Pursuing Fundamental Advances in Human Reasoning

Timothy van Gelder(✉) and Richard de Rozario

University of Melbourne, Parkville, VIC 3010, Australia
tgelder@unimelb.edu.au

Abstract. The IARPA CREATE program's aim to produce "fundamental advances" in human reasoning may provide a new sufficiency test for AGI and insights for the evaluation of AGI performance. The approach of one CREATE program team, the SWARM Project, is outlined.

Keywords: Reasoning · Intelligence · Evaluation

1 Introduction

The standard definition of AGI—the possession, by an artificial system, of "general intelligence at the human level and beyond"—presupposes some understanding of what human-level intelligence actually is. However, this concept has proven elusive. Various tests have been proposed as operational substitutes for a general definition. The Turing Test is the most famous [1], but others include the coffee test [2], and the robot student test [3].

The higher the level of intelligence required to pass a test, the more stringent the test, and the more compelling it would be if an artificial agent passed. Gaining a college degree requires more intelligence than holding an ordinary conversation, and so passing the robot student test is stronger evidence of human-level general intelligence.

It is therefore interesting to ask what the very highest level of general human intelligence might be. The most stringent, and hence compelling, sufficiency test would reference this level.

A research program recently launched by the US Intelligence Advanced Research Projects Activity (IARPA) may shed light on three issues at the heart of this question: (1) the highest level of human performance, (2) how performance at that level can be evaluated, and, (3) how systems achieving that performance might be designed.

This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research projects Activity (IARPA), under Contract 2017-16122000002. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

T. Everitt et al. (Eds.): AGI 2017, LNAI 10414, pp. 259–262, 2017.
DOI: 10.1007/978-3-319-63703-7_24

2 IARPA's CREATE Program

The CREATE (Crowdsourcing Evidence, Argumentation, Thinking, and Evaluation) program aims to produce "fundamental advances in human reasoning" ([4] p. 7) via methods which combine crowdsourcing and structured analytical techniques. Crowdsourcing in this context means collaboration among groups of analysts. Structured analytical techniques are methods intended to produce better analyses [5].

In the CREATE program, four "performer" teams will produce "systems" supporting structured collaboration on difficult reasoning problems. These systems will be rigorously evaluated by an independent testing and evaluation team to determine whether groups of analysts using the systems can meet or exceed prespecified benchmarks for improved reasoning performance, across a wide range of problems, relative to a baseline or control system. Naturally there will be interest in which system performs best overall. However, this "tournament" is friendly and collaborative in nature, with the best outcome being that all systems perform well, though perhaps excelling in different ways or on different types of problems.

The generality of the intelligence required to succeed in the CREATE program is indicated by the range of problems on which systems will be evaluated. "CREATE's methods must be applicable to a wide range of analytic problems, including political, military, economic, scientific and technological questions," such as: Are domestic conflicts in region Y contributing to regional instability? ([4] p. 8).

There is a difficult problem at the heart of the program. How can reasoning performance be measured? If CREATE systems produce superior performance, how can this be reliably demonstrated? This is a problem because, despite all the work over the centuries in logic (broadly speaking) there does not currently exist any widely accepted methodology for rigorously evaluating the quality of complex reasoning.

Some features of the CREATE program make this an especially difficult challenge. First, many of the types of problems CREATE hopes to tackle lack any objective yardsticks. For example, in the case of the sample problem given above, involving the causal explanation of geopolitical instability, there is no gold standard against which answers can be measured. Conclusions can only be evaluated via more reasoning, whose quality is just as questionable as that of the original reasoning.

Second, CREATE is intended to produce reasoning of a higher quality than is achievable by any other method. However, reasoning will necessarily be involved in the evaluation of reasoning produced by CREATE systems. How can superior reasoning be evaluated using (by hypothesis) inferior approaches or methods?

The problem of rigorous evaluation thus involves difficult conceptual issues. It also involves tricky questions of experimental design. The four performer teams, and the test and evaluation team, are collaborating to develop a solution. These efforts are critical to the success of the CREATE program, but they also potentially bear on the problem of rigorously determining whether an artificial agent is engaging in general reasoning at or beyond the highest level of human performance.

3 The SWARM Project

One of the performer teams is the SWARM Project. The acronym stands for Smartly-Assembled Wiki-Style Argument Marshalling. Argument marshalling is a structured analytical technique, similar to argument mapping [6] but not diagrammatic, less rigid, and closer to the natural reasoning behaviors of sophisticated analysts. In the SWARM approach analysts marshal reasoning on a wiki-style platform, i.e., one that supports collaborative and even simultaneous editing of pages. Finally, "smartly assembled" refers to a range of ways the platform supports the production of high-quality reasoning, such as incorporating workflow based on the IDEA protocol [7], or aggregating contributions using information derived from deliberation analytics [8].

At a higher level, the SWARM team aims to succeed by maximizing the collective intelligence of analyst groups using the system; or, in simpler terms, building "super-reasoning teams," analogous to the "superforecasting" teams developed in a previous IARPA program [9]. This challenge is analogous to that of maximizing elite group performance in other contexts, such as sports, military special operations, and surgery. There is an extensive literature on group or team performance. Drawing on recent syntheses (e.g., [10]), it is useful to frame the challenge of maximising the collective intelligence of reasoning groups as one of optimizing the team and its activities along six dimensions or "enabling conditions" [11] of strong group performance:

- **Composition**. Who belongs to the group? More specific issues include: How large should the group be? What attributes should individual members possess? How should attributes be distributed across the group?
- **Processes**. How does the group go about its tasks? What processes, procedures or methods does the team utilize?
- **Resources**. What is provided to the group to enable stronger performance? This includes anything the group can draw in performing their tasks, including equipment, consumables (food, fuel etc.), and information.
- **Motivation**. What drives the group? High-performing groups require strong motivation at individual and group levels. Motivation can be enhanced via good choices on all the other dimensions.
- **Culture**. How can performance be enhanced by means of positive culture? A team's culture consists of its distinctive shared values, standards and practices, over and above what has been made explicit in the team processes.
- **Coaching**. How can performance be enhanced through guidance, feedback, training and conditioning provided to the team as a whole and to its members?

The SWARM team is developing and testing answers to these high-level questions, and many more detailed ones, for the specific case of groups whose mission is to engage in general reasoning. To take one example, in the Resources dimension, a general reasoning group needs a high-quality platform for collaborating in the development of documents expressing their reasoning. To this end, SWARM is developing and testing a new online platform supporting wiki-style argument marshalling.

4 Conclusion

With the CREATE program, IARPA aims to break entirely new ground with regard to human general reasoning capability. If successful, this new standard of performance would arguably define the most stringent sufficiency test for the creation of AGI. That is, if an artificial system could compete at or beyond the level of human groups in a CREATE-style competition, this would represent the most compelling possible evidence that general intelligence had been achieved.

The CREATE program's need to rigorously evaluate whether high levels of performance have been achieved requires it to address some difficult problems in the evaluation of high-level general reasoning. Any progress in this area will also apply to the evaluation of reasoning performance by AGI systems.

Another possible outcome of the CREATE program is to yield insights into how high levels of performance on general reasoning problems can be achieved. These insights might inform the design of AGI systems. For example, it may be that an AGI system could benefit from being designed as a collaboration of artificial agents working together using some of the principles discovered in the CREATE program to result in the highest levels of human performance.

References

1. Pinar Saygin, A., Cicekli, I., Akman, V.: Turing test: 50 years later. Minds Mach. **10**(4), 463–518 (2000)
2. Moon, P.: Wozniak on Apple, AI, and future inventions. In: The Washington Post (2007)
3. Goertzel, B.: What counts as a thinking machine–and when will we meet one? New Sci. **215** (2881), 18 (2012)
4. Intelligence Advanced Research Projects Activity: Crowdsourcing Evidence, Evaluation, Argumentation, Thinking and Evaluation, IARPA-BAA-15-11 (2016)
5. Heuer, R.J., Pherson, R.H.: Structured Techniques for Intelligence Analysis, 2nd edn. CQ Press, Los Angeles (2015)
6. van Gelder, T.J.: Argument mapping. In: Pashler, H. (ed.) Encyclopedia of the Mind. SAGE, Thousand Oaks (2013)
7. Hanea, A.M., McBride, M.F., Burgman, M.A., Wintle, B.C.: Classical meets modern in the IDEA protocol for structured expert judgement. J. Risk Res. **9877**, 1–17 (2016). doi:10.1080/13669877.2016.1215346
8. Shum, S.B., et al.: DCLA meet CIDA: collective intelligence deliberation analytics. In: 2nd International Workshop on Discourse-Centric Learning Analytics, LAK14: 4th International Conference on Learning Analytics & Knowledge (2014)
9. Tetlock, P., Gardner, D.: Superforecasting: The Art and Science of Prediction. Random House, London (2015)
10. Salas, E., Shuffler, M.L., Thayer, A.L., Bedwell, W.L., Lazzara, E.H.: Understanding and improving teamwork in organizations: a scientifically based practical guide. Hum. Res. Manag. **54**(4), 599–622 (2014)
11. Hackman, J.R.: Collaborative intelligence: using teams to solve hard problems. Berrett-Koehler Publishers, San Francisco (2011)