I. Opportunity Description

The Defense Advanced Research Projects Agency (DARPA) is issuing an Artificial Intelligence Exploration (AIE) Opportunity, inviting submissions of innovative basic or applied research concepts in the technical domain of designing for accountability in conversational artificial intelligence (AI).

This AIE Opportunity is issued under the Program Announcement for AIE, DARPA-PA-23-04, and was informed by workshops as part of DARPA’s AI Forward initiative. All awards will be made in the form of an Other Transaction (OT) for prototype project. The total award value for the combined Phase 1 base (feasibility study and algorithm development) and Phase 2 option (proof of concept and evaluation) is limited to $1,000,000. This total award value includes Government funding and performer cost share, if required or if proposed.

To view the original DARPA Program Announcement for the AIE Opportunity visit www.SAM.gov under solicitation number DARPA-PA-23-04: https://sam.gov/opp/b04934de80af4f5b92ab51852266c8e7/view

A. Introduction

The FACT AIE effort will explore human-AI dialogue-based methods that avoid overtrust through reflective reasoning (“friction”) that reveals implicit assumptions between dialogue partners, enabling accountable decision-making in complex environments.

Today, AI models (e.g., large language models or LLMs) are easy to use with natural language (“frictionless”), leading users to uncritically use their output and not examine unintended consequences [1, 2, 3]. In addition, the commercial sector is incentivized to provide frictionless experiences.

The technical hypothesis of FACT is that accountable AI can be built through appropriate friction created through dialogue between AI-systems and humans. FACT aims to enable DoD workflows to accountably exploit LLMs, especially in high-stakes situations like strategic planning, intelligence analysis, and reconnaissance.

Background: Dialogue is often central for human teams to solve complex problems where at each stage they understand each other’s intentions, assumptions, and the accountability built into the framing of their proposals and positions; all of which are conveyed explicitly or implicitly during the course of interactions, such as in a search and rescue operation, or in developing a research report on food climate impact in a subcontinent. Such dialogue also affords a ‘learning and verification’ opportunity where one member can point to defects in others’ positions (i.e., playing the devil’s advocate) and present alternatives in terms of relaxed or changed assumptions or emphasis [4, 5]. In dialogue, when all information necessary is not available in advance,
members posit, refine, and assess their own and each other’s positions, thus arriving at solutions that are accountable, avoid known undesirable outcomes, and which also are likely to be more efficient. This sort of dialogue also promotes the Kahneman two-level thinking of fast (to cater to immediate action) and slow (thinking driven by more encompassing considerations) [6]. FACT explores the possibilities for such accountability in human-machine conversational transactions.

State-of-the-art (SOTA): Most of the SOTA efforts on human-machine interactions are framed with well-formulated problems that are presented at each stage to the machine to search for a solution, such as planning, verifying an item in a checklist in which the human has to accept without much dialogue on the basis of the solution, and without regard to potential consequences and accountability. While a successful paradigm in constrained settings (e.g., toy domains, games), it is less appropriate for many realistic settings (e.g., physical environments or open-ended analysis). Such settings require discussion of options and their applicability. Even dialogue-capable systems such as LLMs do not have the capability to reflect on the basis of their responses and the key underlying assumptions.

One elementary way in which some current systems deal with accountability is to design a dialogue that builds “friction” into the flow [7]. A good example is CAPTCHAs that build a pause for human reflection (in addition to bot-exclusion by the machine). Currently, there are no systems that identify the moments and contents for dialogue that promotes accountability and ensures solutions meet implicit considerations not always enumerated at the start. LLMs have no intrinsic mechanism to determine when the tokens produced are simply retrieved facts versus decision points that should be evaluated. This problem is further complicated by the challenges of aggregation and unification of multimodal signals and outputs, as a system may need to indicate what evidence or subsystem is the basis of the decision rather than simply providing a text span or similar explanation.

### B. Objective/Technical Scope

The FACT effort seeks to explore, develop, and evaluate human-AI conversation-shaping algorithms that: 1) capture mutual assumptions, views, and intentions based on dialogue history, 2) auto-assess the consequences of potential actions and the level of accountability for responses, and 3) reveal implicit costs and assumptions to the user, prompting critical analysis, and proposing course changes as appropriate.

Proposals should aim to develop the ideas of technology for the above capabilities, and demonstrate a proof-of-concept in any application domain, such as embodied AI and robotics, Intelligence, Surveillance, and Reconnaissance (ISR), planning, etc.

Proposals should address all aspects described in Section D, Technical Area Description, and also consider other aspects which proposers deem necessary to explore, develop, evaluate, and demonstrate powerful FACT-powered models.

Proposals should clearly describe the fundamental ideas explored, and how the various elements come together to demonstrate the value of FACT in applications. Proposals must clearly define and justify the proposed notion of accountability, argue how the proposed approach can induce mutual accountability, and propose methods for measuring accountability when compared to the
state-of-the-art.

To these ends, proposals should: clearly describe 1) the domain(s) chosen for the evaluation, 2) datasets available, and 3) the state-of-the-art processes and methods and their baseline performance. See Section D for more details.

C. Structure

The FACT effort is planned as an 18-month effort divided into two Phases, with a 12-month Phase 1 (base) focusing on algorithm development and feasibility studies, followed by a 6-month Phase 2 (option) focusing on detailed evaluation of proof-of-concepts in an application domain.

Proposals submitted in response to this AIE Opportunity must be UNCLASSIFIED. Proposals must address the two independent and sequential project phases (a Phase 1 feasibility study (base) and a Phase 2 proof of concept (option)). The Phase 1 (base) award value is limited to $700,000. The Phase 2 (option) award value is limited to $300,000. Both Phase 1 and Phase 2 value limits include performer cost share, if required or if proposed. The total award value for the combined Phase 1 and Phase 2 is limited to $1,000,000. This total award value includes Government funding and performer cost share, if required or if proposed.

All teams will be required to participate in Principal Investigator (PI) meetings and virtual working groups. For planning and cost estimation purposes, all teams should plan to participate in the following events:

- Two-day kick-off meeting in the Washington, DC area;
- Three-day PI meeting in month 6;
- Three-day PI meeting in month 11;
- Two-day Phase 2 kickoff meeting in month 13; and,
- Three-day final PI meeting with evaluation and demonstration event in months 17-18.

D. Technical Area Descriptions

The proposals should identify conversation shaping techniques with AI agents that:

1. Capture mutual assumptions, salient points, views/intentions based on dialogue history (see for example, [8, 9]);
2. Auto-assess consequences of potential actions/level of accountability for AI responses; and
3. Reveal implicit costs/assumptions, and propose course changes where appropriate [10].

In exploring and developing the above conversational techniques, proposals should address:

1. The technical challenges of combining the encoding of dialogue history (extracting key assumptions and causal factors) with decision theory into large pre-trained models, and methods for a priori assessment of the level of friction needed.
2. The questions of: When to start/end dialogue? What are practicable reflective reasoning methods for accountability in interaction with low computational burden? Can friction be dynamically adjusted to the context?
3. The expected computational burden of the proposed system.
4. The metrics to be used for accountability, and the rationale for the choice of metrics.
5. Demonstration and evaluation of the proof-of-concept of the proposed system in an application domain.

Elaborating on the above considerations, the proposal should clearly describe 1) the innovations in the methods to track framing (of each other’s intent, assumptions, saliency) from dialogue history, 2) methods to discriminate whether framing is mutually in accordance with expectations at a given stage of a task execution (or aligned at that stage) and/or is within expected bounds of the task, and c) the conditions for initiating dialogue and the determination of the content of dialogue. For example, a machine may point out that the data/model it has is inconsistent with one of the assumptions of the human partner, who may then relax or change it. In the reverse direction, a machine may present its views or findings (or ways to do a subtask) along with the underlying views/models/assumptions, and the benefits and the cost; and the human may suggest a selection or modification [11, 12, 13]. A technical challenge is to develop new natural language processing and conversational artificial intelligence methods (e.g., using transformers, knowledge graphs) that augment their answers with the underlying assumptions (based on causal analysis) and rank-ordered recommendations. Previous research has indicated that merely providing explanations for the prediction an AI system makes is not sufficient to force people to critically question the AI system in general [14], but forcing people to “stop and think” reduces over-trust in AI systems [15] and can even improve how humans perceive the quality of AI systems [16]. Friction when using AI systems does have the effect of facilitating more reflection on the task and the role of the AI system [16]. These references provide some evidence that more generally adding friction to AI systems would be feasible to mitigate over-trust and improve the quality of AI system output.

Proposals should outline how the system automatically identifies cases where dialogue-driven friction should be minimal and when dialogue-driven friction should be intense. The identification of such cases should explore the influence of the conditions at the start and the application context, for example in a strategic planning context where no urgent decisions need to be made, adding friction universally might be appropriate; whereas in a battlefield context there may be almost no cases where friction should be added because life-and-death decisions need to be made in real time and delaying a decision is as bad as making an incorrect decision.

Proposals for Phase 1 should address the development of algorithms, develop metrics for accountability, and demonstrate the feasibility and efficacy of dialogue-driven friction for accountability in conversational transactions in a laboratory or simulation. Phase 2 must aim to demonstrate the generalizability of solutions to multiple contexts in laboratory or field environments that mimic real-world applications such as planning, collaborative execution of tasks (such as clearing an area, or arranging a room), etc.
References:

15. Z. Buçınca, M. Malaya and K. Gajos, "To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making," in *CSCW*, 2021.

E. Schedule/Milestones

Proposers must address the following fixed payable milestones in their proposals. Proposers must complete the “Schedule of Milestones and Payments” Excel Attachment provided with this AIE Opportunity as part of submitting a complete proposal and fulfilling the requirements under Volume 2, Price Volume. If selected for award negotiation, the fixed payable milestones provided will be directly incorporated into Attachment 3 of the OT agreement (“Schedule of Milestones and Payments”). Proposers must use the Task Description Document template provided with the Program Announcement DARPA-PA-23-04, which will be Attachment 1 of the OT agreement.

Phase 1 fixed milestones for this effort must include, at a minimum, the following:

1. Detailed research plan for Phase 1 which includes updates from kickoff and plans to address Phase 1 evaluation. Milestone Report (Month 2)
2. Report on the application domain to be used for Phase 1 evaluation, and baseline data to be used. Milestone Report (Month 4)
3. Report detailing the progress on algorithm and architecture design. Milestone Report
(Month 6)
4. Preliminary results on performance of components, and plan dates based on first PI meeting discussions. Milestone Report (Month 7)
5. Report detailing target metrics for end of Phase 1, and progress on proof-of-concept. Milestone Report (Month 8)
6. After 12-month end-of-phase evaluation and second PI meeting, deliver report on results and working version (v1) of code and documentation. Milestone Report/Code (Month 12)

Phase 2 fixed milestones for this effort must include, at a minimum, the following:

1. Report on research and evaluation plan for Phase 2 updated after Phase 2 kickoff; software development. Milestone Report (Month 13)
2. Report finalizing evaluation for Phase 2 to demonstrate accountability and dialogue and target metrics. Milestone Report (Month 15)
4. Final report on research accomplishments and lessons learned along with software deliverables (v2). Milestone Report/Code (Month 18)

For planning and budgetary purposes, proposers should assume a performance start date of March 1, 2024. Schedules will be synchronized across performers, as required, and monitored/revised as necessary throughout the program.

All proposals must include the following meetings and travel in the proposed schedule and costs:

- To foster collaboration between teams and disseminate program developments, a two-day PI meeting will be held approximately every six months, with locations split between the East and West Coasts of the United States. For budgeting purposes, plan for three two-day meetings over the course of 18 months: two meetings in the Washington, D.C. area and one meeting in the San Francisco, CA area.
- Regular teleconference meetings will be scheduled with the Government team for progress reporting as well as problem identification and mitigation. Proposers should also anticipate at least one site visit per phase by the DARPA Program Manager during which they will have the opportunity to demonstrate progress towards agreed-upon milestones.

F. Deliverables

Performers will be expected to provide at a minimum the following deliverables:

- Negotiated deliverables specific to the objectives of the individual efforts. These may include registered reports, experimental protocols, publications, intermediate and final versions of software libraries, code, and APIs, including documentation and user manuals, and/or a comprehensive assemblage of design documents, models, modeling data and results, and model validation data.
II. Award Information

Selected proposals that are successfully negotiated will result in award of an OT for prototype project. See Section 3 of DARPA-PA-23-04 for information on awards that may result from proposals submitted in response to this notice.

Proposers must review the model OT for prototype agreement provided as an attachment to DARPA-PA-23-04 prior to submitting a proposal. DARPA has provided the model OT to expedite the negotiation and award process and ensure DARPA achieves the goal of AIE which is to enable DARPA to initiate a new investment in less than 90 calendar days from idea to inception. The model OT is representative of the terms and conditions that DARPA intends to award for all AIE awards. The task description document, schedule of milestones and payments, and data rights assertions requested under Volumes 1, 2, and 3 will be included as attachments to the OT agreement upon negotiation and award.

Proposers may suggest edits to the model OT for consideration by DARPA and provide a copy of the model OT with track changes as part of their proposal package. DARPA may not accept suggested edits. The Government reserves the right to remove a proposal from award consideration should the parties fail to reach an agreement on OT award terms and conditions. If edits to the model OT are not provided as part of the proposal package, DARPA assumes that the proposer has reviewed and accepted the award terms and conditions to which they may have to adhere and the sample OT agreement provided as an attachment, indicating agreement (in principle) with the listed terms and conditions applicable to the specific award instrument.

To ensure that DARPA achieves the AIE goal of award within 90 calendar days from the posting date (November 15, 2023) of this announcement, DARPA reserves the right to cease negotiations when an award is not executed by both parties (DARPA and the selected organization) on or before March 1, 2024.

III. Eligibility

See Section 4 of DARPA-PA-23-04 for information on who may be eligible to respond to this notice.

IV. AIE Opportunity Responses

A. Proposal Content and Format

All proposals submitted in response to this notice must comply with the content and format instructions in Section 5 of DARPA-PA-23-04. All proposals must use the templates provided as Attachments to the PA and the “Schedule of Milestones and Payments” Excel Attachment provided with this AIE Opportunity and follow the instructions therein.

Information not explicitly requested in DARPA-PA-23-04, its Attachments, or this notice may not be evaluated.

B. Proposal Submission Instructions

DARPA will acknowledge receipt of complete submissions via email and assign identifying numbers that should be used in all further correspondence regarding those submissions. If no confirmation is received within two (2) business days, please contact FACT@darpa.mil to verify receipt.

When planning a response to this AIE Opportunity, proposers should take into account the submission time zone and that some parts of the submission process may take from one business day to one month to complete (e.g., registering for a SAM Unique Entity ID (UEI) number or Tax Identification Number (TIN)).

**Electronic Upload**

First time users of the DARPA BAA Portal must complete a two-step account creation process. The first step consists of registering for an extranet account by going to the URL listed above and selecting the “Account Request” link. Upon completion of the online form, proposers will receive two separate emails; one will contain a user name and the second will provide a temporary password. Once both emails have been received, the second step requires proposers to go back to the submission website and log in using that user name and password. After accessing the extranet, proposers may then create a user account for the DARPA Submission website by selecting the “Register your Organization” link at the top of the page. Once the user account is created, proposers will be able to see a list of solicitations open for submissions, view submission instructions, and upload/finalize their proposal.

Proposers who already have an account on the DARPA BAA Portal may simply log in at https://baa.darpa.mil, select this solicitation from the list of open DARPA solicitations and proceed with their proposal submission. Note: proposers who have created a DARPA Submission website account to submit to another DARPA Technical Office’s solicitations do not need to create a new account to submit to this solicitation.

All full proposals submitted electronically through the DARPA Submission website must meet the following requirements: (1) uploaded as a zip file (.zip or .zipx extension); (2) only contain the document(s) requested herein; (3) only contain unclassified information; and (4) must not exceed 100 MB in size. Only one zip file will be accepted per full proposal. DARPA will reject full proposals not uploaded as zip files. Technical support for the DARPA Submission website is available during regular business hours, Monday – Friday, 9:00 a.m. – 5:00 p.m. EDT. Requests for technical support must be emailed to BAAT_Support@darpa.mil with a copy to FACT@darpa.mil. Questions regarding submission contents, format, deadlines, etc. should be emailed to FACT@darpa.mil. Questions/requests for support sent to any other email address may result in delayed/no response.

Since proposers may encounter heavy traffic on the web server, DARPA discourages waiting until the day proposals are due to request an account and/or upload the submission. Note: Proposers submitting a proposal via the DARPA Submission site MUST (1) click the “Finalize” button in order for the submission to upload AND (2) do so with sufficient time for the upload to
C. Proposal Due Date and Time

Proposals in response to this notice are due no later than 4:00 PM on December 14, 2023. As described in Section 5 of DARPA-PA-23-04, full proposal packages must be submitted per the instructions outlined in this AIE Opportunity and received by DARPA no later than the above time and date. Proposals received after this time and date may not be reviewed.

Proposers are warned that the proposal deadline outlined herein is in Eastern Time and will be strictly enforced. When planning a response to this notice, proposers should take into account that some parts of the submission process may take from one (1) business day to one (1) month to complete.

V. Proposal Evaluation and Selection

Proposals will be evaluated and selected in accordance with Section 6 of DARPA-PA-23-04. Proposers will be notified of the results of this process as described in Section 7.1 of DARPA-PA-23-04.

VI. Administrative and National Policy Requirements

Section 7.2 of DARPA-PA-23-04 provides information on Administrative and National Policy Requirements that may be applicable for proposal submission as well as performance under an award.

VII. Point of Contact Information

Matthew Marge, Program Manager, DARPA/I2O, FACT@darpa.mil

VIII. Frequently Asked Questions (FAQs)

All technical, contractual, and administrative questions regarding this notice must be emailed to FACT@darpa.mil. Emails sent directly to the Program Manager or any other address may result in delayed or no response.

All questions must be in English and must include the name, email address, and telephone number of a point of contact. DARPA will attempt to answer questions publicly in a timely manner; however, questions submitted within seven (7) calendar days of the proposal due date listed herein may not be answered.

DARPA will post an FAQ list under the AIE Opportunity on the DARPA/I2O Opportunities page at (https://www.darpa.mil/work-with-us/opportunities). The list will be updated on an ongoing basis until one (1) week prior to the proposal due date. In addition to the FAQ specific to this notice, proposers should also review the Program Announcement for AIE General FAQ list on the DARPA/Office Opportunities page under the Program Announcement for AIE (DARPA-PA-23-04).