Automation Bias with a Conversational Interface:

User confirmation of misparsed information

Erin Zaroukian & Jonathan Z. Bakdash
Human Research and Engineering Directorate
US Army Research Laboratory
Aberdeen Proving Ground, USA
erin.g.zaroukian.ctr@mail.mil
jonathan.z.bakdash.civ@mail.mil

Alun Preece & Will Webberley
Crime & Security Research Institute
Cardiff University
Cardiff, UK
PreeceAD@cardiff.ac.uk
WebberleyWM@cardiff.ac.uk

Abstract—**We investigate automation bias for confirming erroneous information with a conversational interface. Participants in our studies used a conversational interface to report information in a simulated intelligence, surveillance, and reconnaissance (ISR) task. In the task, for flexibility and ease of use, participants reported information to the conversational agent in natural language. Then, the conversational agent interpreted the user's reports in a human- and machine-readable language. Next, participants could accept or reject the agent's interpretation. Misparses occur when the agent incorrectly interprets the report and the user erroneously accepts it. We hypothesize that the misparses naturally occur in the experiment due to automation bias and complacency because the agent interpretation was generally correct (92%). These errors indicate some users were unable to maintain situation awareness using the conversational interface. Our results illustrate concerns for implementing a flexible conversational interface in safety critical environments (e.g., military, emergency operations).**

Keywords—automation bias; complacency; conversational interface; human-machine interaction; controlled natural language

I. Introduction

Conversational interfaces (e.g., Apple's Siri, Amazon's Alexa, Microsoft's Cortana, Google Now, and Tecent's WeChat) are growing rapidly and are being used for a variety of tasks, such as making payments, scheduling and reminders, and searching [1]. For conversational interfaces to allow novel interactions, they must be flexible enough to allow for user-driven input, such as new concepts and terms.

Using a prototype interactive conversational interface called Mobile Information Reporting App (MORIA)[1], we investigate user acceptance of computational misparses. Misparses occurred when users confirmed reports that were interpreted *incorrectly* by the conversational agent. First, a user's report is interpreted by the agent. Then, if the user confirms the agent's interpretation, it is then added to a knowledge base. We hypothesize that misparses result from automation bias and complacency [2], [3], a lack of situation awareness. Misparses were naturally occurring and a subset of data from simulated intelligence, surveillance, and reconnaissance (ISR) tasks in the real world collected using the SHERLOCK (Simple Human Experiments Regarding Locally Observed Collective Knowledge) platform [4]–[6].

II. SHERLOCK

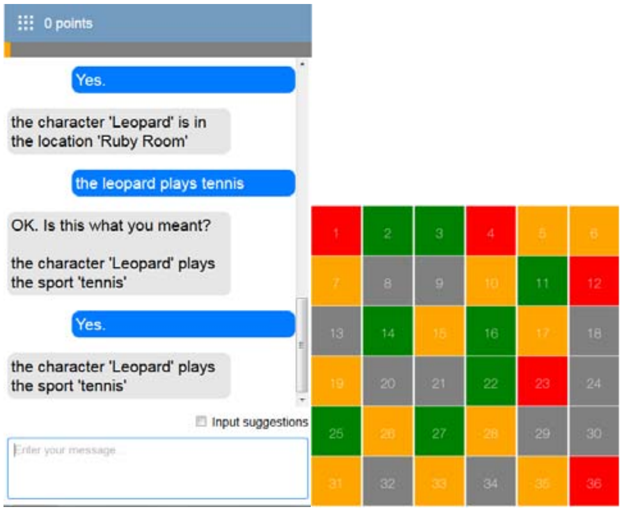
SHERLOCK was designed to support simple situation awareness tasks and the automated fusion of information from human tactical intelligent team members, see [7]. In SHERLOCK, users visited physical locations and reported information to a computational agent on their smartphones. Participants reported information using natural language, which the agent parsed into a Controlled Natural Language (CNL) for the participant's approval or rejection. CNLs such as this one are designed to be both human- and machine-readable. So, if a CNL is designed appropriately, even with minimal or no training, users should be able to recognize whether a CNL translation of their natural language input is correct (and should be "accepted") or incorrect (and should be "rejected"). This would enable the CNL conversational agent to effectively assist the participants. The sample size was N = 161 participants, and they entered a total of 2,482 reports.

A. The task

The ISR task we describe here asked participants to visit posters distributed within a building in order to learn and report the answers to 36 questions they were given (e.g., "What character eats pineapples?", "What sport does the Elephant play?").

This research was sponsored by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. This research was supported in part by an appointment to the U.S. Army Research Laboratory Postdoctoral Fellowship Program administered by the Oak Ridge Associated Universities through a cooperative agreement with the U.S. Army Research Laboratory.

[1] http://orca.cf.ac.uk/79456/



Yes.

the character 'Leopard' is in the location 'Ruby Room'

the leopard plays tennis

OK. Is this what you meant?

the character 'Leopard' plays the sport 'tennis'

Yes.

the character 'Leopard' plays the sport 'tennis'

□ Input suggestions

Enter your message...

1	2	3	4	5	6
7	8	9	10	11	12
13	14	15	16	17	18
19	20	21	22	23	24
25	26	27	28	29	30
31	32	33	34	35	36

Fig. 1 Example smart phone display (left) and dashboard (right).

Participants' progress was shown on a dashboard (see Figure 1) that color-coded the state of each question.

Grey: No information received to answer this question;

Amber: Some information received, but insufficient to give a conclusive answer to this question;

Green: Sufficient information received to give a conclusive answer to this question;

Red: Conflicting information received in answer to this question.

Participants worked as part of a team and were challenged to turn their team's dashboard green.

Five pairs of teams participated, each with one team in the "online" condition and one team in the "offline" condition, which represented unreliable connectivity at the edge. For a team in the online condition, participants' confirmed reports were added to the team's shared knowledge base in real time, and their dashboard reflected the state of the shared knowledge base (i.e., confirmed reports from all members of the team). For an offline team, confirmed reports were added to the team's shared knowledge base only at the end of the experiment once participants returned to a meeting point; until connectivity was established at the end of the experiment, each participant's dashboard reflected only their local knowledge base (i.e., confirmed reports made by the participant him/herself).

Overall, we found that participants in the offline condition contributed more reports than participants in the online condition. The likely explanation for this asymmetry is that online participants' dashboards acted as a common operating picture and allowed participants to stop contributing as soon as the desired dashboard state was obtained. Participants in the offline group, however, were not provided a common operating picture until the end of the task and so did not know if there were questions that they could safely ignore.

B. *Misparses*

A number of the participants' natural language inputs were parsed into CNL in improbable ways, which we refer to as misparses. Specifically, we consider a misparse to be a parse containing a word or phrase that the CNL agent was not preprogrammed with, as the CNL agent was preprogrammed with every entity and relation needed to complete the task. Misparses typically occur when a participant misspells an entity's name or uses an entity or relation that was not preprogrammed, causing the agent to attempt to create a new entity. Two examples are given below (→ = "parsed as"):

NL: Zebra is in the solver room →
CNL: there is a room named 'solver' [should be "silver"]

NL: Lion does not play sport →
CNL: there is a sport named 'does not play'

If the participant confirms these assertions, a new entity is created in the knowledge base, allowing for further misparses, as demonstrated below:

NL: The apple is in the solver room →
CNL: the fruit 'apple' is in the location 'solver'

Some misparses were correctly rejected by participants, but some appear to have been erroneously accepted.

Figure 2 shows the per-participant mean of all reports submitted (which participants then accepted/added to the knowledge base or rejected) on the left. On the right is the per-participant mean of a subset of these reports, the misparses.

As mentioned previously, the participants in the offline groups submitted more reports overall, and the accept/reject rate is similar across conditions (1251/1330 = .94 accepted offline, 1071/1152 = .93 accepted online), as shown in Figure 3. Among the misparses, however, the offline participants had a higher rejection rate, with the online participants more likely to accept misparses (28/65 = .43 accepted offline, 77/119 = .65 accepted online).

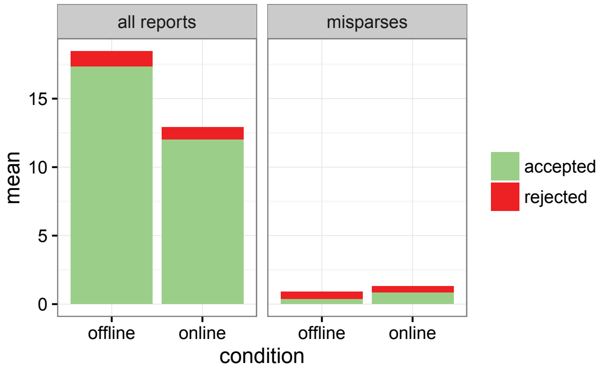


Fig. 2 Mean number of accepted and rejected reports per participant. Left: all reports; Right: misparsed reports.

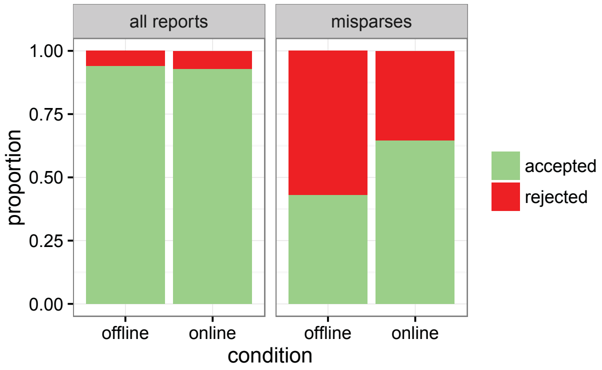


Fig. 3 Proportion of accepted vs rejected reports per participant. Left: all reports; Right: misparsed reports.

A chi-square test found that the number of accepted vs rejected total reports did not meaningfully differ between the online and offline conditions ($\chi^2(1, N = 2,482) = 1.045$, $p = 0.307$, Odds Ratio (OR) = 0.835, OR 95% CI = 0.605 — 1.152). The odds ratio reflects the incidence rate, with a ratio of 1 indicating an equal probability. For misparses, using the probabilities from the total assertions as the null hypothesis, the number of accepted vs rejected misparses differs between the online and offline conditions, ($\chi^2(1, N = 184) = 8.027$, $p = 0.005$, Odds Ratio (OR) = 2.407, OR 95% CI = 1.299 — 4.514). That is, misparses were more than twice as likely to be rejected in the offline condition compared to the online condition. These analyses examine reports and misparses at the overall level and thus assume reports are independent. In the future, we plan to conduct more detailed analyses at the group and individual levels. Results are reproducible; data and analysis can be found at https://osf.io/myv4r/.

III. Conclusions

Two potential complementary explanations for misparses are:

1) Participants thought these parses were correct because of lack of familiarity with the CNL and/or the conversational interface.

2) Automation bias and complacency: Users tended to accept the agent interpretation because it was generally correct (92%).

In the first case, the solution may be to mitigate misunderstanding with more training in the CNL or the interface (e.g., the way reports are accepted/rejected), introduce new feedback and confirmation mechanisms, or in the extreme, to redesign the CNL itself. Less radical fixes include adding a spell-check or term-merging feature so that, for example, the "solver" room and the "silver" room represent the same location in the knowledge base. Autocorrect with spelling may help reduce such errors. However, introducing additional automation may also have unintended consequences, such as user acceptance of erroneous autocorrections. Giving users control over when to use autocorrection has been shown to decrease mistakes in autocorrection [8]. Another fix includes adding reserved words so that, similarly to how most computer languages do not allow "if" to be redefined, predefined CNL entities and relations cannot be accidentally altered[2].

In the second case, participants appear to trust the CNL agent too much. Furthermore, the higher proportion of accepted misparses in the online condition suggests that the connected dashboard showing an up-to-date common operating picture increases trust in the agent, or perhaps is simply distracting. A better understanding of the underlying mechanism(s) for misparses is needed to help users maintain situation awareness when using a conversational interface.

As society's adoption of conversational agents such as Alexa, Cortana, and Siri continues to grow, these results have broader applicability, raising issues of whether, as trust in our agents increases, perhaps so too does the potential for unnoticed misunderstandings.

Acknowledgment

Special thanks are owed to Dave Braines (IBM UK).

References

[1] T. Scherba. (2016, September 15). Conversational Interface Is the New Face of Your Apps | UX Magazine. [Online]. Available: https://uxmag.com/articles/conversational-interface-is-the-new-face-of-your-apps.

[2] L. Bainbridge, "Ironies of automation," Automatica, vol. 19, no. 6, pp. 775–779, Nov. 1983.

[3] R. Parasuraman and D. H. Manzey, "Complacency and Bias in Human Use of Automation: An Attentional Integration," Human Factors: The Journal of the Human Factors and Ergonomics Society, vol. 52, no. 3, pp. 381–410, Oct. 2010.

[4] A. Preece, W. Webberley, D. Braines, E. G. Zaroukian, and J. Z. Bakdash, "SHERLOCK: Experimental Evaluation of a Conversational Agent for Mobile Information Tasks," under review.

[5] A. Preece *et al.*, "SHERLOCK: Simple Human Experiments Regarding Locally Observed Collective Knowledge," U.S. Army Research Laboratory, Aberdeen Proving Ground, MD, Tech Rep. ARL-RP-0560, Dec 2015.

[6] A. D. Preece, W. Webberley, D. Braines, E. Zaroukian, and J. Bakdash, "Human computer collaboration at the edge: Enhancing collective situation understanding with controlled natural language," in Proceedings of the 21st International Command and Control Research and Technology Symposium, London, 2016.

[7] D. L. Hall and J. M. Jordan, *Human-Centered Information Fusion* (Artech House Electronic Warfare Library, 1st ed.). Norwood, MA: Artech House, Inc., 2010.

[8] D. Weir, H. Pohl, S. Rogers, K. Vertanen, and P. O. Kristensson, "Uncertain text entry on mobile devices," in Proceedings of the SIGCHI conference on human factors in computing systems, Toronto, ON, 2014, pp. 2307–2316.

[2] One participant added a new entity "eats" so that a report like "The leopard eats lemons" would include both the relation "eats" and this new unhelpful entity "eats". Because this participant was in the online condition, this new entity then infected the parses of the other teammates.