Towards Detecting and Mitigating **Cognitive Bias** in **Spoken Conversational Search**

Extended Version Available

Kaixin JI[1,3], Sachin Pathiyan Cherumanal[1,3], Johanne Trippas[1], Danula Hettiachchi[1,3], Flora Salim[2,3], Falk Scholer[1,3], Damiano Spina[1,3]

[1]RMIT University, [2]UNSW University, [3]ARC Centre of Excellence for Automated Decision Making + Society

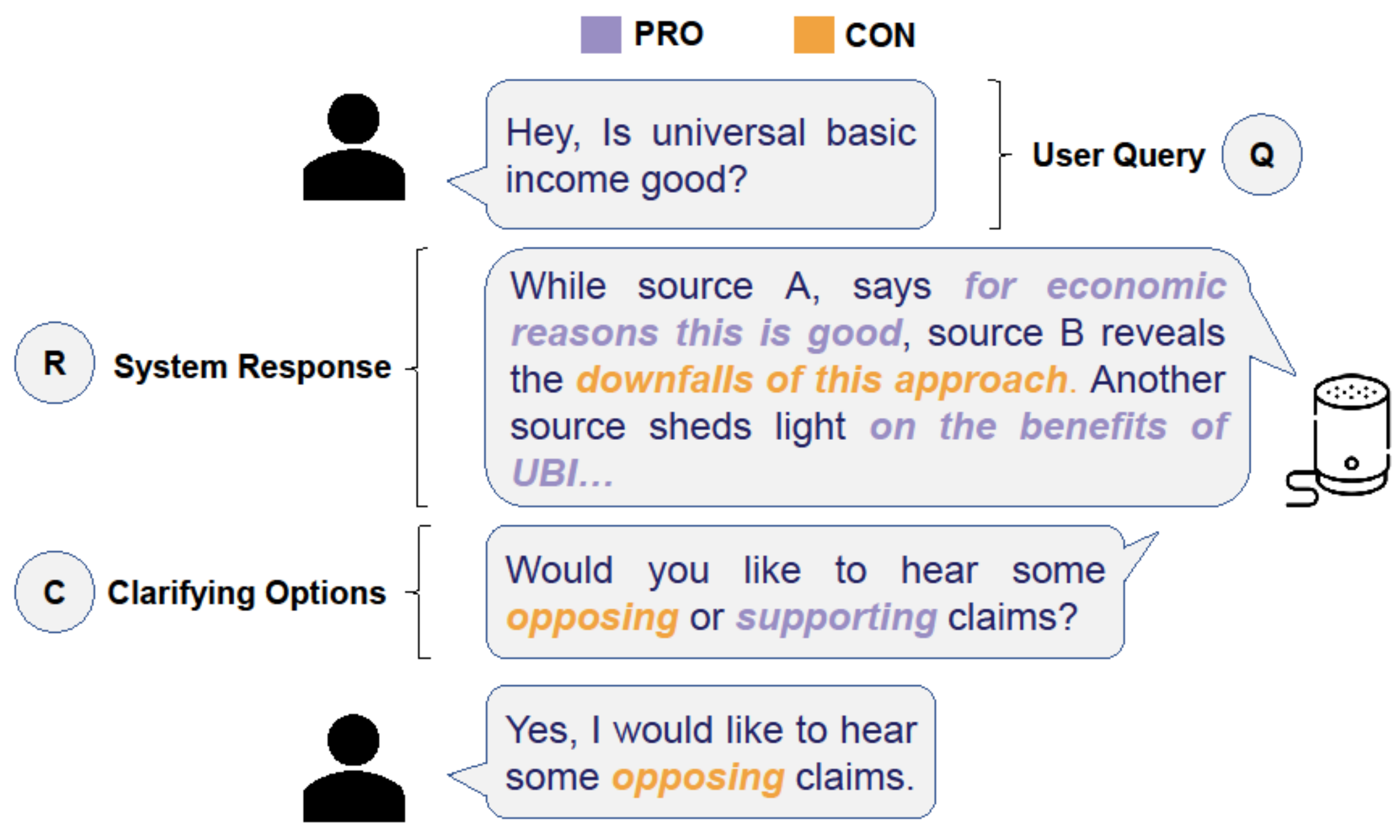
1 INTRODUCTION

The rapid advancement of generative AI has been swiftly integrated into our everyday systems and acted as our personal assistants, and marks a transition **from traditional query-list examine to conversational question-answering** in information searches. Such interaction is evolving towards multimodal capabilities in personal devices, exemplified by the partnership between GPT-4o and Apple. This offers broader accessibility through voice-based interaction, paving the way for **Spoken Conversational Search (SCS).**

Delivering user-friendly yet relevant responses remains a challenge, especially due to limitations in cognition (in processing, analyzing, and interpreting information) and that of the voice channel itself.

2 CASE STUDY: ARGUMENT SEARCH

SCAS systems respond to a user's spoken query on controversial topics with multiple argument stances or view-points (i.e., PRO and CON). Users can rely on SCAS to provide them with balanced arguments on topics of interest.



Hey, Is universal basic income good?

While source A, says *for economic reasons this is good*, source B reveals the *downfalls of this approach*. Another source sheds light *on the benefits of UBI...*

Would you like to hear some *opposing* or *supporting* claims?

Yes, I would like to hear some *opposing* claims.

3 RESEARCH OPPORTUNITIES

1. How to Characterize Cognitive Bias at the Different Stages of the SCS Process?
2. What Is the Role of Clarifying Questions in SCS? How Is It Related to Cognitive Bias?
3. Can Voice Modulation Be Used to Characterize Cognitive Bias?
4. How to Leverage Content Manipulation to Mitigate Harms of Cognitive Bias?

5 LESSONS LEARNED

1. To collect reliable data, these factors should be considered when designing the experiment:

    (i) Data offers direct insights but involves noise and requires specialized designs and expertise, while with fewer channels is easier to analyze;

    (ii) Longer activities (10+ minutes) provide more reliable data, but SCS often involves short tasks (~ 1 minutes);

    (iii) Confounding variables like fatigue, interest, health, and specific activities (e.g., speech) may significantly impact;

2. Ensure optimal contact between sensors with specific body areas.

3. Biases are abstract concepts, the related hypotheses should be deconstructed into specific constructs, like engagement or cognitive load, and further into direct indicators that are measurable, reliable, and objective, e.g., skin conductance or reaction time.

4. During analysis, the requirements of signal processing on frequency can make certain features unavailable or distorted, especially those associated with high frequency in PPG.

5. Analyzing SCS transcripts requires extensive effort and qualitative approaches.

ETHICAL CONSIDERATIONS

- Informed consent and participant awareness of the exposure levels as physiological data could compromise privacy
- Cognitive Liberty

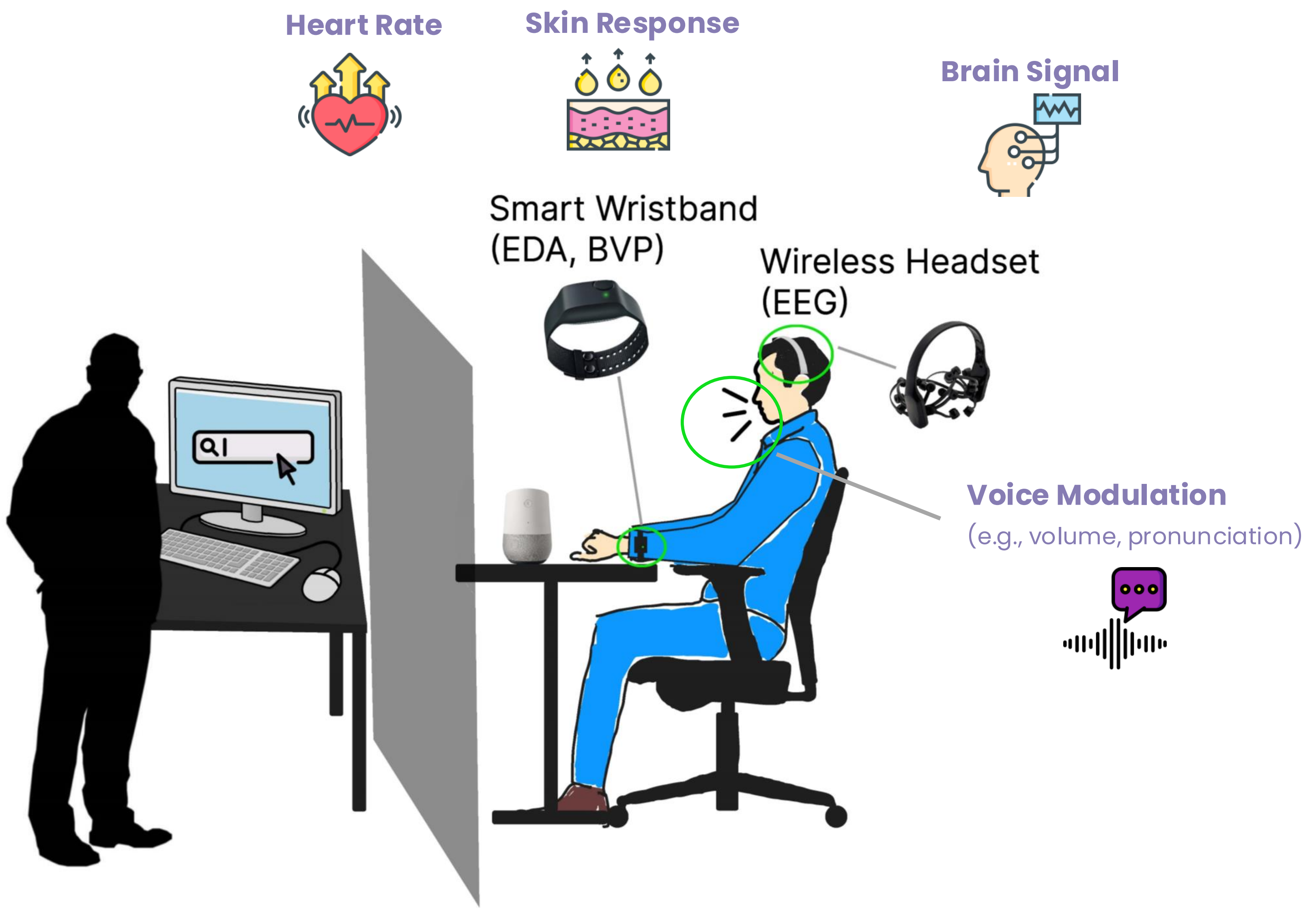
4 PROSPECTIVE METHODOLOGIES

Features	Lab Study	Field Study	Crowdsourced Study
Control	High	Low; unobserved factors in real-world	Moderate; depends on the design of platform or task
Data Quality	High and detailed; due to highly controlled and optimal environment	Low; real-world noise and factors may affect data	Moderate; less controlled than lab studies.
Scalability	Low; requires physical attendance on both participants and researchers	Moderate; enables more participants than lab studies but still limited	High; enables larger participant pool from diverse locations. LLM applications like Retrieval Augmented Generation (RAG) [61] show potential for controlled studies [83, 87]
Ecological Validity	Low; the artificial setting may influence behavior	High; since participants are in natural environments	Moderate; the absence of a physical entity (e.g., smart speaker) may influence user information perception [57]
Setup	Wizard of Oz (WOZ) [27, 111, 116]	Participants are provided with pre-configured voice agents and wearable devices to take home [120]. Comfortable and portable devices may facilitate longitudinal studies.	Crowdsourcing platforms like Prolific enable simulating always-on voice assistants for hypothetical scenarios. Consumer products like Apple AirPods with EEG [7] will make crowdsourced studies more feasible.
Related Works	[13, 45, 79, 109, 111, 113]	[118–120]	[43, 108]

Table 1: A breakdown of different experiment set-ups (i.e., Lab, Field, and Crowdsourced) in *SCS*. LLM: large language model

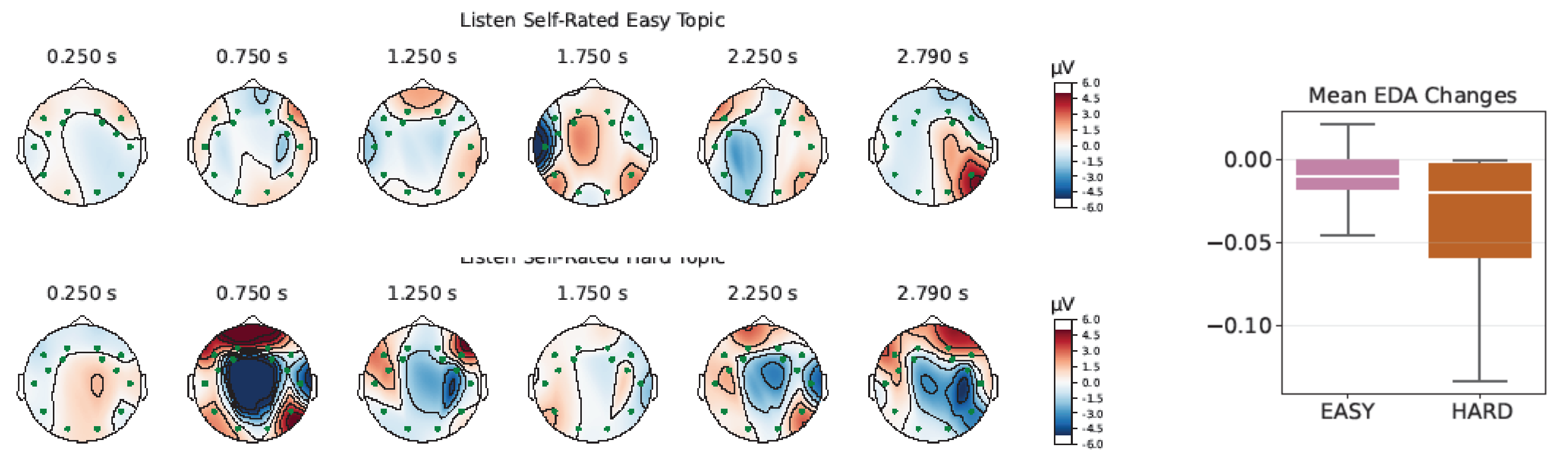
	Data Type	Screen-based		Voice	
		Construct	Related Work	Construct	Related Work
Behavioral	Web-logging (e.g., dwell time, clicks)	**Cognitive Bias**	[26, 59, 106]	–	
	Transcripts & Voice Modulation (e.g., pitch, speed)	–		Perceived Trust	[38, 63]
	Task Performance (e.g., sentiments of query/utterance, recall rate)	**Cognitive Bias** Search Experience	[30] [68, 98]	Listening Effort Search Experience	[16, 49, 51, 89, 97] [49, 98]
	Motion, Facial Expression, Gaze	–		Engagement	[81, 84, 84]
Physiological	Brain Signals (e.g., EEG)	Cognitive Workload Search Experience **Cognitive Bias**	[47, 72] [4, 41, 70, 75, 123] [10, 71, 74, 125]	Perceived Trust	[46]
	Peripheral Sensing (e.g., EDA, PPG)	**Cognitive Bias**	[14, 71, 90]	–	
	Pupillary Responses	Selective Attention	[41, 93]	Selective Attention Distraction Listening Effort	[93] [65] [89]

Table 2: A Breakdown of studied measures by data type (Behavioral vs. Physiological) and user interaction mode(screen-based vs. voice). Bold text highlights studies on cognitive biases, emphasizing the limited research on cognitivebiases in voice search (i.e., SCS).



WoZ Setup + Wearable Physiological Sensors

Preliminary Results



Preliminary EEG (left) and EDA (right) results ($N = 7$) of grand average on listening to search results (about 1 minute) on self-rated *easy* and *hard* topics. In the left figure, deeper colors indicate greater neural activity. Cool colors (negative voltage) represent inhibitory, i.e., suppressing or restricting neural responses, while warm colors (positive) represent excitatory, i.e., promoting or enhancing responses. The dots represent the placement of 14 electrodes.

This research was conducted by the ARC Centre of Excellence for Automated Decision-Making and Society (CE200100005), and funded (fully) by the Australian Government through the Australian Research Council.