Designing Human-Centered Autonomous Agents

Gregory Dorais
NASA Ames Research Center
Moffett Field CA, 94035
gadorais@ptolemy.arc.nasa.gov
and
David Kortenkamp
NASA Johnson Space Center – ER2
Houston, TX 77058
kortenkamp@jsc.nasa.gov

1 Introduction

Human-centered automation (HCA) maximizes the goals of humans by supporting a full range of interactions between humans and autonomous systems. The key goal of this research is to minimize the *necessity* for human interaction, but maximize the *capability* to interact. Within HCA we define *adjustable autonomy* as the ability of autonomous systems to operate with dynamically varying levels of independence, intelligence and control. HCA encompasses adjustable autonomy and interfaces it with Human-Computer Interaction (HCI).

The motivations for human-centered automation are many. They include technical issues such as an inability to automate certain aspects of a task because of its complexity or because of an uncertain or changing operating environment. They also include non-technical issues such as a desire to allow for human intervention even when full autonomy is possible. These latter motivations may include safety, training, maintenance or calibration.

The benefits of human-centered automation include the ability to partially automate a system when full automation is not possible. Other benefits are lower costs because difficult-to-automate parts of the system can be left to humans and increased operator acceptance of an autonomous system.

Early work in human-centered autonomous systems has been conducted at NASA Johnson Space Center [2, 4], at NASA Ames Research Center [3], at the Honeywell Technology Center [5] and the University of Texas [1].

2 A simple example

To demonstrate adjustable autonomy, let's take the example of a tank that needs to be kept at a specific pressure. The tank has two ways of pressurizing: a motor that is controlled electrically and a crank that is turned manually. The tank has two sensors: an analog pressure gauge that must be read by a human and a digital pressure gauge that can be read by a computer. Finally, the system has a controller. The system is summarized as follows:

- Controller
  - human who decides whether pressure should be increased
  - computer that decides whether pressure should be increased
- Actuator
  - pump that human cranks
  - pump that motor activates increased
- Sensor
  - analog pressure sensor that human reads
  - digital pressure sensors that computer reads

This translates into eight different autonomy modes. For example, a fully autonomous system would use computer control with the motorized pump and the digital pressure sensor. A fully manual system would use the human controller, hand-cranked pump and analog pressure sensor. However, in addition to these two extremes there are six other partially autonomous configurations. For example, the computer could decide when the human should turn the crank by watching the digital pressure sensor. In this case, the computer is the controller and the human is the actuator. Or the human could read the analog pressure sensor and enter the value into the computer, which would do the control. From this simple example you can see that adjustable autonomy is the ability of the system to move amongst these different configurations as required.

3 Requirements

Building an adjustably autonomous control system to support HCA places severe requirements on the underlying control system. Without a properly designed control system, adjustable autonomy can be ineffective and even dangerous. It can be difficult to "retrofit" adjustable autonomy into existing autonomous or tele-operated control systems. There are several key requirements for a well-designed adjustably autonomous control system:

- A human (or other agent) who wants to adjust autonomy needs to know what the control system knows and what it is doing. This includes the following:
  - *State*: The state is a set of values that represent the current abstraction of the system (internal state) and its environment (external state). The human should be able to read and update the internal states of the control system and the control system's perceived state of the world.
  - *Models*: Models define the set of possible states and their relationships. Models should be presented in a way that is easily understood by humans.
  - *Goals*: A goal is a desired set of states. The user needs to know the system's current goals and its progress in achieving those goals. The system may need to explain non-linearities, e.g., backtracking.
  - *Tasks:* The tasks are the actions the system is currently taking to achieve its goals. The human needs to be able to see those tasks, adjust them and add to them if necessary.
- A system can be adjustably autonomous only if it can be commanded by an outside agent. Commanding can take many forms, including physical actuation, setting goals, changing models or executing procedures.
- Adjustable autonomy only applies when there are multiple methods (paths) to accomplish system tasks. If there are no choices then there is no autonomy to adjust.
- The human (or agent) that is adjusting autonomy must have knowledge of the capabilities of the other agent(s) and be able to recognize success and failure.
- The protocol for changing responsibility (or the level of autonomy) must be clear and must support both requesting a change in autonomy and accepting a change in autonomy.

4 Conclusions

We have developed the following list of questions that must be asked (and answered) when developing a human-centered autonomous system.

- What tasks can be done only by humans? Only by automation? By both?
- Who can set the level of autonomy for a task? Can the level of autonomy change at any time or only under certain circumstances?

- What are the timing issues with respect to a change in autonomy?
- Can an autonomy setting at one level of a hierarchical task be applied to all descendants?
- What are the possible autonomy level transitions? What transitions are not permitted?
- Is information necessary to control the system available to the user or to other agents?
- Are there multiple ways to accomplish the same task? Are they selectable by the user? By a planner?
- What parts of the system are commandable from the outside?
- How is success and failure of other agents recognized?

References

[1] Suzanne Barber, Anul Goel, and Cheryl Martin. The motivation for dynamic adaptive autonomy in agent-based systems. In *Proceedings of the First Asia-Pacific Conference on Intelligent Agent Technology (IAT '99*, 1999.

[2] R. Peter Bonasso, David Kortenkamp, and Troy Whitney. Using a robot control architecture to automate space shuttle operations. In *Proceedings of the 1997 Innovative Applications of Artificial Intelligence Conference*, 1997.

[3] Gregory Dorais, R. Peter Bonasso, David Kortenkamp, Barney Pell, and Debra Schreckenghost. Adjustable autonomy for human-centered autonomous systems on mars. In *Proceedings of the Mars Society Conference*, 1998.

[4] David Kortenkamp, Debra Keirn-Schreckenghost, and R. Peter Bonasso. Adjustable control autonomy for manned space flight. In *Proceedings of the IEEE Aerospace Conference*, 2000.

[5] David Musliner and Kurt Krebsbach. Adjustable autonomy in procedure control for refineries. In *Working Notes of the AAAI Spring Symposium on Agents with Adjustable Autonomy*, 1999.