Three Challenges for Research on Integrated Cognitive Systems

Pat Langley
Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA

Abstract

In this paper, we propose three novel challenges that we hope will encourage progress toward human-level intelligent systems. Each problem – one in entertainment, one in law, and one in politics – 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 human avatars in simulated environments. In each case, we 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 intelligent systems. We also consider the reasons that both researchers and the public will find the challenge tasks interesting and worthwhile.

Introduction

One initial goal of artificial intelligence was to construct complete intelligent agents with the same 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 proposed it, both the task and the associated metrics have been criticized as problematic on many fronts (Shieber, 2004). We will not review the many critiques here; instead we will propose tasks that have a similar flavor but that are more focused and tractable.

In the remaining pages, we pose three distinct challenge problems that share two key features with the Turing test. First, they all rely at least partly on conversational interaction with humans. We 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 metrics for evaluation depend on human reactions to the agents' behaviors. We argue that this is reasonable because, ultimately, we want intelligent systems that humans find compelling.

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 aims 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 three problems involve embodied conversational agents (Cassell, 2001), reflecting the emerging consensus that intelligence (at least the human variety) benefits from incarnation in physical form. Finally, each task concerns a competitive scenario that provides both an overall measure of success and component metrics that offer more detailed information.

In the sections that follow, we present challenge problems that involve entertainment, law, and politics, respectively. For each one, we describe the generic challenge and its component tasks. We also discuss existing work on component technologies that make an integrated intelligent agent possible. After this, we 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, we propose some measures for evaluating the synthetic agents, as well as reasons why the research community and broader public will find the problem appealing.

Before starting, we should add a word of caution. At first sight, some readers may conclude that the challenges we pose are too difficult and the current state of AI and cognitive systems does not support them. We agree that they cannot be solved immediately, but we also claim that the field can make substantial progress toward each of them by extending and combining existing methods. We also claim that, until we tackle such daunting problems, progress on intelligent systems will continue to be incremental and piecemeal, rather than making serious advances toward synthetic agents with human-level capabilities.

A Synthetic Entertainer

Our first challenge is an example of AI for interactive digital entertainment, a topic that has received increasing attention in recent years and 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 intelligence. 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 (e.g., Cassell et al., 1994), gaze (e.g., Thiebaux et al., 2009), expression (e.g., Cassell et al., 1994), and posture, 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 interactive entertainment, there have been no efforts to develop genuine AI *entertainers*. As our first challenge, we propose the task of constructing a singer-songwriter.[1] This synthetic character would have a simulated human body, a distinctive personality, the basic competencies needed for its profession, and a memory for previous performances and interactions. The component tasks it 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 spoken language.

We have already noted some of the existing progress on component technologies for virtual humans. We will not claim these results are ready to produce a high-fidelity pop star, but the pieces appear mature enough to support primitive performers of this sort. There have also been successful efforts at writing poetry (e.g., Gervás, 2001) that could produce song lyrics and significant advances in music composition (e.g., Cope, 2006) and performance (e.g., Weinberg et al., 2009), although combining these abilities raises new hurdles. Music videos involve aspects of song performances and theater, which also makes research on synthetic agents for virtual drama (e.g., Hayes-Roth et al., 1997) relevant.

Dealing with interviews is perhaps 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 answers the performer generates should have a distinctive style, but they need not use many syntactic forms provided the content is reasonable. Techniques for embodied conversational agents will play a key role here, although they must be combined mechanisms for episodic memory and reasonably deep language understanding.

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 track. Moreover, the integration issues should be offset by the relative independence of the four component tasks, so they can be worked on separately even though tasks later in the list build on earlier ones. For instance, the performer 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 dialogues rely on general conversational skills that do not require such content. 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 we will pose, is that we can measure success in the same manner as for human performers. The natural metric in this case is the number of songs, videos, and albums sold or, since the public may at first be reluctant to buy material produced by synthetic agents, the number of recordings viewed on sites like YouTube. This would let the developers follow progress of a particular character, whose popularity might grow as its capabilities improve 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 (possibly named 'American Aidoll') in which a panel of human judges rate performers along dimensions like originality and expressiveness. Synthetic characters could compete not only against each other, but also against human-controlled avatars who would provide useful control conditions. The latter would operate in the same virtual environment to ensure use of 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.

A Synthetic Attorney

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

For this challenge, we 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 pre-trial 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 could be provided at the outset, to avoid the need for agent-controlled selection.

[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).

This problem seems more difficult than building a synthetic entertainer because it requires substantially more reasoning and interaction, but nevertheless we believe the component technologies needed to enable progress already exist. As we 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, 1981; 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. We 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 certain 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. We 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 produce 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.

A Synthetic Politician

Our final challenge falls in the area of politics, a topic that has received little attention within AI and cognitive systems but that seems ripe for study. In keeping with our focus on virtual embodied agents, we propose the task of constructing a synthetic politician who runs for public office. The character would control a simulated human body, draw upon knowledge about relevant political issues and memory for events in its career, and incorporate the abilities needed to run for election.

Component tasks needed for this activity would include (1) reasoning about current issues, (2) writing and delivering speeches, (3) answering questions from the press, and (4) participating in debates with other candidates. The agent should be able to formulate abstract plans that would address the issues to achieve public goals, defend those plans against critiques, and argue for their superiority over opponents’ proposals. Elections might focus on national, state, or local issues and, to keep the competition on a high plane, we could forbid comments about candidates’ personal lives or abilities.

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. (1997) 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, expressive, 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 of the component tasks (e.g., not all elections need to involve debates) and by providing human assistance to the candidate (e.g., many politicians depend heavily on speech writers). We can also limit questions asked by reporters at a press conference 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 participants agree. We can also provide a party platform that the agent can use when formulating plans to present in speeches and debates. One could make this content available in a standardized logical notation for synthetic agents and in English for human competitors.

Again, we can measure the overall success of our political agent in the same manner as 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 artificial intelligence. They could also garner increased attention for the field among the general public, as well as shed light on the political process.

Concluding Remarks

In this paper, we proposed three challenge problems that could drive future research on integrated intelligent systems. To our knowledge, work on these overall tasks has not appeared in the literature, nor has anyone explicitly suggested them as targets. In each case, we described the general problem and its component tasks, variants that would enable incremental advances, and approaches to measuring such progress. We believe that each proposal has made a convincing case, but, before concluding, we should address some key questions that are shared by all the tasks.

Are these good choices for challenge problems? We claim that all three tasks are inherently interesting and 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 having some hope of success. This hope is connected to each problem's need for integrated systems that can build on well-defined component technologies. Finally, each challenge task lends itself to competitions that could generate excitement, and each incorporates a virtual embodied agent which could be captured in videos that demonstrate its capabilities.

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

Are these problems tractable? Our three challenge problems are intentionally audacious, but we have already outlined ways to make each of them manageable by decomposing them into subtasks. We have also described reduced versions of the problems that would allow incremental progress. And we should note that not all popular performers produce 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 require much greater depth than competitions like the Loebner Prize.

Nevertheless, progress on any of the challenges will require extensions to existing technologies. Mechanisms for embodied conversational agents that combine language, gesture, gaze, and facial expressions have been tested in many experimental systems, but applying them to entertainment, courtrooms, and politics will undoubtedly reveal limitations that must be overcome. Similarly, these tasks are likely to raise unforseen obstacles for existing approaches to language understanding and generation, argumentation, and planning. Moreover, integrating these capabilities into coherent systems will introduce additional complications. Yet leaping such hurdles is a requirement for long-term progress, and the problems we have posed can serve to galvanize the community to address them.

In closing, let us clarify that we are not proposing that every AI 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 such 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. In addition, readers should recall that our synthetic agents need not be perfect. Even a mediocre singer-songwriter, defense attorney, or politician would constitute significant intellectual progress. We should apply one of the earliest insights about cognitive systems – Simon's (1956) notion of *satisficing* – to our field's aspirations for integrated intelligent systems.

Acknowledgements

This material is based on research sponsored by ONR under agreement N00014-10-1-0487. The U. S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are the authors' and should not be interpreted as representing the official policies or endorsements, either expressed on implied, of ONR or the U. S. Government.

References

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

[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 they are nearly different professions.

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. (2001). Embodied conversational agents: Representation and intelligence in user interfaces. *AI Magazine*, *22*, 67–83.

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 SIGGRAPH '94* (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.

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-Verlag.

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*. Darmstadt, Germany.

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

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., Miceli, M., & Cesta, A. (1997). Personality-driven social behaviors in believable agents. *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*. Cambridge, MA: AAAI Press.

Shieber, S. M. (Ed.). (2004). *The Turing test: Verbal behavior as the hallmark of intelligence*. Cambridge, MA: MIT Press.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*, 129–138.

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).

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.

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.