Between Instruction and Reward: Human-Prompted Switching

Patrick M. Pilarski and **Richard S. Sutton**
Reinforcement Learning & Artificial Intelligence Laboratory
Department of Computing Science, University of Alberta

Abstract

Intelligent systems promise to amplify, augment, and extend innate human abilities. A principal example is that of assistive rehabilitation robots—artificial intelligence and machine learning enable new electromechanical systems that restore biological functions lost through injury or illness. In order for an intelligent machine to assist a human user, it must be possible for a human to communicate their intentions and preferences to their non-human counterpart. While there are a number of techniques that a human can use to direct a machine learning system, most research to date has focused on the contrasting strategies of instruction and reward. The primary contribution of our work is to demonstrate that the middle ground between instruction and reward is a fertile space for research and immediate technological progress. To support this idea, we introduce the setting of human-prompted switching, and illustrate the successful combination of switching with interactive learning using a concrete real-world example: human control of a multi-joint robot arm. We believe techniques that fall between the domains of instruction and reward are complementary to existing approaches, and will open up new lines of rapid progress for interactive human training of machine learning systems.

Smarter, Stronger, More Productive

Humans make use of automated resources to augment and extend our physical and cognitive abilities. Machine-based augmentation is especially prominent in the setting of rehabilitation medicine—assistive devices like artificial limbs and cochlear implants have taken on a central role in restoring biological functions lost through injury, illness, or congenital complications. In particular, robotic prostheses have made significant improvements to the quality of life and functional abilities achievable by amputees (Williams 2011).

However, as prosthetic devices increase in power and complexity, there is a resulting increase in the complexity of the control interface that binds a prosthesis to a human user. Despite the potential for improved abilities, many amputees find the control of multi-function robotic limbs frustrating and confusing; non-intuitive control is a principal cause of prosthesis rejection by amputees (Peerdeman et al. 2011).

Starting with work in the 1960s, a number of increasingly successful control paradigms have been developed to help amputees direct their powered robotic prostheses. While classical control remains the mainstay for current commercial prostheses, machine learning has provided some of the most successful methods for controlling next-generation robot limbs. Examples of machine learning for multifunction prosthesis control include support vector machines, artificial neural networks, linear discriminant analysis, and reinforcement learning (Scheme and Englehart 2011; Micera, Carpaneto, and Raspopovic 2010; Pilarski et al. 2011, 2012).

The use of artificial intelligence and machine learning is a natural trajectory for automation: in an applications context, we strive to make machines more intelligent so that we can improve our control abilities, achieving greater power and precision when addressing our goals.

Directing an Intelligent System

One consequence of human-machine collaboration is that humans must find ways to successfully communicate their intentions and goals to learning machines. Humans must take on the challenge of directing intelligent automated systems. Interaction is one way of addressing this challenge. Through ongoing interactions, a human can direct and mould the operation of a learning system to more closely match his or her intentions. Information from a human trainer has been shown to allow a system to achieve arbitrary user-centric goals, improve a system's learning speed, increase asymptotic performance, overcome local optima, and beneficially direct a system's exploration (Judah et al. 2010; Kaplan et al. 2002; Knox and Stone 2012; Lin 1991–1993; Pilarski et al. 2011; Thomaz and Breazeal 2008).

It is natural to expect that providing added instructional information to a learning system will help drive the learning process (Lin 1992; Thomaz and Breazeal 2008). Interactive teaching is a dominant approach to human and animal learning, and techniques from these biological domains seem to transfer well to the machine learning case. Building on a basis in biological learning, many approaches operate within the framework of reinforcement learning (Sutton and Barto 1998) and deliver direction by way of generalized scalar feedback signals known as *reward*; others provide explicit *instruction* in the form of demonstrations, performance critiques, or semantically dense training interactions.

The use of reward and instruction during interactive learning has produced a number of important milestones. Previ-

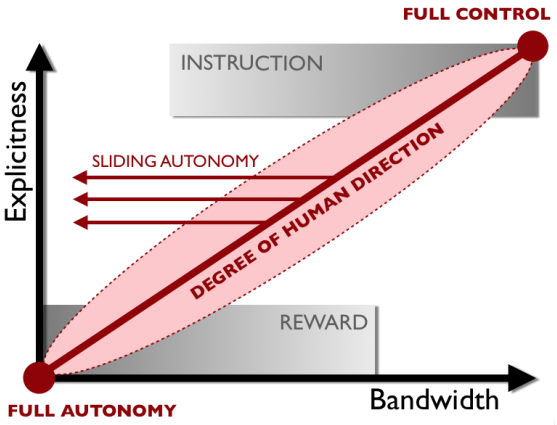

Figure 1: The continuum of interactive training and direction methods. There are a number of ways that human-generated signals have been used to direct a learning machine. We can characterize these interactions as lying within a two-dimensional space. One dimension corresponds to how explicit the human signals are and the other corresponds to the overall bandwidth or information density of the signals. Most application domains lie along the diagonal between full autonomy and full control, shown in red.

ous work has shown how trial-and-error machine learning can be enabled or accelerated through the presentation of human-delivered rewards and intermediate reinforcements. Examples include the use of shaping signals (Kaplan et al. 2002), the combination of human and environmental reward (Knox and Stone 2012), multi-signal reinforcement (Thomaz and Breazeal 2008), and our preliminary work on human-directed prosthesis controllers (Pilarski et al. 2011). The presentation of interactive learning demonstrations or instructions has also been shown to help teach a collection of viable sub-policies even when a globally optimal policy is challenging to achieve (e.g., Chao, Cakmak, and Thomaz 2010; Judah et al. 2010; Kaplan et al. 2002; Lin 1991–1993). As such, leading approaches to the human training of a machine learner almost exclusively involve the presentation of new information in the form of instruction or reward. These human directions and examples supplement the signals already occurring in a machine learner's sensorimotor stream.

Work on instruction and reward is representative of a growing body of literature on interactive learning, and there are a number of ways that non-interactive human guidance has been used to direct learning machines. We suggest that the continuum of human training and direction methods can be usefully separated along three main axes:

Explicitness: Explicitness describes the degree to which the signals from a human user contain explicit semantics, and relates to the detail of voluntary human involvement. At one end of this axis is reward, as in the reinforcement learning case of a scalar feedback signal (e.g., Knox and Stone 2012) or a binary shaping signal (e.g., Kaplan et al. 2002). At the other extreme is instruction in the form of demonstration learning and supervisory control (e.g., Lin 1991–1993), performance critiques following a period of action by the learner (e.g., Judah et al. 2010), and the Socially Guided Machine Learning of Chao, Cakmak, and Thomaz (2010).

Bandwidth: Bandwidth refers to the rate and density with which information is passed to a learning system, in terms of signal frequency, signal complexity (binary, scalar, vector, nominal), and the number of distinct signalling channels. Directive information may be as simple as a single binary reinforcement or shaping signal (Kaplan et al. 2002), can involve multiple signals or signal types being presented to the learning system (Thomaz and Breazeal 2008), or can involve processing with verbal or non-verbal cues (Chao, Cakmak, and Thomaz 2010). Signals may be presented regularly during real-time operation (Knox and Stone 2012; Pilarski et al. 2011), or may be sparse and irregularly spaced with gaps between signals (e.g., Chao, Cakmak, and Thomaz 2010).

Immediacy: Interaction may vary in terms of its timeliness, from real-time interactions between a human and a learner, to episodic, asynchronous, or offline interactions. Highly interactive direction involves instantaneous or immediate feedback signals about what the agent has done or is about to do (Kaplan et al. 2002; Thomaz and Breazeal 2008; Knox and Stone 2012). Less interactive direction occurs when human signals are presented before or after a learner's operation—e.g., the *a priori* presentation of temporally extended training sequences (Lin 1991–1993) or *a posteriori* performance evaluation (Judah et al. 2010). Fixed control schemes, such as classical PID control and pre-defined reward functions, occupy the far end of the immediacy axis.

We are interested in human-robot control settings where a machine learner improves through ongoing, real-time interaction with the human user over an extended period of time. Artificial limbs and assistive rehabilitation devices fall into this category. As such, for the remainder of this work we will deal with the case of interactive direction and therefore focus on ideas of bandwidth and explicitness.

The two dimensional space formed by combining bandwidth and explicitness is shown in Figure 1. The bottom left of this continuum represents fully autonomous operation (no human direction), while the top right represents full human control (explicit high-bandwidth supervision; no automation). The notion of sliding control between a human and an autonomous system can also be represented on the continuum shown in Figure 1. As one example, a reduction in the number or frequency of signals needed from a human user takes the form of a shift in communication bandwidth (Figure 1, red arrows pointing left).

A critical region of the bandwidth/explicitness continuum is the spectrum we define as the *degree of human direction*—the line on the diagonal between full control and full autonomy (Figure 1, red envelope). This spectrum represents a natural relationship between the complexity and semantic content of human-derived signals. As shown by a survey of recent literature, most human-machine training and direction work to date has focused on either reward or instruction—the lower left and upper right regions of the ex-

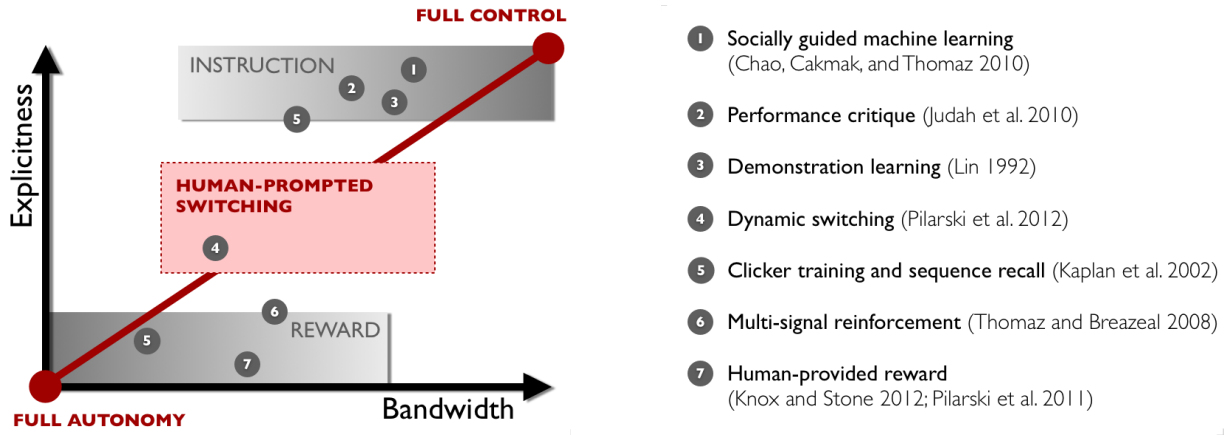

Figure 2: The setting of human-prompted switching (shown in red) is located in the middle ground between reward feedback and human instruction. Markers suggest the positions of representative research examples in terms of the bandwidth/explicitness continuum. Marker relationships are not to scale; locations are only intended to imply general characteristics of the methods.

plicitness/bandwidth continuum. The markers in Figure 2 suggest the position of representative research in terms of the continuum between full autonomy and full human control. With few exceptions, the middle of the space between instruction and reward remains surprisingly unexplored.

In the work that follows, we suggest that—for concrete applications—the space between instruction and reward embodies a fruitful area for the human direction of machine learners, encompassing a rich array of techniques that balance human effort with communication efficiency. To provide traction in this domain, we now introduce the idea of human-prompted switching.

Human-Prompted Switching

In many assistive technologies, human direction takes the form of manual prompts that cause a machine to switch between its different functions or actions. We call this setting *human-prompted switching*. In human-prompted switching, a user is often faced with the task of controlling a machine that has more functions than the user can independently actuate. In other words, the number of channels that can be controlled by the user is smaller than the number of controllable dimensions available within the mechanical or computational system. This disparity may be due to cognitive, physical, mechanical, or computational constraints on the human user's ability to direct the system. As such, the human user is required to switch their control between the system's different functions, actuating only a subset of the available control dimensions at any given time. Switching-based interactions of this kind occur frequently in human-machine contact; notable examples include surgical robotics, teleoperation and telepresence, semi-autonomous vehicle control, and assistive biomedical robotics.

Human-prompted switching is a versatile setting for studying human direction and sliding autonomy in a machine learning system. Switching prompts by the user are light in terms of cognitive demand, being less explicit than most instructional interactions. At the same time, prompting and switching signals from the user contain more semantic content than a classical reward signal—a user's switching prompts uniquely specify both the correct switching action and also the timing of a functional shift. As such, human-prompted switching represents one example of a learning scenario that lies in a central position between the extremes of full control and full autonomy (Figure 2).

A key feature of switching-based interactions is that a user's control intent is implicitly presented by way of the user's switching actions. Directions from the human user are available *in-channel*—i.e., within the normal sensorimotor stream of the learning machine. In other words, action or lack of action by the human side of the human-machine interface provides an implicit directive to the learner. In the ideal case, no input (or minimal input) from the user is required—the fact that the human is interacting with the system smoothly is positive, reinforcing feedback, and manual changes or non-intuitive use is implicit negative feedback that the system needs to adapt its behaviour. Feedback and reinforcement may therefore be extracted automatically from the human's use of the system. This idea has similarities to the work of Kaplan et al. (2002); in their work, confirmation or lack of confirmation by a human trainer was used by a robot learning system to generate alternate proposals for action sequences.

The combination of prompted switching with interactive machine learning can take several forms. In this work we discuss two cases. In the first case, a system learns to passively anticipate how a human will switch between functions. Learned predictions about a user's switching behaviour in different contexts are used to suggest appropriate controllable functions to the user—the system will switch only when prompted by the user, but will autonomously select the switching target. In the second case, the system will switch functions autonomously according to its observations about itself and its user. The system takes the initiative in switching, but can have its actions overridden by a human's

manual switching actions—if the system delays too long in switching, the user will manually prompt it to switch functions. In both situations, if the system selects an incorrect function, the user will manually switch to the correct function. Human switching signals therefore provide implicit feedback that can be used by a learning system to both alter its immediate behaviour and update its learning parameters. Prompting also provides a natural way to effect sliding autonomy—the better a system performs, the less prompting is required from its human user.

In what follows, we examine the combination of human-prompted switching and learning via a representative thought experiment (motor vehicle control) and one immediately applicable case study: recent work by our group on the human control of an assistive biomedical robot.

Thought Experiment: The Intelligent Semi-Automatic Transmission

The automobile transmission is an excellent example of an automated system that can occupy different locations along the spectrum of human direction. With the manual transmission found in most sports cars, the human driver is given full control over when and how the car will shift gears. Conversely, in cars with automatic transmissions, the car has full autonomy over the gearbox and the human does not participate in switching gears. While relatively uncommon, some cars have what is called a semi-automatic transmission, or clutchless manual transmission. With the semi-automatic transmission, the driver is able to manually switch gears by pressing a toggle switch; the driver specifies the need to switch gears, but does not actuate the clutch or explicitly move the transmission to the desired gear.

We now imagine the intelligent semi-automatic transmission. It is able to perform a switching action autonomously in response to observations from its environment (e.g., speedometer or tachometer readings). Our intelligent transmission begins with a reasonable policy for switching gears in different situations. However, at any time the driver may manually force the system to switch gears using a physical button on the steering wheel. In order to improve the car's performance in different situations—e.g., passing another car or going up hill—the driver uses their input signal to prompt the system to switch gears. Prompting may occur if the system delays too long or if it switches to an inappropriate gear. Each time the user prompts the system, the system updates its knowledge about the user's preferences.

Over time, the learning system within the transmission adjusts its policy so that the user is required to provide less and less manual prompting; control slides from frequent prompting on the part of the driver to mostly autonomous operation. However, should the situation change—for example, should the car now be in a different environment, like icy roads during the car's first winter—the driver may resume prompting as often as needed to ensure the high performance of the car. In this collaboration between a human and a machine, the human is able to optimize the system's performance by providing simple, momentary prompts. Learning is a continuous, ongoing process.

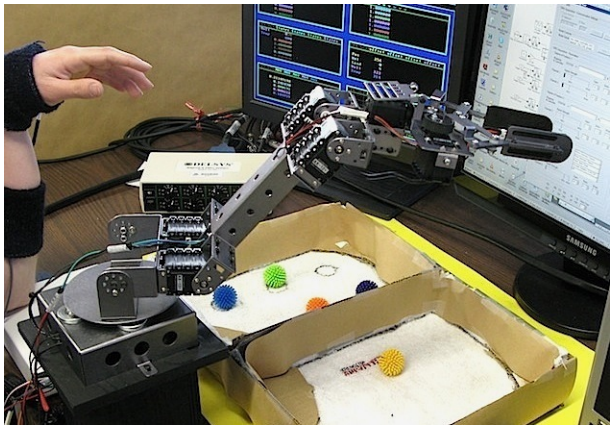

Figure 3: A table-top robot arm used to train new amputees. The multiple functions of the robot are selected and controlled using electrical recordings from three of the user's muscles, as processed by a real-time control computer.

Case Study: Dynamic Switching for an Assistive Biomedical Robot

We now turn to a real-world example that combines learning and prompting for the human control of a biomedical robot with multiple functions. As described above, robot arms are now being used as assistive devices by upper-limb amputees. Control information for these devices is typically recorded from the muscle groups on an amputee's body. As a patient's level of amputation increases, there is a corresponding decrease in the number of available muscle recording sites; viable control sites on an amputee's body are often severely limited in number (Williams 2011).

A lack of discrete control sites restricts the number of prosthesis functions that can be simultaneously controlled by an amputee (Scheme and Englehart 2011). The standard approach in this situation is to give an amputee the ability to manually switch how their available control channels influence the different joints or functions of their prosthesis. User switching is therefore a key feature in most commercial prostheses with more than one function.

In recent work, we presented an interactive, learning-based approach for dynamically switching between the different joints of a robot arm (Pilarski et al. 2012). The platform for our experiments was a table-mounted robot with motors located at its hand, wrist, elbow, and shoulder joints (Figure 3). This system was designed by Dawson, Fahimi, and Carey (2012) to help prepare new amputees for the control of a powered commercial prosthesis. As in commercial artificial limbs, muscle activity recorded from the user's body was converted into control commands for the robot limb. Two recording channels were combined to control the velocity of a user-selected joint. A third recording channel was used as a manual switch to shift control between the arm's different joints according to an expert-designed (fixed) cyclic order. This configuration allowed a user to sequentially control the arm's four degrees of freedom by contracting and relaxing different muscle groups.

Algorithm 1 Online Learning of General Value Functions

initialize: $\mathbf{w}, \mathbf{e}, \mathbf{s}, \mathbf{x}$
repeat:
  observe $\mathbf{s}$
  $\mathbf{x}' \leftarrow \mathsf{approx}(\mathbf{s})$
  for all joints j **do**
    observe joint activity signal r_j
    $\delta \leftarrow r_j + \gamma \mathbf{w}_j^T \mathbf{x}' - \mathbf{w}_j^T \mathbf{x}$
    $\mathbf{e}_j \leftarrow \min(\lambda \mathbf{e}_j + \mathbf{x}, 1)$
    $\mathbf{w}_j \leftarrow \mathbf{w}_j + \alpha \delta \mathbf{e}_j$
    $\mathbf{x} \leftarrow \mathbf{x}'$

The prediction of future joint activity p_j at any given time is sampled using the linear combination: $p_j \leftarrow \mathbf{w}_j^T \mathbf{x}$

One downside to a switching-based myoelectric control approach is that a user can spend an unacceptably large portion of their time shifting their control between different joints or functions. In a 20.7min task with the robot arm, we found that a skilled non-amputee subject spent more than 10.4min ($\sim 50\%$) of their time switching between different joints. We explored the use of a machine learning system to reduce the magnitude of these switching-related delays.

In contrast to systems with a fixed switching order, our system learned to dynamically adapt the switching suggestions presented to the human user (Pilarski et al. 2012). By interactively learning the user's switching preferences, we hoped to minimize the number of required switching prompts. In our approach, dynamic switching order suggestions were formed out of learned knowledge about the user's control behaviour. Switching prompts and joint actuation by the user provided the interactive direction signals needed to shape the system's learning. Human use of the system's switching suggestions implicitly confirmed their correctness; prompting in the form of additional switching actions served as corrective feedback, reinforcing the user's intended switching decision in a given situation.

Learned knowledge took the form of temporally extended anticipations and predictions about the user's joint control preferences and switching prompts. Learning occurred online using a reinforcement learning approach, in this case an implementation of General Value Functions (Sutton et al. 2011) and Nexting (Modayil, White, and Sutton 2012). As described in Algorithm 1, predictions about joint activity were represented using generalized value functions, and learned in real time using temporal difference methods. The weight vectors $\mathbf{w}_j$ for each general value function were updated at each time step using new information about the current state of the system ($\mathbf{x}'$) and the signals of predictive interest (r_j). The binary state vector $\mathbf{x}'$ was approximated from the real-valued input space $\mathbf{s}$ using a tile coding function approximation method (Sutton and Barto, 1998), here denoted $\mathsf{approx}(\mathbf{s})$. For each joint j, a temporal difference error signal (denoted δ) was formed using a joint activity signal r_j and the difference between the current and future predicted values for this signal (computed from the weight vector using the linear combinations $\mathbf{w}_j^T \mathbf{x}$ and $\gamma \mathbf{w}_j^T \mathbf{x}'$, respectively). Next, a trace $\mathbf{e}_j$ of the current feature vector was updated in a replacing fashion, where $\mathbf{e}_j$ was an eligibility trace vector with λ as the corresponding eligibility trace decay rate. This trace $\mathbf{e}_j$ was used alongside the error signal δ to update the weight vector $\mathbf{w}_j$ for each value function. Here $\alpha > 0$ was a scalar step-size parameter. Learned predictions p_j for each signal r_j were sampled from the weight vector $\mathbf{w}_j$ and current feature vector $\mathbf{x}$ according to $p_j = \mathbf{w}_j^T \mathbf{x}$.

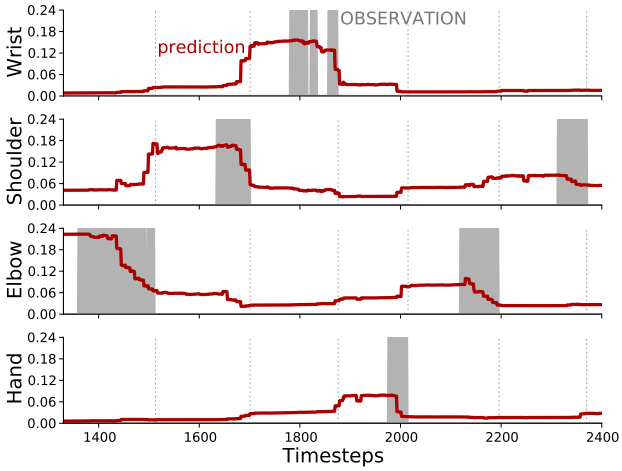

Figure 4: Predictions about human joint control activity. Our system learned in real time to anticipate a user's switching preferences during their operation of a multi-joint robot arm. Anticipation is shown in terms of the system's predictions about the human's joint control signals and switching signals. Shaded areas of the plot indicate human control of a given joint (the observed signal of interest). The system's temporally extended predictions about this control activity are shown as solid red lines. These results show prediction accuracy after 15min of online learning.

Four predictions were learned in parallel, with one prediction for the motion of each user-driven joint. By ranking the magnitude of these predictions prior to manual switching by the user, the learning system was able to form an adaptive switching order. In other words, simple relationships between the predictions were used to formulate the system's switching suggestions. These suggestions depended on both context and learned knowledge about a user's preferences. Learning updates occurred in real time (every 20ms), and took into account an observation space comprised of all four joint angles on the robot arm and 28 other sensors relating to the robot arm and the human's recorded muscle activity. For a full description of the learning mechanisms and experimental procedures summarized in this case study, please see Pilarski et al. (2012).

Figure 4 shows an example of the system's predictions after 15min of real-time learning from its interactions with a human user. The predictions made by the learner correctly matched the true (computed) observations, and rank ordering these predictions accurately anticipated the next joint or joints to be used by the human. We compared the system's adaptive (dynamic) switching order to both the expert-

designed switching order mentioned above, and the optimal switching order as computed post-hoc from the interaction data. Our dynamic switching approach was found to significantly decrease the switching time and number of switching prompts required from a human user. The switching orders generated by the learning system indicated a 14% decrease in switching time on the target task, as compared to the optimal fixed switching order computed for this data (a savings of 1.5min out of the total 20.7min task time). When compared to the fixed, expert-designed switching order, the dynamically generated switching order indicated a potential decrease of 16% (1.7min) in total transition time. This advantage is expected to increase as the number of actuators or control options in the switching order increases.

In this case study, the user retained control over switching actions and their timing. Learning allowed the system to anticipate the user's needs, streamlining the control interface. As learning progressed, the system was able to reduce the number of manual switching prompts required from the user. In 70% of all switching events in the testing data, the user's desired switching choice appeared first in the dynamically generated switching order, requiring only a single manual prompt to achieve control of the desired joint (Pilarski et al. 2012). In the remaining 30% of all switching events, the user's desired choice appeared second or third in the dynamic switching order, requiring one or more additional switching prompts to achieve control of the desired joint. Based on our observations about the average time needed for a user to deliver multiple switching prompts, increasing the first-choice suggestion accuracy to 100% would have led to an additional time savings of 1.47min. Thus, even with a perfect oracle for switching suggestions, transitions would still have occupied 7min (30%) of the 21min task time.

To further improve these savings and allow switching time to approach zero, the requirement for manual switching prompts would need to be reduced and eventually removed. As was suggested by our thought experiment on the intelligent semi-automatic transmission, this requires a learning system that is able to initiate switching actions in an autonomous or semi-autonomous way, and that can directly impact both the nature and timing of switching events. Our preliminary results indicate that switch timing can be predicted to a high degree of certainty using the same learning methods described in this case study. We therefore believe that it would be straightforward to shift the described experiments into a setting where the system takes on some degree of autonomy over the switching actions.

Taken as whole, our experiments with dynamic switching on the robot arm indicated a clear area of application for learning within the context of prompted switching. We are currently evaluating our dynamic switching approach with a population of upper-limb amputees, and are preparing to translate our techniques to a second problem domain involving a dexterous hand prosthesis with multiple grip patterns. In addition to time savings, we expect dynamic switching to reduce user frustration and cognitive load, and qualitative studies with users are underway. Future work will expand our amputee studies to include sliding autonomy with system-initiated switching.

Figure 5: Example of application switching on a Mac OS X desktop environment. Switching is a common aspect of our day-to-day human-computer interaction.

Between Instruction and Reward

In the work presented above, we explored one specific example of interactive machine learning in the middle ground between instruction and reward. Instead of trying to increase the number or magnitude of human interactions with the system during learning, we suggest that instruction can occur implicitly from human prompting within the sensorimotor stream. In other words, the action or lack of action by the human side of the human-machine interface provides implicit feedback. By combining prompted switching and learning, a human retains the ability to exert direct control over their assistive device. At the same time, the learning system is able to adapt its choices and actions to better meet the human's control intentions, without the need for training information or explicit direction from the user. The human's use or disuse of a proposed function or mode becomes a directive signal to the learner.

A similar approach is widely used in existing personal computing interfaces (Figure 5). Mobile and desktop computers include mechanisms to prioritize, promote, and highlight control options that are frequently or recently deployed by a user—e.g., applications, contacts, or menu options. Online learning of the kind described in the present work provides one clear way to extend interface adaptation mechanisms into the domain of real-time human-robot interaction, and to give existing techniques the ability to interpret context and human intent at greater levels of detail.

Prompted machine learning has the added advantage that human direction does not need to be demarcated into periods of training and execution. The amount of prompting required can also change over time, opening up the possibility for reducing the required direction (or getting more out of the same level of direction) as learning progresses. Put differently, the combination of interactive learning with prompting allows human input to shift over time with respect to the signalling continuum formed by bandwidth and explicitness (Figure 1). Prompted switching therefore facilitates a natural approach to sliding autonomy, incrementally reducing the cognitive and supervisory load on a human user.

The key message of the case study presented above is that prompted machine learning has powerful and immediate application to real-world problem domains. One readily accessible domain is video conferencing and robot gaze control, following the recent work of Denk (2012). Another domain is the task of efficiently switching between applications or

menus on a personal computer or mobile computing device, as discussed above (Figure 5). However, the potential utility of learning in a prompted switching setting is not limited to standard assistive tasks and robotic platforms. More whimsical examples include the human control of a distributed swarm of flying semiautonomous vehicles, actuation of a non-physiological prosthesis (for example, a robotic octopus arm or a telescoping gripper), and augmenting cognitive abilities with a tightly coupled memory or computation device. As such, we suggest that there is an exciting space of applied research to be found between the cases of instruction and reward; regions of this space may be immediately explored using a combination of switching and real-time machine learning. In addition, we believe prompted machine learning can be combined with existing interactive learning approaches for added benefit.

Conclusions

Artificial intelligence and machine learning provide the tools to amplify human capacity, bringing power and precision to our control of assistive technologies. The principal contribution of our work is to show that the middle ground between instruction and reward is an important territory for human-machine collaboration. To support this idea, we introduced a setting that balances between the extremes of full autonomy and full control: human-prompted switching. The combination of prompted switching with interactive machine learning works in practice to streamline the operation of a multi-joint robot arm, and is immediately applicable to other challenging real-world problems. Prompting also enables a natural form of sliding autonomy. Based on these observations, we believe that human-prompted switching—and other methods that occupy the space between instruction and reward—will soon yield powerful new ways for humans to direct their assistive machines.

Acknowledgements

The authors acknowledge support from Alberta Innovates – Technology Futures, the Alberta Innovates Centre for Machine Learning, and the Glenrose Rehabilitation Hospital Foundation. Thanks as well to Joseph Modayil for a number of helpful discussions relating to this work.

References

Chao, C., Cakmak, M., Thomaz, A. L. (2010). Transparent active learning for robots. In *Proc. of the 5th ACM/IEEE International Conference on Human-robot Interaction (HRI '10)*, 317–324.

Dawson, M. R., Fahimi, F., Carey, J. P. (2012). The development of a myoelectric training tool for above-elbow amputees. *The Open Biomedical Engineering Journal* 6: 5–15.

Denk, C. (2012). Reinforcement learning for robotic gaze control. M.Sc. diss., Department of Data Processing, Technische Universitaet Muenchen, Muenchen, Germany.

Judah, K., Roy, S., Fern, A., Dietterich, T. G. (2010). Reinforcement learning via practice and critique advice. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10)*, 481–486.

Kaplan, F., Oudeyer, P. Y., Kubinyi, E., Miklosi, A. (2002). Robotic clicker training. *Robotics and Autonomous Systems* 38: 197–206.

Knox, W. B., Stone, P. (2012). Reinforcement learning from simultaneous human and MDP reward. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (eds.), June 4–8, Valencia, Spain.

Lin, L.-J. (1991). Programming robots using reinforcement learning and teaching. In *Proceedings of AAAI-91*, 781–786.

Lin, L.-J. (1992). Self-improving reactive agents based on reinforcement learning, planning and teaching. *Machine Learning* 8: 293–321.

Lin, L.-J. (1993). Hierarchical learning of robot skills by reinforcement. In *Proceedings of the International Joint Conference on Neural Networks*, 181–186.

Micera, S., Carpaneto, J., Raspopovic, S. (2010). Control of hand prostheses using peripheral information. *IEEE Reviews in Biomedical Engineering* 3: 48–68.

Modayil, J., White, A., Sutton, R. S. (2012). Multi-timescale nexting in a reinforcement learning robot. In *Proceedings of the 2012 Conference on Simulation of Adaptive Behavior*.

Peerdeman, B., Boere, D., Witteveen, H., Huis in 't Veld, R., Hermens, H., Stramigioli, S., Rietman, H., Veltink, P., Misra, S. (2011). Myoelectric forearm prostheses: state of the art from a user-centered perspective. *Journal of Rehabilitation Research and Development* 48(6): 719–738.

Pilarski, P. M., Dawson, M. R., Degris, T., Fahimi, F., Carey, J. P., Sutton, R. S. (2011). Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *Proceedings of the 2011 IEEE International Conference on Rehabilitation Robotics (ICORR)*, June 29–July 1, Zurich, Switzerland, 134–140.

Pilarski, P. M., Dawson, M. R., Degris, T., Carey, J. P., Sutton, R. S. (2012). Dynamic switching and real-time machine learning for improved human control of assistive biomedical robots. In *Proceedings of the 4th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, June 24–27, Roma, Italy, 296–302.

Scheme, E., Englehart, K. B. (2011). Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use. *Journal of Rehabilitation Research and Development* 48(6): 643–660.

Sutton, R. S., Barto, A. (1998). *Reinforcement learning: an introduction*. Cambridge, Massachusetts: MIT Press.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., Precup, D. (2011). Horde: a scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, May 2–6, Taipei, Taiwan, 761–768.

Thomaz, A. L., Breazeal, C. (2008). Teachable robots: understanding human teaching behaviour to build more effective robot learners. *Artificial Intelligence* 172: 716–737.

Williams, T. W. (2011). Guest editorial: progress on stabilizing and controlling powered upper-limb prostheses. *Journal of Rehabilitation Research and Development* 48(6): ix–xix.