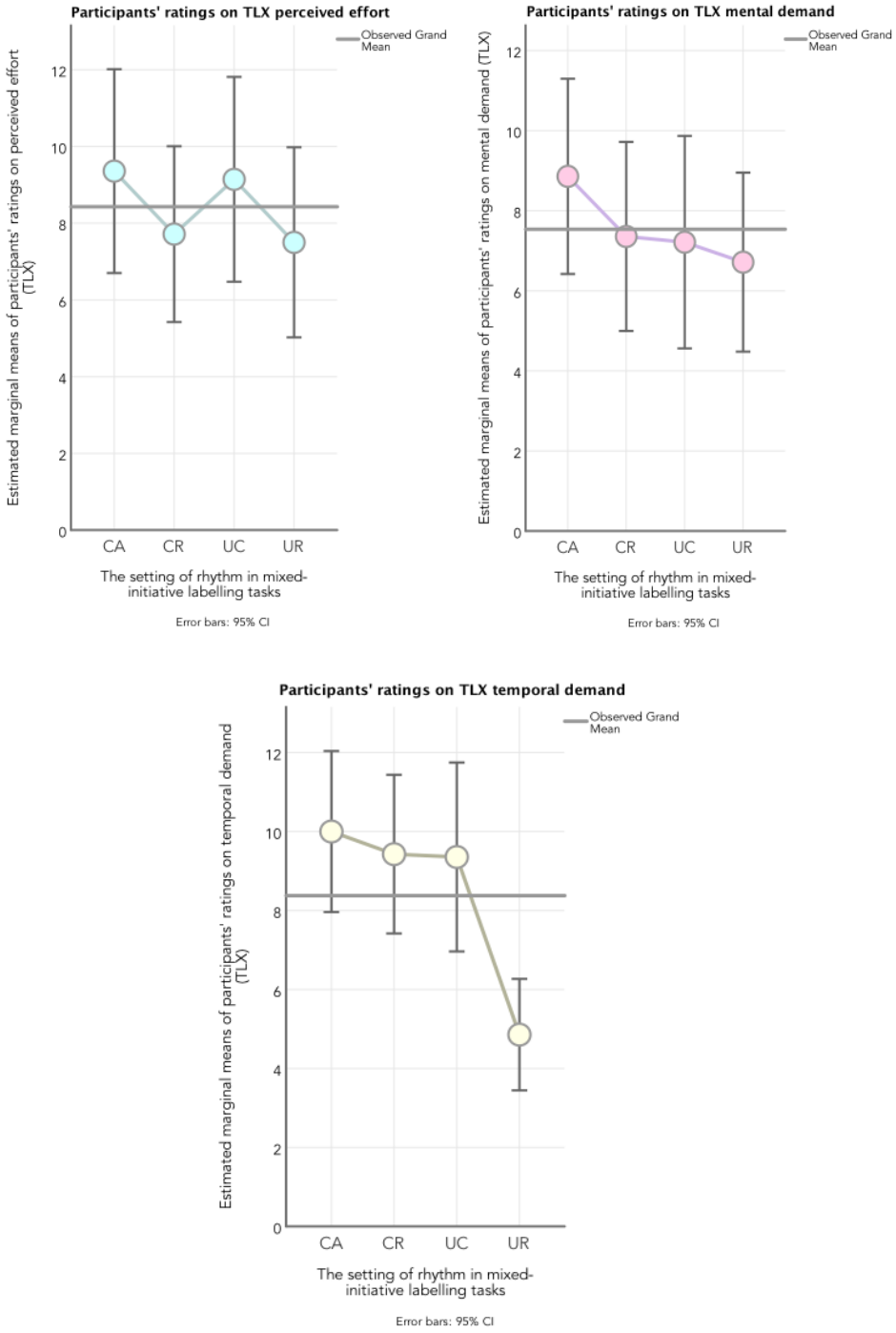

Figure 3 - TLX ratings for a) perceived effort, b) mental demand and c) temporal demand - marginal means for contrast analysis between four rhythm conditions

Perceived helpfulness: An important goal in mixed-initiative interaction is that users should perceive the system as helping them in their tasks. We evaluated this perception in the post-task questionnaire by asking participants to report whether "the system was helping me" vs. "the system was challenging me". Ratings on this scale were normally distributed after removing outliers as before. An omnibus ANOVA test confirmed that the experimental condition did have a significant main effect on this measure

(F=MS/MSe = 300.875/90.439 =3.327, df=3, p=0.029). As shown in Figure 4, participants were more likely to report that the AI-assisted labeling system was challenging them in the conditions when the computer imposed the labeling rhythm (CA and CR), and were more likely to report that the system was helping them in the conditions when they set the rhythm (UC and UR). Contrast analysis with a linear model for decreasing challenge / increasing helpfulness (CA > CR > UC > UR) confirmed that this tendency is statistically significant (F(1,13) =MS/MSe = 17643.500/2407.500 = 7.329, p = 0.018)

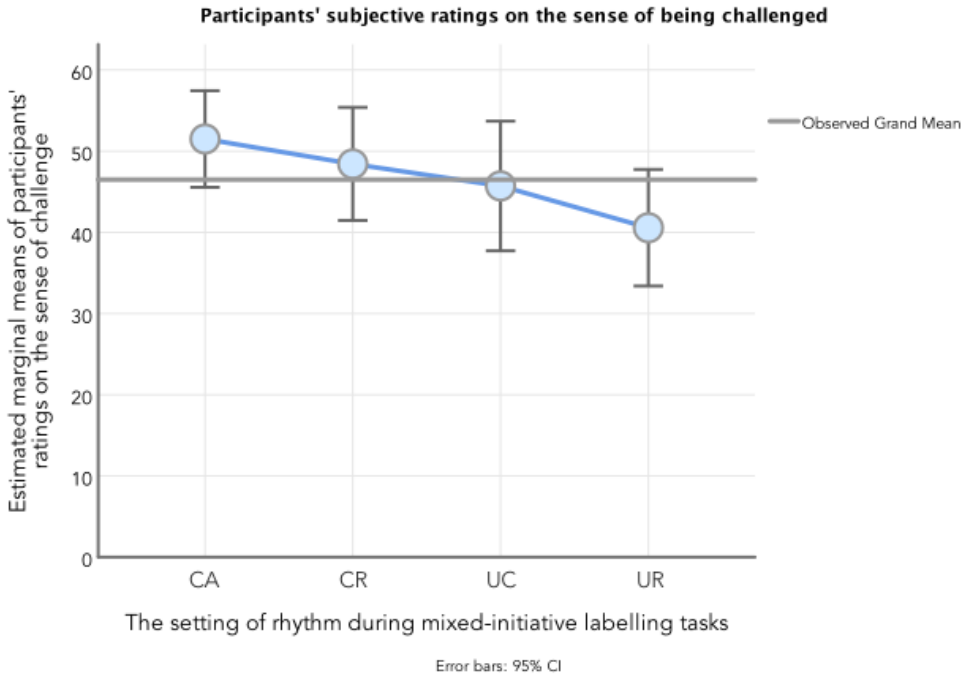

Figure 4 - subjective reports on scale from the system helping the user (0) to challenging the user (100) - marginal means for contrast analysis between four rhythm conditions

Accumulated task load: In many of the routine judgment tasks where mixed-initiative interaction techniques might be applied, task load can be measured in terms of the "backlog" of decisions that need to be taken by the human agent. In our design case of an AI-assisted labeling tool, the accumulated task load is directly visible as the length of the queue of messages that are waiting to be labeled. For each experimental block we calculated the average length of the queue over the duration of the block. This measure was not normally distributed. An omnibus Friedman Test showed that the experimental condition had a very significant main effect on average queue length ($\chi 2 = 25.039$, $p<0.001$). Our expectation was that the accumulated task load would be high in the two conditions where messages were pushed at a rate determined by the computer, and low in the two conditions where the user set the pace (CA ~= CR > UC ~= UR). However, contrast analysis with this model did not find a significant effect (F(1,10) =MS/MSe = 0.555/0.152 = 3.648, p = 0.085). Because the omnibus test had shown a highly

significant main effect, we performed further post-hoc analysis. As shown in Figure 5, the accumulated load was lowest in the UR condition, and highest in the UC condition. Contrast analysis with the model (UR < CA < CR < UC) shows that this linear trend is highly significant ($F(1,10) = MS/MSe = 33.005/4.939 = 66.831$, $p < 0.001$). We consider the implications of this observation further in the discussion section below.

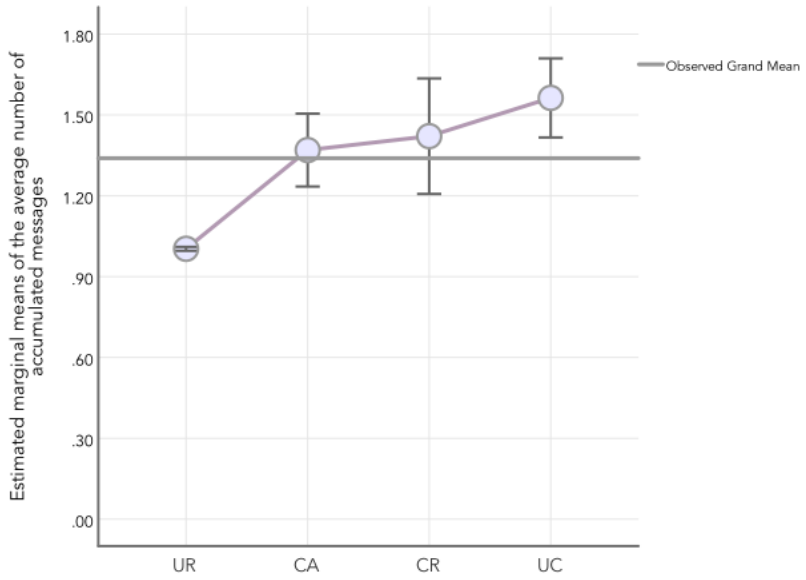

Figure 5 - accumulated task load of messages waiting to be labeled - marginal means for contrast analysis between four rhythm conditions

Accuracy: Recall that in each block of items presented to the participant, 50% of the labels have been designed to be correct, and 50% to be incorrect. The user task was to identify and reject the incorrect labels by clicking the "wrong" button, and to confirm the correct labels by clicking the "correct" button. To analyze accuracy, we counted the number of false negatives (where the user clicked the "correct" button when the label was actually incorrect) and the number of false positives (where the user clicked the "wrong" button when the label was actually correct). The overall numbers of true/false positives and true/false negatives, across all participants and all conditions, is summarized in Table 2.

		User judgement of the label		
		Wrong	Correct	Totals
Accuracy of system label	Wrong	857 (TP)	43 (FN)	900
	Correct	24 (FP)	876 (TN)	900
	Totals	881	919	1800

Table 2 - confusion matrix of labels presented for 30 test data items in each block

The overall error rate (FP + FN) is 67 out of 1800 decisions, or 3.72%. This means that on average, for each block of 30 trials (a single participant in a single experimental condition), participants typically made about one error - either a false negative or false positive. This average of 1 observation per condition does not allow us to characterize variance of the distribution or carry out ANOVA between the experimental conditions, but we include some descriptive statistics as a basis for further investigation (requiring substantially larger samples) in future.

Table 3 summarizes in a confusion matrix the proportion of errors observed in each of the four conditions, across all participants. This shows that the greatest number of false negatives (where the label is wrong, but the user says it is correct) is seen in the CA condition, while the greatest number of false positives (where the label is correct, but the user says it is wrong) is seen in the UR condition. Based on these observations, the large variation in the relative proportion of false positives and false negatives between the conditions is intriguing, and deserves future investigation - for example, to determine whether there might be differential levels of *agreement bias* between the conditions. However, a Chi-squared test shows that, while these overall differences in accuracy between conditions tend toward significance, no statistically significant effect can be confirmed across all four conditions ($p=0.11$).

		User judgement of the label	
		Wrong	Correct
Accuracy of system label	Wrong	211 / 215 / 216 / 215	14 / 10 / 9 / 10
	Correct	3 / 8 / 3 / 10	222 / 217 / 222 / 215
	Totals	214 / 223 / 219 / 225	236 / 227 / 231 / 225

Table 3 - confusion matrix with breakdown by condition CA / CR / UC / UR

As a measure of overall performance in a labeling task, we can use the Matthews Correlation Coefficient (MCC), which is commonly applied to machine learning classifiers. Possible values of MCC range from 1 (perfect correspondence) to -1 (no correspondence at all). We calculated MCC for each experimental block (30 trials, of which 15 had incorrect labels). We observed the best performance in the UC condition (MCC = 0.95) and the worst performance in the UR condition (MCC = 0.91), as shown in Figure 6. This difference in means is sufficiently large to suggest that the effect would be worth investigating in future, especially given the possibility that rhythmic entrainment in a conversational labeling interface may offer superior accuracy to conventional interfaces based on self-paced rhythm. However, a Chi-squared test again shows that, while differences between the four conditions tend toward significance, no statistically significant effect can be confirmed ($\chi 2 = 5.755$, N=15, df=3, p=0.124).

As discussed further below, the difference between UC and UR is particularly interesting, because UR represents current standard design practice, in which the rhythm of interaction is simply determined by the user, while UC reflects a potential conversational approach to labeling, in which the computer follows the rhythm set by the user in an analogy to conversational entrainment. The CA and CR conditions, in contrast, are experimental manipulations that have informed our theoretical agenda from the first experiment, but are not suggested as practical design approaches. We therefore carried out a post-hoc within-subjects comparison of the UC and UR conditions using the Wilcoxon Signed-Rank test, which confirmed that the difference between these two conditions is statistically significant (V = 42, p=0.023) with a large effect size ($r= |Z| / \sqrt{N} = 2.269/ \sqrt{15} = 0.59$) (Rosenthal, Cooper & Hedges, 1994; Maher, Markey & Ebert-May, 2013). Given that this was a post-hoc test, further experimental investigation is justified. Nevertheless, the finding suggests interesting implications for application of this work, as we discuss below.

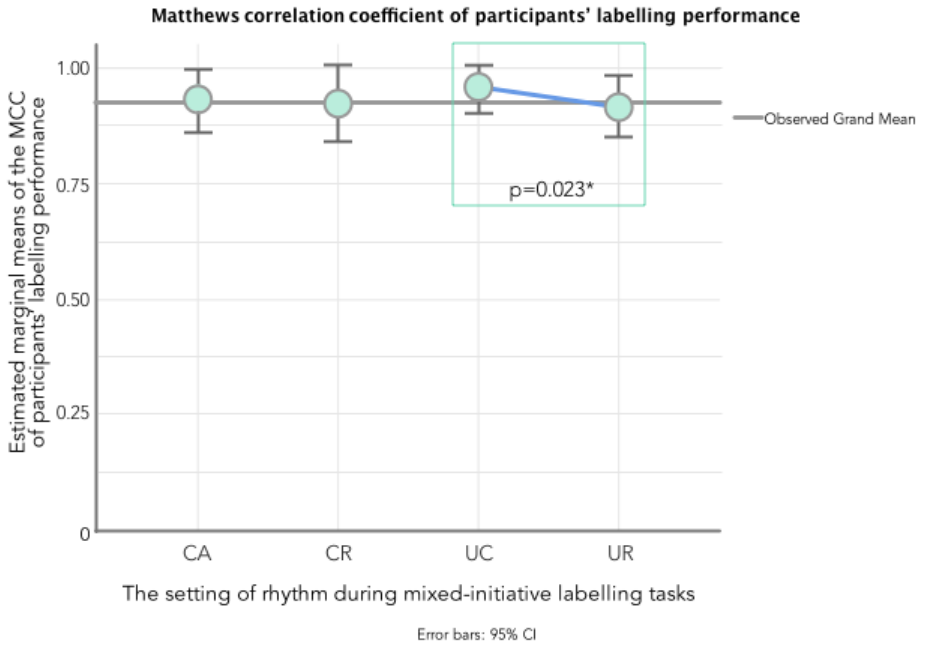

Figure 6 - Matthews Correlation Coefficient (MCC) - marginal means for four rhythm conditions, indicating post-hoc paired-sample comparison between UC and UR

Discussion

Our main focus in this research has been to understand the changes in user perceptions that result from introducing a conversational approach to the temporal design of mixed initiative interaction. Experiment 1 demonstrated that interaction following predictable rhythms, incorporating elements of human conversation such as rhythmic entrainment, is perceived as less stressful and effortful, providing greater sense of agency, locus of control and task confidence.

The tasks in Experiment 1 were based on previous research that measured sense of agency via the intentional binding illusion, drawing on previous suggestions that such methods could be used to assess the sense of agency in HCI, since sense of agency is a key element underlying the locus of control in interaction (Coyle et al., 2012). While our main focus in Experiment 1 was to demonstrate that such effects can be modified by manipulating the rhythmic properties of interaction, it is important to ask whether the effects of rhythm are purely subjective and attitudinal (properties which are of course valuable in themselves), or whether they are also associated with improvements in task performance in realistic mixed-initiative settings.

We therefore moved beyond controlled experiments that had been designed specifically to measure sense of agency, and proceeded to assess whether the same effects would be observed in the more realistic context of a prototype application whose design reflected that of an actual mixed initiative labeling application. This application-oriented experiment showed that the perceptual effects seen in laboratory-style tasks are

replicated in a realistic labeling application. Furthermore, the results of this experiment suggest that user accuracy in labeling decisions can be enhanced by temporal interaction design that imitates conversational entrainment between humans -- having the system follow the rhythm of actions made by the user. It is particularly interesting that, while users find it least stressful when they are able to fully control the rhythm of interaction by the timing of their own actions (the UR condition), labeling accuracy is not optimal in this more comfortable design approach. A conversational rhythm, where the timing of the actions initiated by the system is modified to follow rhythmic cues from the user's own actions, results in greater accuracy than self-paced interaction design where the user sets their own rhythm.

This has two important implications. The first is that subjective sense of agency, by itself, is not necessarily a good predictor or proxy for task performance. The self-paced UR condition resulted in the highest measured values for sense of control (in both Experiment 1 and Experiment 2), but the worst accuracy on an actual labeling task. The implications for design are firstly, that user comfort may not always be the ultimate goal of interaction design in mixed initiative systems, and secondly, to remind us that system evaluation must consider actual performance metrics, not simply subjective self-assessment by users.

The second implication is that the improvement in accuracy seen in the more conversational UC condition may reflect enhanced mutual partnership between user and system, drawing on the expectations of naturalistic dialog. It is interesting to consider what mechanism may be involved in the extremely poor performance of the UR condition. One possible explanation, for which further investigation would be beneficial, draws on expectation states theory (EST), which considers the position that a person holds within the "power-and-prestige" order of a group. Those in higher positions are likely to be more assertive, more critical of others' performance, give others fewer opportunities to speak, and attribute less credit to others' contribution. If self-paced interaction rhythm gives users the impression that they are in full control, and hence more "powerful" than the system in terms of EST theory, this would result in them being more critical of the system's suggestions, and giving less credit to the system's contribution (Bonito, Burgoon & Bengtsson, 1999; Fişek, Berger & Norman, 1995). This is indeed the pattern observed in the confusion matrix for the self-paced UR condition, which resulted in a far higher proportion of False Positives (where the system is right, but the user says it is wrong) by comparison to the more conversational UC condition. It is also notable that the accumulated task load is lowest in the UR condition, indicating that decisions are being made quickly in order to reduce queue length, giving the system less "opportunity to speak", while the accumulated load is highest in the UC condition, indicating that participants spent more time considering the system contributions as a conversational partner.

Further investigation of this conversational design strategy seems likely to offer a valuable resource for future mixed initiative systems, where adapting system behavior to follow the rhythmic conventions of conversation results in an effective combination

that is both comfortable and empowering in terms of sense of agency, while also offering the optimum level of human judgment as a contribution to overall system accuracy in mixed initiative tasks.

Limitations

There are limitations in these experiments that should be addressed in future research. The most important is that our experimental designs always using four different rhythmic conditions CA / CR / UC / UR, while providing useful control conditions to validate the effects of interaction rhythm, lead to complex and time-consuming experimental designs with loss of statistical power. Having demonstrated that the basic effects of conversational interaction can be reliably replicated, we suggest that further investigation should focus only on comparison of the self-paced UR interaction style (which is effectively the design approach used in all mixed-initiative labeling systems currently in use) to the novel UC interaction style, which employs conversational entrainment to create interaction that appears optimum in terms of labeling performance, while also offering comfort and sense of control to the user. The CA and CR conditions, on the other hand, are not representative of any realistic system design (other than unintentionally, in the rather perversely arrhythmic CA case). While useful as an experimental control for theoretical comparison, it is difficult to interpret the implications of these two conditions for design applications, making them less valuable in an HCI context.

We also note that the limitations arising from this experimental complexity in Experiment 2 leave room to question the statistical strength of the findings. The relatively low frequency of observed errors, combined with expected variation within the four different conditions, meant that omnibus tests of labeling accuracy had only marginal levels of statistical significance. The paired-sample comparison of the conventional UR and novel UC design approaches, on the other hand, did find that this difference was significant, even though only explored as a post-hoc comparison. We have a plausible (also post-hoc) theoretical explanation for this effect, and this explanation is further validated by comparison of agreement bias as revealed in relative proportions of false negative and false positive decisions by the user. If repeating this experiment, we would certainly design it to test these effects as our central research question, especially considering their practical relevance to the design of conversational labeling interfaces.

Ethical questions in conversational labeling

Research like this, which blurs the boundaries between human and machine behavior, often raises longer-term ethical considerations that we would like to draw attention to.

Firstly, it is important to ask whether the intentional mimicking of human social behavior is intended to establish any kind of moral equivalence between human agents

and interactive systems (especially where those systems might be deployed by organizations whose own moral purposes could be mixed or unclear). Our own goal has been to create interaction dynamics that are comfortable and productive for the user, and we strongly recommend this as the priority for other such research in future. There is a further danger, as seen in some kinds of social media, that routine adoption of particular interpersonal cues in technical contexts might devalue those cues, damaging social cohesion as a result. User experience researchers must be alert to possible social harms arising from technology adoption.

Secondly, a possible future extension of this approach might be to further emulate human conversational behavior, in order to deceive system users, and have them believe they are interacting with another person when they are not. Deception of this kind would obviously be unethical, not only for the falsehood itself, but also for creating a situation where a person invests emotional resources into interaction that appears to be reciprocal and mutually affiliative, but is actually a form of dominance, controlled by hidden actors (cf Rule 4 of Boden *et al.*, 2017). Sadly, there are many circumstances in which it can be tempting for businesses to substitute automated systems for human staff as a cost-saving measure. The ethics of such situations are already problematic for our societies, and subject to open debate. Unethical deception in combination with those business practices might (adversely) clarify their moral status, but could also be used to avoid the obligations of scrutiny according to accepted social norms.

Finally, the widespread use of manual labeling as the unseen human labor underlying contemporary AI systems is associated with numerous ethical risks. The development of surveillance capitalism, in which users are compelled to submit to extractive terms of use in exchange for “free” services, has until now been unregulated and only occasionally subject to critical attention (Zuboff, 2019). The labor dynamics of cognitive capitalism as a basis for the future of work are also problematic, raising serious questions about underlying transgressions of human rights, and exploitative terms of employment (Irani & Silberman, 2013; Blackwell, 2019). If such labor is recast as a “conversation”, it is important to acknowledge that this conversation would be taking place within a context where there is an extreme imbalance of economic and contractual power between the conversational “partners". It would be ethically wrong to use the *illusion* of agency as a design trick in order to disguise lack of *actual* agency.

Conclusion

The principles of mixed-initiative interaction are now well understood. In designing intelligent interactive systems, it is important to understand the dynamic of back-and-forth contributions between system and user. However, in human-to-human conversation (as in music), the temporal properties of the back-and-forth exchanges are also important (see Richardson, Dale & Kirkham, 2006; Hawkins, 2014). If we play the “notes” in the right order, but without attending to the rhythm, it will not be the same

conversation. The studies that we have reported investigate the ways in which these principles might apply to HCI. In the two studies of the first experiment, we explored the ways in which varying the rhythm of interaction influences the user's perception of their relationship with the system, including task stress, confidence, and most importantly the sense of agency in which they perceive themselves as being able to control and contribute to the system behavior. These user perceptions are directly related to deep questions around the future influence of intelligent systems in our lives, suggesting that design for agency and control will be an important priority in many domains.

In our second experiment, we explored whether the user experiences that are associated with rhythmic action in a purely experimental task would also be preserved in a realistic application scenario having genuine conversational elements. In intelligent labeling systems, it is a fundamental requirement that the user should be able to contribute actively (as a source of human judgments and behavioral data), and also that they should be able to perceive, respond to and correct the machine learning models being constructed by the system. We therefore simulated an intelligent labeling interface, following the design of an actual deployed product, that implemented a range of different interaction rhythms. Evaluation of this interface demonstrated that the observed effects of rhythm on user stress and confidence are preserved in a realistic task context. Although our main focus was on showing that these effects can potentially be replicated in mixed initiative applications, we also observed a substantial impact on labeling accuracy. When our novel implementation of labeling interaction using a conversational rhythm is compared to the currently established design practice in which the user determines the rhythm without any kind of conversational entrainment or other rhythmic response from the system, we found evidence that users may take more care in their judgments, while still finding the interaction comfortable. This requires further investigation in design-oriented research, with closer calibration of accuracy measurement over a wider range of tasks and labeling duration.

Overall, this research demonstrates the opportunity for mixed initiative interaction design to become more acceptable to users, and also more effective, through design elements that draw on understanding of rhythmic entrainment in human-to-human conversation. We expect that as a result, interactive rhythmic agency will become an important element of more conversational intelligent labeling systems, and potentially also in other domains.

Acknowledgments

We would like to thank participants in the experiments described. This work was supported by the Cambridge Commonwealth European and International Trust Scholarship, and the student research grant from the Department of Computer Science and Technology, University of Cambridge.

Declaration of interest statement

The authors have no financial interest or benefit to declare.

References

Afzal, S., & Robinson, P. (2014). Emotion data collection and its implications for affective computing. In R. A. Calvo, S. K. D'Mello, J. Gratch, & A. Kappas (Eds.), *The Oxford handbook of affective computing* (pp. 359–369). Oxford University Press.

Akamatsu, M., & MacKenzie, I. S. (1996). Movement characteristics using a mouse with tactile and force feedback. *International Journal of Human-Computer Studies*, 45(4), 483–493. doi: 10.1006/ijhc.1996.0063

Benus, S., Gravano, A., & Hirschberg, J. (2011). Pragmatic aspects of temporal accommodation in turn-taking. *Journal of Pragmatics, 43*(12), 3001-3027. doi: 10.1016/j.pragma.2011.05.011

Bernstein, M. S., Brandt, J., Miller, R. C., & Karger, D. R. (2011). Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST'11)* (pp. 33-42). doi: 10.1145/2047196.2047201

Berthaut, F., Coyle, D., Moore, J. W., & Limerick, H. (2015). Liveness through the lens of agency and causality. *Proceedings of the 15th International Conference on New Interfaces for Musical Expression (NIME'15)* (pp. 76–79). Retrieved from http://hdl.handle.net/10197/6635

Binns, R., Veale, M., Van Kleek, M., & Shadbolt, N. (2017, September). Like trainer, like bot? Inheritance of bias in algorithmic content moderation. *Proceedings of the International Conference on Social Informatics* (pp. 405-415).

Blackwell, A. F. (2019). Artificial intelligence and the abstraction of cognitive labour. In M. Davis (Ed.), *Marx200: The significance of Marxism in the 21st century*. (pp. 59-68) Praxis Press.

Blackwell, A. F. (2015). Interacting with an inferred world: the challenge of machine learning for humane computer interaction. *Proceedings of the 5th Decennial Aarhus Conference on Critical Alternatives (AA'15)* (pp. 169–180). doi: 10.7146/aahcc.v1i1.21197

Blackwell, A. F., Church, L., Hales, I., Jones, M., Mahmoudi, M., Marasoiu, M., Meakins, S., Nauck, D., Prince, K., Semrov, A., Simpson, A., Spott, M., Vuylsteke, A. and Wang, X. (2018). Computer says 'don't know' - interacting visually with

incomplete AI models. In S. Tanimoto, S. Fan, A. Ko and D. Locksa (Eds), *Proceedings of the Workshop on Designing Technologies to Support Human Problem Solving*. University of Washington. (pp. 5-14).

Boden, M., Bryson, J., Caldwell, D., Dautenhahn, K., Edwards, L., Kember, S., et al. (2017). Principles of robotics: regulating robots in the real world. *Connection Science*, 29(2), 124-129. doi: 10.1080/09540091.2016.1271400

Boker, S. M., Rotondo, J. L., Xu, M., & King, K. (2002). Windowed cross- correlation and peak picking for the analysis of variability in the association between behavioral time series. *Psychological methods*, 7(3), 338-355. doi: 10.1037/1082-989X.7.3.338

Bonito, J. A., Burgoon, J. K., & Bengtsson, B. (1999). The role of expectations in human-computer interaction. *Proceedings of the international ACM SIG-GROUP Conference on Supporting Group Work (GROUP'99)* (pp. 229–238). ACM Press. doi: 10.1145/320297.320324

Bratman, M. (1999). *Faces of intention: Selected essays on intention and agency*. Cambridge University Press.

Brennan, S. E., & Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-based systems*, 8(2-3), 143-151. doi: 10.1016/0950-7051(95)98376-H

Brodley, C. E., Rebbapragada, U., Small, K., & Wallace, B. (2012). Challenges and opportunities in applied machine learning. *AI Magazine*, 33(1), 11–24. doi: 10.1609/aimag.v33i1.2367

Chafe, C. (1993). Tactile audio feedback. *Proceedings of the International Computer Music Conference (ICMC)* (pp. 76–79).

Chang, K. S.-P., & Myers, B. A. (2014). Creating interactive web data applications with spreadsheets. *Proceedings of the 27th annual ACM Symposium on User Interface Software and Technology (UIST'14)* (pp. 87–96). doi: 10.1145/2642918.2647371

Clayton, M., Sager, R., & Will, U. (2005). In time with the music: the concept of entrainment and its significance for ethnomusicology. *Proceedings European meetings in Ethnomusicology* (Vol. 11, pp. 1–82). Retrieved from http://oro.open.ac.uk/2661/1/InTimeWithTheMusic.pdf

Coyle, D., Moore, J. W., Kristensson, P. O., Fletcher, P., & Blackwell, A. F. (2012). I did that! measuring users' experience of agency in their own actions. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'12)* (pp. 2025–2034). doi: 10.1145/2207676.2208350

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & L., F.-F. (2009). ImageNet: A large-scale hierarchical image database. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'09)*. (pp. 248–255). doi: 10.1109/CVPR.2009.5206848

Dyson, M. C., & Haselgrove, M. (2001). The influence of reading speed and line length on the effectiveness of reading from screen. *International Journal of Human-Computer Studies*, 54(4), 585–612. doi: 10.1006/ijhc.2001.0458

Ebert, J. P., & Wegner, D. M. (2010). Time warp: Authorship shapes the perceived timing of actions and events. *Consciousness and cognition*, 19(1), 481–489. doi: 10.1016/j.concog.2009.10.002

Fails, J. A., & Olsen Jr, D. R. (2003). Interactive machine learning. *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI'03)* (pp. 39–45). doi: 10.1145/604045.604056

Farrer, C., Bouchereau, M., Jeannerod, M., & Franck, N. (2008). Effect of distorted visual feedback on the sense of agency. *Behavioural Neurology*, 19(1, 2), 53–57. doi: 10.1155/2008/425267

Fişek, M. H., Berger, J., & Norman, R. Z. (1995). Evaluations and the formation of expectations. *American Journal of Sociology*, 101(3), 721–746. doi: 10.1086/230758

Furr, R. M., & Rosenthal, R. (2003). Evaluating theories efficiently: The nuts and bolts of contrast analysis. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2 (1), 33–67. doi: 10.1207/S15328031US020103

Gallagher, S. (2007). Sense of agency and higher-order cognition: Levels of explanation for schizophrenia. *Cognitive Semiotics*, 1, 33–48.

Gray, W. D., John, B. E., & Atwood, M. E. (1993). Project Ernestine: Validating a GOMS analysis for predicting and explaining real-world task performance. *Human-Computer Interaction*, 8(3), 237-309.

Grice, H. P. (1975). Logic and conversation. In *Speech Acts* (pp. 41-58). Brill. doi: 10.1163/9789004368811_003

Gulwani, S. (2011). Automating string processing in spreadsheets using input-output examples. *ACM Sigplan Notices*, 46(1), 317-330.

Haans, A. (2018). Contrast analysis: A tutorial. *Practical Assessment Research & Evaluation*, 23(9), 1–21.

Haggard, P. & Tsakiris, M. (2009). The experience of agency: Feelings, judgments, and responsibility. *Current Directions in Psychological Science* 18(4), 242–246.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology* 52, 139-183.

Holliday, D., Wilson, S., & Stumpf, S. (2016). User trust in intelligent systems: A journey over time. *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI'16),* (pp. 164-168).

Hon, N., Poh, J.-H., & Soon, C.-S. (2013). Preoccupied minds feel less control: Sense of agency is modulated by cognitive load. *Consciousness and Cognition*, 22(2), 556–561. doi: 10.1016/j.concog.2013.03.004

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'99)* (pp. 159–166). http://doi.org/10.1145/302979.303030

Irani, L.C., & Silberman, M.S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'13)* (pp. 611–620).

Kehler, A. (2000). Cognitive status and form of reference in multimodal human-computer interaction. In *Proceedings of the 17th National Conference on Artificial Intelligence* (*AAAI/IAAI'00)* (pp. 685-690).

Knight, S., Spiro, N., & Cross, I. (2017). Look, listen and learn: Exploring effects of passive entrainment on social judgements of observed others. *Psychology of Music*, 45(1), 99-115. doi: 10.1177/0305735616648008

Krishna, R. A., Hata, K., Chen, S., Kravitz, J., Shamma, D. A., Li, F.-F., & Bernstein, M. S. (2016). Embracing Error to Enable Rapid Crowdsourcing. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI'16)* (pp. 3167–3179). New York, New York, USA: ACM Press.

Kulesza, T., Amershi, S., Caruana, R., Fisher, D., & Charles, D. (2014). Structured labeling for facilitating concept evolution in machine learning. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems (CHI'14)* (pp. 3075–3084). New York, New York, USA: ACM Press. doi: 10.1145/2556288.2557238

Kulesza, T., Burnett, M., Wong, W.-K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI'15)* (pp. 126–137). ACM Press. doi: 10.1145/2678025.2701399

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential) the unconscious initiation of a freely voluntary act. *Brain*, 106(3), 623–642. doi: 10.1093/brain/106.3.623

Lieberman, H. (2000). *Your wish is my command: Giving users the power to instruct their software*. Morgan Kaufmann.

Limerick, H., Moore, J. W., & Coyle, D. (2015). Empirical evidence for a diminished sense of agency in speech interfaces. *Proceedings of the 33rd annual ACM conference on human factors in computing systems (CHI'15)* (pp. 3967–3970). ACM Press. doi: 10.1145/2702123.2702379

Maher, J. M., Markey, J. C., & Ebert-May, D. (2013). The other half of the story: effect size analysis in quantitative research. *CBE—Life Sciences Education*, 12(3), 345-351. doi: 10.1187/cbe.13-04-0082

McCann, H. (1998). *The works of agency: On human action, will, and freedom*. Ithaca, NY, USA: Cornell University Press.

Menon, A., Tamuz, O., Gulwani, S., Lampson, B., & Kalai, A. (2013). A machine learning framework for programming by example. *Proceedings International Conference on Machine Learning (ICML)* (pp. 187-195).

Miller, R. B. (1968). Response time in man-machine conversational transactions. In: *Proceedings of AFIPS Fall Joint Computer conference* (Atlantic City, NJ, 30 April – 2 May), New York: ACM, vol. 33, (pp. 267–277). doi: 10.1145/1476589.1476628

Moore, J. W., Wegner, D. M., & Haggard, P. (2009). Modulating the sense of agency with external cues. *Consciousness and cognition*, 18(4), 1056–1064. doi: 10.1016/j.concog.2009.05.004

Moore, J. W., & Obhi, S. S. (2012). Intentional binding and the sense of agency: a review. *Consciousness and Cognition*, 21(1), 546–561. doi: 10.1016/j.concog.2011.12.002

Muter, P., & Maurutto, P. (1991). Reading and skimming from computer screens and books: The paperless office revisited. *Behaviour & Information Technology, 10*(4), 257–266. doi: 10.1080/01449299108924288

Nielsen J. (1993). *Usability Engineering*. San Francisco, CA: Morgan Kaufmann.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2006). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407-413. Doi: 10.1111/j.1467-9280.2007.01914.x.

Rosenthal, R., Robert, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance*. Cambridge University Press.

Rosenthal, R., Cooper, H., & Hedges, L. (1994). Parametric measures of effect size. *The handbook of research synthesis,* 621(2), 231-244.

Patel, A. D. (2010). *Music, language, and the brain*. Oxford University Press.

Sarkar, A., Jamnik, M., Blackwell, A. F., & Spott, M. (2015). Interactive visual machine learning in spreadsheets. *Proceedings of IEEE symposium on Visual Languages and Human-Centric Computing (VL/HCC 2015)* (pp. 159–163). doi: 10.1109/VLHCC.2015.7357211

Shneiderman, B. (2010). *Designing the user interface: strategies for effective human-computer interaction*. Pearson Education India.

Synofzik, M., Vosgerau, G., & Newen, A. (2008). Beyond the comparator model: a multifactorial two-step account of agency. *Consciousness and Cognition, 17*(1), 219–239. doi: 10.1016/j.concog.2007.03.010

Ware, M., Frank, E., Holmes, G., Hall, M., & Witten, I. H. (2001). Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies, 55*(3), 281–292. doi: 10.1006/ijhc.2001.0499

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54*(7), 480–492. doi: 10.1037/0003-066X.54.7.480

Yu, G., & Blackwell, A. F. (2017). Effects of timing on users' agency during mixed-initiative interaction. *Proceedings of the 31st British Human Computer Interaction Conference (BHCI'17)* (pp. 1–12). doi: 10.14236/ewic/HCI2017.35.

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Profile Books.