Towards Natural Cognitive System Training Interactions: A Preliminary Framework

Erik Harpstead*[1], Christopher J. MacLellan*[2], Robert P. Marinier III[2], Kenneth R. Koedinger[1]

[1]Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, 15213
[2]Soar Technology, Inc. 3600 Green Court, Suite 600 Ann Arbor, MI 48105
* These authors contributed equally and should be considered co-first authors.
eharpste@cs.cmu.edu, chris.maclellan@soartech.com, bob.marinier@soartech.com, koedinger@cmu.edu

Abstract

Researchers have developed cognitive systems capable of human-level performance at complex tasks (e.g., Watson and AlphaGo), but constructing these systems required substantial time and expertise. To address this challenge, a new line of research has begun to coalesce around the concept of cognitive systems that users can teach rather than program. A key goal of this research is to develop **natural** approaches for end users to directly train these systems to perform new tasks. However, what makes training interactions natural remains an open research question that we begin to explore in this paper. To lay the foundation for this exploration, we review the human-computer interaction literature to identify characteristics of systems that have historically been natural for end users to interact with. Based on this review, we propose a framework for cognitive system training interactions that decomposes interaction into *patterns*, *types*, and *modalities*, all of which support the acquisition of different kinds of *knowledge*. Finally, we discuss how this framework characterizes existing research within this space and how it can guide future research.

Introduction

In recent years, there has been a growth of research and development in the area of cognitive systems, or systems capable of higher-level processing and reasoning with structured representations using techniques informed by cognitive science (Langley, 2012). For example, IBM's Watson and Google's AlphaGo systems have demonstrated that it is possible for cognitive systems to achieve human-level performance at complex tasks. However, cognitive systems still remain largely out of reach for the general public (Laird et al., 2017). A major factor contributing to this disconnect is that our daily lives are filled with a wide range of tasks across multiple domains, whereas today's state-of-the-art cognitive systems are implemented to perform specific tasks in specific domains. Extending specialized cognitive systems to support a wider range of tasks requires substantial time and expertise (e.g., the base IBM Watson system that famously beat two Jeopardy! champions required over a century of AI expert development time).

To address this challenge, cognitive systems researchers have begun exploring approaches for users to create and extend the capabilities of cognitive systems by teaching them, rather than by programming them. This emerging area of research, which has been referred to as Interactive Task Learning (Kirk & Laird, 2014; Laird et al., 2017) and Apprentice Learning (MacLellan, 2017; MacLellan, Harpstead, Patel, & Koedinger, 2016), aims to develop the computational and cognitive theory needed for building systems that support natural interactions and that possess general capabilities for learning across a wide range of domains and contexts. Similar to how research and development on computing hardware enabled the transition from corporate mainframes to personal computers, this research area aims to support the transition from monolithic cognitive systems (e.g., Watson) to personal cognitive systems (e.g., Forbus & Hinrichs, 2006).

The longer-term goal of our research program is to develop a user-centered approach for teaching cognitive systems. For the moment, we will focus on the issue of naturalness and in particular the naturalness of the training interactions these systems afford. In doing so, we draw on the human-computer interaction perspective that an understanding of interaction is central to the design and development of usable technology. In this paper, we first review commonly recognized characteristics of natural interaction from the HCI literature and propose a preliminary framework that characterizes the space of training interactions that cognitive systems could support. Ultimately, we intend this work to lay the foundation for the development of personal cognitive systems that users can naturally teach.

What Makes an Interaction Natural?

In order to create an initial framework for natural training interactions, we must first contend with what it means for an interaction to be natural. While it is common to think of gesture and speech as lending naturalness to an interaction, the prior literature highlights that an interaction is not necessarily natural by virtue of its physical modality. Norman (2010) argues that so called natural user interfaces (e.g., speech- and gesture-based) are not inherently more natural than graphical user interfaces (e.g., screen-based widgets). For example, gestural interfaces lack the affordances to let users know what gestures they support, whereas graphical user interface widgets, such as buttons, readily advertise their supported interactions. In general, this work suggests that the naturalness of a modality alone is neither necessary nor sufficient for making an overall interaction natural.

Given that naturalness does not derive from modality, then what makes interaction natural? To address this question, we reviewed the HCI literature on natural interactions and identified four common characteristics of systems that support naturalness: they (1) support the goals of the user, (2) do what the user expects, (3) allow the user to work the way they want, and (4) leverage users' experience to minimize training. In this section, we review each of these characteristics.

Supports the goals of the user. Systems supporting natural interactions should be able to support what users want to do (i.e., their goals). One temptation in developing these systems is to overemphasize ease of use at the expense of limiting what users can achieve. Myers, Hudson, and Pausch (2000) refer to this balance as the threshold and ceiling of tools. Thresholds refer to the barriers a user must overcome to use a tool, whereas the ceiling describes what the tool enables users to do. Many systems attempting to support natural interactions emphasize a low threshold, but often ignore the ceiling. For example, it is easy to interact with Siri, but it only supports built-in commands—it is unable to learn new commands. To overcome this risk, systems should be developed with end-user goals and intents in mind (e.g. the desire to teach Siri new user-defined commands), so that the developers can ensure the system does not limit users' capabilities.

Does what the user expects. A common theme in research on natural interactions is an emphasis on the expectations users have for a system (Myers, Pane, & Ko, 2004). Humans typically follow patterns, scripts, or norms when engaging in everyday interactions (Bicchieri, 2006), which make it possible for the humans involved in the interaction to know how to respond. For example, tutors generally expect that their pupils will attempt to solve problems before asking for help. Systems that aspire to naturalness should support naturally occurring patterns of interaction and be aware of users' expectations within these patterns. It is worth noting that these patterns may arise from a user's particular cultural background (e.g., what roles their culture ascribes to teachers and students) or from their personal experiences (e.g., whether they are a Mac or PC user). Additionally, systems attempting to be natural should not require users to learn new (unnatural) patterns of interaction—deviations from typical scripts make it difficult for users to know what the system will do next and how to respond accordingly.

Allows the user to work the way they want. Given that natural systems support users' goals they should also let users execute those goals the ways they prefer or expect to. A key idea from the ubiquitous computing literature is that computing systems should become invisible because they seamlessly support the ways users want to do something (Weiser & Brown, 1996). They should not impede users or force them to achieve goals in unpreferred ways. For example, a common trend is to build systems around a speech interaction paradigm, but there are many situations where speech is an unnatural form of communication. In his study of architectural designers, Schön (1982) found that sketches of designs often better supported communication and reasoning than verbal articulations. This finding suggests that systems aiming to support natural architectural design should prefer sketch-based interactions over speech.

Leverages users experience to minimize necessary training. One of the most pervasive ideas within research on natural user interfaces is the idea of instant expertise (Wigdor & Wixon, 2011), or the idea that users should not have to learn how to control a system because the modality used is one they have immediate familiarity with. In the words of Buxton (Larsen, 2010), "[*natural user interfaces] exploit skills that we have acquired through a lifetime of living in the world, which minimizes the cognitive load and therefore minimizes the distraction"*. Common approaches within this space include voice- and text-based natural language and gestural interfaces that take advantage of users' lived experiences interacting with other people. Additionally, many users have extensive training with artificial interfaces, such as QWERTY keyboards, that may be natural for many application contexts, so it is worth noting that these artificial modes of interaction should not be discounted.

A Preliminary Framework for Cognitive System Training Interactions

In order to design cognitive systems that support natural training interactions, we require a better understanding of how these systems could hypothetically interact. In this section, we will propose a framework for cognitive system training interactions that aligns with the four characteristics noted in the previous sections. We do not intend for this

Table 1. A Framework for Designing Natural Training Interactions for Cognitive Systems

Knowledge	Patterns	Types	Modalities
• Goals • Beliefs • Concepts • Experiences • Skills • Dispositions	• Passive Learning • Operant Conditioning • Direct Instruction • Apprentice Learning • After-Action Review • Socratic Learning • Collaborative Learning	• Command • Clarify • Acknowledge • Inform • Spotlight • Annotate • Reward • Demonstrate • Request <type>	• Command-Line Interface • Control device • GUI • Sketch • API • Gesture • Speech • Text • Multi-modal

work to be complete but hope that it provides a useful language to start talking about naturalness in the context of cognitive systems and their instructional interactions with users.

The framework characterizes four dimensions of training interactions between an agent and a human. First, we assume the goal of an interaction is to change some aspect of an agent's **knowledge**. The interplay between agents and trainers follow instructional **patterns**. Within patterns, trainers engage in several **types** of interaction, and these interactions can be done through various **modalities**. Table 1 shows these four aspects of training interactions and presents examples of each.

Knowledge. The goal of any training interaction is to update the learner's knowledge. There are many types of knowledge that might be included in a cognitive system. However, within the literature, there are several generally accepted types of knowledge (Laird, Lebiere, & Rosenbloom, n.d.). For our preliminary framework, we include six such kinds of knowledge: *goals*, which fully or partially describe desirable states of the world; *beliefs*, which represent an agent's current worldview; *concepts*, which support semantic inference and enable an agent to augment its worldview with additional non-observable information; *experiences*, which organize past situations and problem-solving episodes; *skills*, which describe procedures for changing the world and updating beliefs; and *dispositions*, which specify an agent's problem-solving orientations (e.g., whether to explore or exploit). Our current focus is primarily on symbolic forms of knowledge arising from interactions with a trainer, but future extensions of the framework might also include sub-symbolic knowledge (e.g., learning probabilistic grammar knowledge for parsing English sentences or equations as in Li et al. (2015)). Further, we do not mean to imply that all cognitive systems must support all of these knowledge categories but rather that the nature of the knowledge being changed will likely dictate choices across the other dimensions of the framework.

Patterns. Within human-human instructional settings there are many naturally occurring interaction and training patterns. These patterns govern the relationship between trainer and trainee and establish the contours for how training interactions play out. Inspired by existing systems (Hinrichs & Forbus, 2014; Kirk & Laird, 2014; MacLellan et al., 2016) and instructional practice (Chi & Wylie, 2014; Koedinger, Booth, & Klahr, 2013), our framework highlights several possible patterns. At its most simple, learning could be primarily passive, with agents observing training behaviors without active input from instructors. Increasing complexity, agents can have some control over their actions and receive rewards from the environment or an instructor (operant conditioning) or instructors can explicitly coach an agent, without requiring agent decision making (direct instruction). An even more complex pattern, apprentice learning (MacLellan et al., 2016), incorporate aspects of both of these approaches—both explicit instruction and feedback on agent actions. Additionally, many other instructional patterns are possible, such as after-action review, Socratic learning (Chi & Wylie, 2014), and collaborative learning (Olsen, Belenky, Aleven, & Rummel, 2014).

Types. Within a pattern, an instructor and trainee engage in many types of interactions. For example, within the apprentice learning pattern (MacLellan et al., 2016), an instructor issues a *command*, which specifies the task for an agent to perform. If the agent does not know how to perform the task, then it might *request* a *demonstration* from the instructor, who provides one. On subsequent tasks, the agent might attempt the task (i.e., provide the instructor with a *demonstration*) and *request* feedback (i.e., a *reward*) on this attempt. Finally, the instructor provides the agent with the appropriate *reward*. Under this pattern, this process continues until the agent is correctly performing all of the tasks. Our framework also includes interaction types for supporting Direct Instruction (Hinrichs & Forbus, 2014; Kirk & Laird, 2014), which allow instructors to directly *inform* agents about the world ("TicTacToe is a two-player game"), *spotlight* agents attention on particular parts of the world ("This [pointing] is a block"), and *annotate* demonstrations ("This is the move action [demonstrate drawing of X on board]") to facilitate efficient learning. The types listed in Table 1 are drawn from existing systems as well as the literature on communicative acts (Allen, Blaylock, & Ferguson, 2002; Traum & Hinkelman, 1992). This is not meant to be an exhaustive list, but is representative of the types that commonly occur in current practice. It is important to note that when we refer to interaction types we are interested in the overall instructional act being performed and not how it is

being performed. For example, orders delivered via a command line interface or spoken natural language are both instances of the *command* type.

Modalities. The different types of interactions ultimately ground out in particular modalities of interaction, with many different modalities, or potentially multiple simultaneous modalities, supporting each type. For example, command-line or graphical-user interfaces, are both capable of supporting all of the interaction types listed in Table 1. Typically, systems that claim to support natural interaction leverage modalities commonly used in human-human interaction as the primary modes of interaction. For example, the Microsoft Kinect enables gesture- and speech-based interactions. A key aspect of modalities from our perspective is that they are cast in terms of what the trainer is doing and not necessarily how an action is being detected by an agent. For example, a gesture such as waving could be detected using either visual sensing with a camera or gyroscopic sensing with a wearable device (e.g., Taylor, Quist, Lanting, Dunham, & Muench, 2017); in either case, the trainer would be using a gestural modality.

These four dimensions intentionally map to the four characteristics highlighted in the previous section. In particular, in the context of training, supporting a user's goals consists of supporting of the types of knowledge transference they are trying to achieve. Users' expectations regarding training will derive from the social instructional patterns they have experience with. Thus, in order to naturally support training interactions, it is important for system designers to be aware of the interaction patterns that users expect. Further, users will want to interact in certain ways and system designers should be aware of the different types of interactions they want to perform. Finally, for each type of interaction, system designers should leverage modalities that draw on users' prior experience.

Other Existing Frameworks

The concept of decomposing human-agent interactions using a framework is not new and multiple decompositions exist in the prior literature. For example, Laird et al. (2017) divide interactive task learning systems by the mode of communication used (natural language or demonstration) and the type of knowledge taught (goals, concepts, actions, and procedures). Our work differs in that it also emphasizes the importance of higher-level interaction patterns, such as passive learning, direct instruction, and apprentice learning. Many interactive task learning systems use a pattern similar to apprentice learning, so this dimension may have less variation within that literature. Additionally, we make a distinction between interaction types and modalities because it is possible for interactions to be communicated via different modalities, such as a demonstration (an interaction type) being communicated using sketch, speech, or a graphical user interface (different modalities).

Another related line of work is Bartneck and Forlizzi's (2004) human-robot interaction framework, which, like our framework, has categories for patterns—called norms—and modalities. However, this framework focuses on robot's social interactions with humans more generally, rather than training interactions specifically, and so does not have dimensions for the types of knowledge being taught. Additionally, we emphasize interaction types, which form an intermediate layer of abstraction between patterns and modalities. Finally, as their work emphasizes the physicality of robots, it also distinguishes systems by the form they take (e.g., abstract or anthropomorphic). However, as our work is less concerned with the physical embodiment of agents, we do not make this distinction, but it is not incompatible with our current thinking. In general, while many existing frameworks share commonalities with the one proposed here, their focus is either more general (interaction broadly) or directed toward a different kind of interaction (non-training interactions). Thus, we believe our framework combines prior ideas, but still presents a novel perspective on interaction that is better aligned with our high-level goal of building cognitive systems that are natural for end users to train.

Discussion and Future Work

In proposing this initial framework, we aim to achieve three objectives. First, we attempt to highlight what we view as a key opportunity within cognitive systems research: to better understand the space of training interaction and develop cognitive systems that are natural and efficient for users to teach and interact with. Recent research efforts, such as Rosie (Kirk & Laird, 2014), the Companion Architecture (Forbus & Hinrichs, 2006), and the Apprentice Learning Architecture (MacLellan et al., 2016), have begun exploring different combinations of patterns, types, and modalities to support training interactions with end users. Each of these systems represent particular choices across the dimensions of our framework. To reach a more complete understanding of the space of training interaction design, researchers should explore additional approaches and new combinations of approaches in order to explore the space more broadly and ultimately direct work toward designing more natural means for training cognitive systems.

Second, organizing training interactions along an orthogonal set of dimensions enables a modular approach to the challenge of building cognitive systems to support natural training interactions. Individual researchers or developers need not contend with the whole problem and can instead focus on addressing subproblems. For example, one team of

researchers might investigate which patterns are best for acquiring skills knowledge, whereas another team might investigate which patterns are best for acquiring concepts. Because these decisions are orthogonal, both teams can benefit from each other's work and integrate their findings within the common structure of the framework to support the development of systems that can naturally learn both skills and concepts. Thus, the framework supports the unification of independent research efforts, even if these efforts do not explicitly describe their work within this framework.

Finally, towards the goal of actually building cognitive systems that people can naturally train, we intend our framework to provide a language for formulating scientific hypotheses about how such systems should interact with users to best achieve naturalness. Much of the existing work implicitly assumes that choosing natural approaches for only one of the components of the framework (patterns, types, or modalities) establishes the overall naturalness of a system. For example, Hinrich and Forbus (2014) emphasize the use of multiple natural modalities, such as text and sketching, whereas MacLellan et al. (2016) emphasize the use of a natural pattern. Central to our framework, however, is the hypothesis that different combinations of patterns, types, and modalities of interaction are better suited for updating different kinds of knowledge. Thus, we believe that systems that are natural for users to teach will not only support a wide range of patterns, types, and modalities, but flexibly choose the appropriate combination based on the type of knowledge being communicated, the trainer's preference, and potentially other contextual factors. There is evidence that learning in humans follows a similar logic, in that different kinds of knowledge are best taught by different forms of instruction (Koedinger, Corbett, & Perfetti, 2012). Given that an artificial intelligence need not represent a natural system, there is no inherent reason to transfer this logic (Simon, 1983). However, if we want to support humans in naturally training such systems, then it becomes important to understand these relationships and how they might impact different kinds of training. In conclusion, it is our hope that this framework will focus attention on this issue, provide a language for talking about training interactions and their naturalness, and guide future research on this exciting frontier of personal cognitive systems.

References

Allen, J., Blaylock, N., & Ferguson, G. (2002). A problem solving model for collaborative agents. In *Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems part 2 - AAMAS '02* (p. 774). https://doi.org/10.1145/544862.544923

Bartneck, C., & Forlizzi, J. (2004). A design-centred framework for social human-robot interaction. In *Proceedings of the13th IEEE International Workshop on Robot and Human Interactive Communication - Ro-Man 2004* (pp. 591–594). https://doi.org/10.1109/ROMAN.2004.1374827

Bicchieri, C. (2006). *The grammar of society: the nature and origins of social norms*. Cambridge University Press.

Chi, M. T. H., & Wylie, R. (2014). The ICAP Framework: Linking Cognitive Engagement to Active Learning Outcomes. *Educational Psychologist*, *49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Forbus, K. D., & Hinrichs, T. R. (2006). Companion cognitive systems: a step toward human-level AI. *AI Magazine*, *27*(2), 83–95. https://doi.org/10.1609/aimag.v27i2.1882

Hinrichs, T. R., & Forbus, K. D. (2014). X Goes First: Teaching a Simple Game through Multimodal Interaction. *Advances in Cognitive Systems*, *3*, 31–46.

Kirk, J. R., & Laird, J. E. (2014). Interactive Task Learning for Simple Games. *Advances in Cognitive Systems*, *3*, 13–30.

Koedinger, K. R., Booth, J. L., & Klahr, D. (2013). Instructional Complexity and the Science to Constrain It. *Science*, *342*(6161), 935–937.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, *36*(5), 757–798. https://doi.org/10.1111/j.1551-6709.2012.01245.x

Laird, J. E., Gluck, K., Anderson, J., Forbus, K. D., Jenkins, O. C., Lebiere, C., … Kirk, J. R. (2017). Interactive Task Learning. *IEEE Intelligent Systems*, *32*(4), 6–21.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (n.d.). A Standard Model for the Mind: Toward a Common Computational Framework across Artificial Intelligence, Cognitive Science, Neuroscience, and Robotics. *AI Magazine (Special Issue on The Cognitive System Paradigm: A New Thrust to Attain Human-Level AI)*, In Press.

Langley, P. (2012). The Cognitive Systems Paradigm. *Advances in Cognitive Systems*, *1*, 3–13.

Larsen, L. (2010). CES 2010: NUI with Bill Buxton. Retrieved April 10, 2017, from https://channel9.msdn.com/Blogs/LarryLarsen/CES-2010-NUI-with-Bill-Buxton

Li, N., Matsuda, N., Cohen, W. W., & Koedinger, K. R. (2015). Integrating representation learning and skill learning in a human-like intelligent agent. *Artificial Intelligence*, *219*, 67–91. https://doi.org/10.1016/j.artint.2014.11.002

MacLellan, C. J. (2017). *Computational Models of Human Learning: Applications for Tutor Development, Behavior Prediction, and Theory Testing*. Carnegie Mellon University.

MacLellan, C. J., Harpstead, E., Patel, R., & Koedinger, K. R. (2016). The Apprentice Learner Architecture: Closing the loop between learning theory and educational data. In *Proceedings of the 9th International Conference on Educational Data Mining - EDM '16* (pp. 151–158).

Myers, B., Hudson, S. E., & Pausch, R. (2000). Past, present, and future of user interface software tools. *ACM Transactions on Computer-Human Interaction*, *7*(1), 3–28. https://doi.org/10.1145/344949.344959

Myers, B., Pane, J., & Ko, A. (2004). Natural programming languages and environments. *Communications of the ACM*, *47*(9), 47. https://doi.org/10.1145/1015864.1015888

Norman, D. a. (2010). Natural user interfaces are not natural. *Interactions*, *17*(3), 6. https://doi.org/10.1145/1744161.1744163

Olsen, J. K., Belenky, D. M., Aleven, V., & Rummel, N. (2014). Using an intelligent tutoring system to support collaborative as well as individual learning. In *Proceedings of the International Conference on Intelligent Tutoring systems - ITS 2014* (pp. 134–143). Retrieved from http://link.springer.com/chapter/10.1007/978-3-319-07221-0_16

Schön, D. A. (1982). *The Reflective Practitioner: How Professionals Think in Action*. New York, New York, USA: Basic Books.

Simon, H. A. (1983). Why Should Machines Learn? In R. S. Michalski, J. G. Carbonell, & T. M. Mitchell (Eds.), *Machine Learning*. Springer Berlin Heidelberg.

Taylor, G., Quist, M., Lanting, M., Dunham, C., & Muench, P. (2017). Multi-Modal Interaction for Robotics Mules. In *Proc. SPIE 10195, Unmanned Systems Technology XIX, 101950T (5 May 2017)*. https://doi.org/http://dx.doi.org/10.1117/12.2262896

Traum, D. R., & Hinkelman, E. A. (1992). Conversation Acts in Task-Oriented Spoken Dialogue. *Computational Intelligence*, *8*(3), 575–599. https://doi.org/10.1111/j.1467-8640.1992.tb00380.x

Weiser, M., & Brown, J. S. (1996). Designing Calm Technology. *PowerGrid Journal*, *1*(1), 75–85. https://doi.org/10.1.1.135.9788

Wigdor, D., & Wixon, D. (2011). *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Burlington, MA: Morgan Kaufmann.