Advances in Cognitive Systems 3 (2014) 3–11 Submitted 9/2013; published 7/2014

Four Research Challenges for Cognitive Systems

Pat Langley PATRICK.W.LANGLEY@GMAIL.COM
Department of Computer Science, University of Auckland, Private Bag 92019, Auckland 1142 NZ
Silicon Valley Campus, Carnegie Mellon University, Moffett Field, CA 94305 USA

Abstract

In this essay, I propose four novel challenges that I hope will encourage progress toward human-level cognitive systems. Each problem – one in entertainment, one in law, one in politics, and one in education – involves the integration of components for which initial technologies exist, although the component tasks remain difficult in their own right. Each challenge also revolves around a virtual embodied agent that interacts with humans in a simulated environment. In each case, I describe the overall challenge problem, subtasks on which researchers can make independent progress, graded versions of the problem that would enable incremental improvement, and methods for evaluating the resulting cognitive systems. I also consider the reasons that both researchers and the public will find the challenge tasks interesting and worthwhile.

1. Introduction

One initial goal of artificial intelligence, inherited by the cognitive systems movement, was to construct complete intelligent agents with the same range of abilities as humans. The Turing test, which posed both a challenge problem and an evaluation scheme for measuring success, reflected this early aspiration for the field. Since Turing (1950) proposed it, both the task and the associated metrics have been criticized as problematic on many fronts. I will not review the many critiques here; instead I will propose tasks that have a similar flavor but that are more focused and tractable.

In the remaining pages, I pose four distinct challenge problems that share two key features with the Turing test. First, they all rely at least partly on conversational interaction with humans. I maintain that this capacity is important because the use of language is a hallmark of human intelligence, as is the ability to reason about the beliefs and goals of others. Second, the main metrics for evaluation depend on human responses and reactions to the agents' behaviors. I argue that this is reasonable because, ultimately, we want cognitive systems that humans find compelling, and conversational agents have always done well on this dimension.

One key difference from the Turing test is that each challenge problem revolves around one or more generic but well-specified tasks. Each supports different versions of the generic problem, which is crucial for demonstrating generality, but there is still a clear notion of what the intelligent system desires to achieve. Another distinction is that each problem, although very challenging, has strong constraints that limit the amount and type of knowledge needed to achieve it. Third, all four problems involve embodied conversational agents, reflecting the emerging consensus that intelligence (at least the human variety) benefits from incarnation in some physical form. Finally,

all but one task concern competitive scenarios that provide both overall measures of success and component metrics that offer more detailed information about system performance.

In the sections that follow, I present challenge problems that involve the fields of entertainment, law, politics, and education, respectively. For each one, I describe the generic challenge and its component tasks, in each case discussing existing work on component technologies that make an integrated intelligent agent possible. After this, I consider constraints on the problem that would make it tractable, along with graded versions that the community could tackle to make incremental progress. In addition, I propose some measures for evaluating the synthetic agents, as well as reasons why the research community and broader public should find the problem appealing.

Before starting, I should add a few words of caution. At first sight, some readers may conclude that the challenges are too difficult and the current state of cognitive systems research does not support them. I agree that they cannot be solved immediately, but still hold that the field can make substantial progress toward each of them by extending and combining existing methods. I also claim that, until we tackle such daunting problems, progress in cognitive systems will continue to be incremental and piecemeal, rather than making serious advances toward human-level intelligence.

2. A Synthetic Entertainer

The first challenge is an example of AI for interactive digital entertainment, a topic that has received increasing attention in recent years and that even has its own conference. Much of the research in this area has focused on developing virtual embodied agents to serve as nonplayer characters in computer games, often giving as much attention to affect, emotion, and personality as to traditional facets of cognition. There has been considerable progress on the component technologies for believable virtual humans, including realistic models for bodies and faces (e.g., Thiebaux et al., 2008), methods for controlling gesture, expression, and posture (e.g., Cassell et al., 1994), gaze (e.g., Thiebaux et al., 2009), techniques for carrying out both spoken and textual dialogue (e.g., Gorniak & Roy, 2005; Kopp & Wachsmuth, 2004; Mateas & Stern, 2004), and for coordinating these activities over time (e.g., Cassell et al., 1994).

Yet despite the broad interest in AI for interactive entertainment, there have been no efforts to develop genuine AI *entertainers*. As the first challenge, I propose the task of constructing a synthetic singer-songwriter.[1] This virtual character would have a simulated human body, a distinctive personality, the basic competencies needed for its profession, and an episodic memory for previous performances and interactions. The component tasks the agent should support include (1) writing the music and words for songs, (2) singing this material on a virtual stage in collaboration with a backup band, (3) performing its songs in staged music videos written and directed by humans, and (4) carrying out brief interviews with reporters and talk show hosts, with questions asked in text rather than in spoken language.

I have already noted some of the existing progress on component technologies for virtual humans. These results are not enough by themselves to produce a high-fidelity pop star, but the pieces appear mature enough to support primitive performers of this sort, perhaps at the level of amateurs

1. This idea has been explored by some science fiction authors, in particular William Gibson in his novel *Idoru* (1996) and, to a lesser extent, Norman Spinrad in *Little Heroes* (1987).

who participate in talent competitions. There have also been successful efforts at writing poetry (e.g., Gervás, 2001) that could produce song lyrics, as well as significant advances in music composition (e.g., Cope, 2006) and performance (e.g., Weinberg et al., 2009), although joining these abilities raises new hurdles. Music videos combine aspects of song performance and theater, which also makes research on synthetic agents for virtual drama (e.g., Hayes-Roth et al., 1997) relevant.

At first glance, dealing with interviews might appear to be the most challenging subtask, but we can make it tractable by imposing constrained syntax on questions and by limiting vocabulary to basic English and words that appear in the performer's material. We can also limit the questions to ones about the meaning of particular songs, the character's opinion about previous performances and interviews, and its feelings about fans and critics. The performer's answers should have a distinctive style and they should be consistent with its personality, but they need not use a diverse range of syntactic forms provided their content is reasonable.

Of course, integrating these capabilities poses challenges of its own, but recent years have seen increasing success at building complex systems, as evidenced by AAAI's Integrated Intelligence and Cognitive Systems tracks. Moreover, the integration issues should be offset by the relative independence of the four component tasks, which can be worked on separately even though some tasks build on others. For instance, the agent needs song material before it can perform, but researchers could test this ability by giving the agent existing lyrics, music, or both. Interviews would focus on the performer's career, such as the meaning of particular songs and its feeling about certain performances, but many aspects of dialogue rely on general conversational skills that do not require such content, and initial interviews could be scripted. Thus, researchers could make initial progress by focusing on reduced tasks that nevertheless lead toward a compelling synthetic pop star.

One advantage of this challenge, like the others I will pose, is that we can measure success in the same manner as for human performers. The most natural metric in this case is the number of songs, videos, and albums that have been sold or, since the public may at first be reluctant to buy material produced by synthetic agents, the number of recordings that have been viewed on sites like YouTube. This would let developers follow the progress of a particular character, whose popularity might grow as the agent's capabilities improve, as its accomplisments grow, and as its reputation with audiences increases over time.

But aggregate scores of this sort provide little feedback for assigning credit and blame, so we should also include measures for component tasks. To this end, one could organize an explicit competition (presumably named 'American Aidoll') in which a panel of human judges rates performers along dimensions like originality and expressiveness. Synthetic characters could compete not only against each other, but also against human-controlled avatars that would provide useful control conditions. The latter would operate in the same virtual environment to ensure that all contestants use the same graphics and animation technologies.

Given the popularity in our society of music, music videos, and the performers who deliver them, it seems clear that many researchers, especially younger ones, will be attracted to this challenge problem. Simplified versions of the task would be appropriate for courses on computer music, dialogue systems, and virtual characters, and well-organized competitions could draw on volunteer energy. Successful synthetic performers could even provide supplemental income for follow-on research, and the authors of such systems would garner their own brand of fame.

3. A Synthetic Attorney

The second challenge problem involves the legal field, an area that AI researchers have studied for decades (Rissland, 1990), but not in the context of virtual agents. Rather than focus on some isolated aspect of legal reasoning, I propose the task of designing and creating a synthetic defense attorney. This intelligent agent would retain knowledge about legal procedures and precedents, have the abilities needed to defend its human clients effectively in mock trials, and operate a simulated body that operates in a courtroom settting.

For this challenge, I envision a number of component tasks, including (1) interviewing the client to gather information about the case, (2) planning a defense to use in court, (3) interacting with the judge during the pretrial hearing and the trial, (4) examining and cross examining witnesses, and (5) preparing and presenting a closing argument. Clients, judges, jury members, and (in some cases) the prosecuting attorney would be avatars controlled by human actors. To narrow the problem, both jury members and witnesses would be provided, along with details about them relevant to the case, to avoid the need for agent-controlled selection.

This problem seems more difficult than building a synthetic entertainer because it requires substantially more reasoning and interaction, but nevertheless I believe the component technologies exist to enable progress. As I have noted, there is a long history of AI work on legal reasoning, much of it focused on inference from the precedents that are so central to the British and US legal systems (e.g., Ashley, 1991; Bench-Capon & Sartor, 2003; Rissland, 1990). We can adapt techniques from this literature to represent, store, retrieve, and utilize knowledge about legal procedures and relevant cases. I have already discussed methods for carrying out dialogue (in this context with the client and judge) and for coordinating gesture, expression, and gaze in virtual bodies, all of which would be necessary for a simulated Perry Mason.

We can make this challenge task less daunting by constraining it on a number of fronts. We could provide a set of relevant precedents, stated in a standard format, for each case from which both sides could draw but not go beyond. We could restrict cases to particular classes, such as murder and assault, that emphasize certain patterns of reasoning. We can restrict the syntax and vocabulary used by the client, judge, and human attorney in order to bypass challenging aspects of sentence processing. And we can eliminate some subtasks, say by focusing on pretrial hearings without juries, skipping the closing arguments, and encoding the results of client interviews manually. I should also note that many trials are far less complex than those on television; we could design cases in which reasonably simple arguments (e.g., self defense or alibis) would let any competent attorney achieve a 'not guilty' verdict.

This challenge problem is even more explicitly competitive than the entertainment task, suggesting again that we use the same measures of success as with human attorneys – whether they win their cases. Of course, many factors can influence this outcome, including the people who serve as judge, jury members, and opposing lawyer. The difficulty of the case itself is also important, as some clients will actually be guilty and thus harder to defend. One natural response would be to let the synthetic attorney participate in multiple cases that involve different judges, juries, and prosecutors, although this may only be realistic for pre-trial hearings that involve no juries. Although winning cases is the ultimate goal, post-trial surveys of jury members and judges could provide more detailed metrics that identify strengths and weaknesses of the defending agent.

Courtroom dramas have held a fascination in our society for decades, suggesting that many people, including AI researchers, will find the construction of a synthetic attorney inherently appealing. The process of defending a client against legal charges has many facets, yet it offers a simple measure of success that will let developers track progress. Constrained versions of the task would be ideal for courses on language processing, reasoning, and virtual characters, and focused competitions could engage the excitement of junior researchers who still hope to develop human-level intelligent systems. Progress in this area could also clarify the nature of our legal system, which would be a worthwhile outcome in its own right.

4. A Synthetic Politician

The third challenge falls in the area of politics, a topic that has received little attention within AI but that seems ripe for study. In keeping with a focus on virtual embodied agents, I propose the task of constructing a synthetic politician who runs for a fictitious public office. As before, the character would control a simulated human body, and it would incorporate knowledge about a constrained set of political issues, have access to memory for events from its career, and support the abilities that are needed for election to office.

Component tasks needed for this activity would include (1) reasoning about a specified set of current issues, (2) writing and delivering speeches on these topics, (3) answering questions from the press, and (4) participating in debates with other candidates. The agent should formulate abstract plans that would address the issues to achieve public goals, defend those plans against critiques, and argue for their superiority over its opponents' proposals. Elections might focus on national, state, or local issues and, to keep the competition on a high plane, we would guard against mudslinging by forbidding comments about candidates' personal lives or abilities (incuding criticisms that it is merely a computer program).

One important way in which this challenge differs from the others is the need for a rich system of beliefs that inform political proposals. Fortunately, Carbonell's (1978) POLITICS system provided an early approach to encoding such content and showed its use in drawing inferences, answering questions, and forming plans. There has been little related work in the interim, although Rizzo et al. (1999) reported a similar approach to modeling personality in terms of abstract goals. Researchers could combine these ideas with advances in text generation (e.g., Traum et al., 2003) for speech writing, coordinated gesture, facial expression, and gaze (e.g., Cassell et al., 1994) for speech delivery, question answering (Strzalkowski & Harabagiu, 2006) for press conferences, and argumentation (Rahwan & Simari, 2009) for debates.

As before, we can generate reduced forms of the challenge problem by removing one or more component tasks (e.g., not all elections involve debates) and by providing human assistance to the candidate (e.g., many politicians depend heavily on speech writers). We can limit questions asked by reporters to topics that have been announced in advance, and we can even let a candidate or its developers select which written questions to answer from a pool submitted before the event. Furthermore, we can constrain the set of issues that candidates address by specifying the political and economic context of the election, along with stating high-level goals (such as increasing employment or reducing inflation) on which all parties agree. We can also provide a party platform – encoded as the higher levels of a hierarchical task network – that the agent can use when formulating

plans to present in speeches and debates. We could make this content available in a standardized logical notation for synthetic agents and provide it in English for human competitors.

Again, we can measure the overall success of our political agent in the same manner as for humans – whether it is chosen for office. Primary elections could involve races among a number of synthetic politicians, but more informative competitions would pit the virtual agent against a human-controlled avatar. Rather than relying on results from a single election, we could hold a series of races that involve different political-economic issues, different party platforms, and different human competitors. Finer-grained evaluations would come from electronic polls taken after speeches, press conferences, and debates. These would measure viewers' opinions about candidates along dimensions like responsiveness to the issues and coherence of proposals. Finally, we could augment public feedback with ratings by a panel of informed political experts.

Given the attention that political elections receive in our society and the allure of winning public office, it seems clear that many researchers would find this challenge intriguing.[2] Simplified variants of the problem would be useful for project-oriented courses on planning, dialogue, and virtual characters, and well-designed competitions could attract energetic young scientists and engineers to the cognitive systems movement. They could also garner increased attention for the field among the general public, as well as shed light on the political process.

5. A Synthetic Teacher

The final challenge concerns education, an area that has a long history in AI and that even has its own conferences and journal. The obvious task here is to develop a synthetic teacher. This agent would have knowledge about some domain of educational interest, include abilities for conveying content to students and assessing their mastery, and control a virtual body that gives it a physical presence. There has already been considerable AI work on interactive tutors, but nothing quite as audacious as the virtual teachers we propose.

We can decompose this challenge into five primary component tasks: (1) composing a series of lectures about the instructional content, (2) presenting these lectures to one or more students, (3) answering questions about the material during or after the lecture, (4) generating exercises and tests associated with each lecture topic, and (5) grading students' answers to exercises and test questions. An embodied teacher is not strictly required, but this seems likely to make the agent's lectures more accessible and entertaining, and thus worth including in our task statement.

This problem lacks some complexities that arise in other tasks we have discussed, in that compelling performances, although desirable, are not crucial and in that persuasive arguments are not essential. However, discourse processing remains central to generating and delivering lectures, as well as to answering students' questions. As we have noted, there exists a substantial body of AI work on education, especially on tutoring systems that offer personalized instruction based on inferred models of student knowledge (Wenger, 1987). There has even been work on tutorial dialogue (e.g., Graesser et al., 2001) and on embodied instructors (e.g., Lester, Stone, & Stelling, 1999).

2. The fact that the virtual agent would not actually hold office should not detain us. Human politicians' behaviors before and after election are so disjoint that one might argue they effectively involve different professions.

We can make this challenge more tractable in a number of ways. We might organize course material into a sequence of lectures and even order the content to be presented within each lecture. We can limit domains to ones that involve formal content, like geometry and physics, as most work on intelligent tutoring systems has assumed. We can also forbid questions during lectures and constrain tests to use multiple choice questions. In addition, we can focus on subsets of domain content to alleviate the effort required for knowledge entry. Naturally, we can relax these restrictions over time to require a more complete set of teaching abilities.

This challenge differs from the others in that it is not inherently competitive. We can certainly measure the effectiveness of different synthetic teachers on student groups, both globally and on particular dimensions, as done with many tutoring systems, but this is not central to the task itself. For this reason, constructing synthetic teachers may well generate less excitement than the other tasks we have proposed. The resulting systems might well have a larger impact on society, which could benefit greatly from improved delivery of education, but experimental evaluation would be substantially more difficult. Moreover, teaching is a less prestigious profession than the others we have discussed, so this task could attract fewer students and junior AI researchers.

6. Concluding Remarks

In this paper, I proposed four challenge problems that could drive future research on integrated cognitive systems. To my knowledge, work on these overall tasks has not appeared in the literature, nor has anyone explicitly suggested them as research targets. In each case, I described the general problem and its component tasks, variants that would enable incremental advances, and approaches to measuring such progress. I believe that each proposal has made a convincing case, but, before concluding, I should address some key questions, shared by all the tasks, that relate to their appropriateness as challenge problems. These are important to researchers who want to propose alternative tasks, since they can draw on the same criteria to argue for their relevance.

Are these good choices for challenge problems? I claim that all four tasks are inherently interesting and thus should have wide appeal, as they concern professions that receive considerable attention and even admiration in our society. Moreover, the challenges are audacious but still limited in scope, so that addressing them would force advances over existing methods while still having some hope of success. This hope is connected to each problem's need for integrated systems that can build on well-defined and reasonably mature component technologies. Finally, three of the challenge tasks lend themselves to competitions that could generate excitement, and each incorporates a virtual embodied agent which could be captured in videos that demonstrate its capabilities. The virtual teacher might attract less interest, but it also requires an integrated cognitive system and it could lead to practical educational outcomes.

Are these challenges well-enough defined? Although I have not provided details for any of the challenges, most readers will have seen enough music videos, legal dramas, and televised politicians, and interacted with enough teachers, to understand their intent. Still, it seems clear that additional effort will be needed to make any of these problems fully operational. This will take time and would benefit from multiple rounds of critiques and revision by interested members of the community. But I am confident that this work could produce well-defined problem statements, a series of tasks with graded difficulty, and clear criteria for system evaluation.

Are these problems tractable? The four challenge problems are intentionally audacious, but I have already outlined ways to make each of them manageable by decomposing them into subtasks. I have also described reduced versions of the problems, assuming constrained abilities and knowledge, that would allow incremental progress. And I should note that we need not set our standards outrageously high. Not all popular performers produce songs with deep lyrics, attorneys often use verbal ploys to influence juries, and politicians are well known for superficial proposals and evasive answers. High verbal skills are not required to become pop stars, and some US presidential candidates have had clear language impediments. This suggests that moderately shallow approaches will prove useful for at least some problem facets, although they should still require much greater depth than competitions like the Loebner Prize. Together, these factors suggest that researchers can make progress on the problems without massive amounts of funding, and that even volunteer efforts might contribute to the construction of compelling synthetic entertainers, attorneys, and politicians.

In closing, let me clarify that I am not proposing that every cognitive systems researcher should focus on these or similar challenge problems. There remains an important need for basic research on the component abilities that underlie intelligence, and other tasks are better suited for driving work at that level. But we also need more research on integrated intelligent agents and, as Swartout (2006) has argued, problems that involve the construction of virtual humans are a natural means to increase efforts toward that end. Finally, readers should recall that the synthetic agents we develop need not be perfect. Even a mediocre singer-songwriter, attorney, politician, or teacher would constitute significant intellectual progress for the field. We should apply one of the earliest insights of AI – Simon's (1955) notion of *satisficing* – to our field's aspirations for integrated intelligent systems.

Acknowledgements

This research was supported by Grants N00014-10-1-0487 and N00014-09-1-1029 from the Office of Naval Research, which is not responsible for its content. I thank David Nicholas and Daniel Shapiro for useful discussions about the ideas presented in this paper. A very similar version of this essay appeared in *Proceedings of the Second Annual Conference on Advances in Cognitive Systems*.

References

Ashley, K. D. (1991). Reasoning with cases and hypotheticals in HYPO. *International Journal of Man-Machine Studies*, *34*, 753–796.

Bench-Capon, T., & Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, *150*, 97–143.

Carbonell, J. G. (1978). POLITICS: Automated ideological reasoning. *Cognitive Science*, *2*, 27–51.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., Douville, B., Prevost, S., & Stone, M. (1994). Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. *Proceedings of the Twenty-First Annual Conference on Computer Graphics and Interactive Techniques* (pp. 413–420). Orlando, FL: ACM Press.

Cope, D. (2006). *Computer models of musical creativity*. Cambridge, MA: MIT Press.

Gervás, P. (2001). An expert system for the composition of formal Spanish poetry. *Journal of Knowledge-Based Systems*, *14*, 181–188.

Gibson, W. (1996). *Idoru*. New York: Viking Press.

Gorniak, P., & Roy, D. (2005). Speaking with your sidekick: Understanding situated speech in computer role playing games. *Proceedings of the First Conference on Artificial Intelligence and Interactive Digital Entertainment* (pp. 57–62). Marina del Rey, CA.

Graesser, A. C., VanLehn, K., Rose, C. P., Jordan, P. W., & Harter, D. (2001). Intelligent tutoring systems with conversational dialogue. *AI Magazine*, *22*, 39–52.

Hayes-Roth, B., Gent, R. v., & Huber, D. (1997). Acting in character. In R. Trappl & P. Petta (Eds.), *Creating personalities for synthetic actors*. Berlin: Springer.

Kopp, S., & Wachsmuth, I. (2004). Synthesizing multimodal utterances for conversational agents. *Journal Computer Animation and Virtual Worlds*, *15*, 39–52.

Lester, J. C., Stone, B. A., Stelling, G. D. (1999). Lifelike pedagogical agents for mixed-initiative problem solving in constructivist learning environments. *User Modeling and User-Adapted Interaction*, *9*, 1–44.

Mateas, M., & Stern, A. (2004). Natural language processing in Facade: Surface text processing. *Proceedings of the Third International Conference on Technologies for Interactive Digital Storytelling and Entertainment* (pp. 3–13). Darmstadt, Germany.

Rahwan, I, & Simari, G. R. (Eds.) (2009). *Argumentation in artificial intelligence*. Berlin: Springer.

Rissland, E. L. (1990). Artificial intelligence and law: Stepping stones to a model of legal reasoning. *Yale Law Journal*, *99*, 1957–1981.

Rizzo, P., Veloso, M. M., Miceli, M., & Cesta, A. (1999). Goal-based personalities and social behaviors in believable agents. *Applied Artificial Intelligence*, *13*, 239–271.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics*, *69*, 99–118.

Spinrad, N. (1987). *Little heroes*. New York: Bantam Books.

Strzalkowski, T., & Harabagiu, S. (Eds.), (2006). *Advances in open domain question answering*. Berlin: Springer.

Swartout, W. R. (2006). Virtual humans. *Proceedings of the Twenty-First National Conference on Artificial Intelligence* (pp. 1543–1545). Boston: AAAI Press.

Thibaux, M., Lance, B., & Marsella, S. (2009). Real-time expressive gaze animation for virtual humans. *Proceedings of the Eighth International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 321–328). Budapest, Hungary.

Thiebaux, M., Marshall, A., Marsella, S., & Kallmann, M. (2008). SmartBody: Behavior realization for embodied conversational agents. *Proceedings of the Seventh International Conference on Autonomous Agents and Multi-Agent Systems* (pp. 151–158). Estoril, Portugal.

Traum, D., Fleischman, M., & Hovy, E. (2003). NL generation for virtual humans in a complex social environment. *Papers from the AAAI Spring Symposium on Natural Language Generation in Spoken and Written Dialogue* (pp. 151–158). Stanford, CA: AAAI Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–60.

Weinberg, G., Raman, A., & Mallikarjuna, T. (2009). Interactive jamming with Shimon: A social robotic musician. *Proceedings of the Fourth International Conference on Human-Robot Interaction* (pp. 233–234). La Jolla, CA: ACM Press.

Wenger, E. (1987). *Artificial intelligence and tutoring systems: Computational and cognitive approaches to the communication of knowledge*. San Francisco: Morgan Kaufmann.