Toward a Psychological Framework for Levels of Robot Autonomy in Human-Robot Interaction

Technical Report HFA-TR-1204

Atlanta, GA: Georgia Institute of Technology

School of Psychology – Human Factors and Aging Laboratory

JENAY M. BEER

ARTHUR D. FISK

WENDY A. ROGERS

Correspondence concerning this report should be addressed to Jenay M. Beer or Wendy A. Rogers, School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, 30332-0170. E-mail: smcbride@gatech.edu, wendy@gatech.edu.

Acknowledgments

This research was supported in part by a grant from the National Institutes of Health (National Institute on Aging) Grant P01 AG17211 under the auspices of the Center for Research and Education on Aging and Technology Enhancement (CREATE; www.create-center.org).

Table of Contents

1. Executive Summary

Service robots have the potential to help people in many different tasks and contexts. However, determining the aspects of a task a robot should perform is not an exact science. An important question to ask is “what should a robot do, and to what extent.” Consideration of robot autonomy, that is the extent to which it can carry out its own processes and operations without external control, is important because while the implementation of a robot may supplement a task it may also impose new demands on the user. Therefore, the way in which a human and robot interact varies along a continuum of autonomy, where a robot may vary from teleoperated to fully autonomous.

The purpose of this paper was to examine relevant research in primarily the fields of human-automation interaction and human-robot interaction (HRI), and create a framework of levels of robot autonomy in HRI. The field of robotics is a separate but related to automation. Throughout this investigation, the two fields were compared and contrasted, demonstrating that literature from both fields can certainly inform the other. In particular, two large literature reviews were conducted.

First, an in-depth multidisciplinary systematic review of the literature investigated how autonomy has been conceptualized and categorized in automation, robotics, and HRI. This review revealed many inconsistencies in the way the construct is considered specifically within the field of HRI. However the models and frameworks from the automation literature (i.e., models of levels of automation, LOA) can serve as a guide in conceptualizing robot autonomy.

Next, a review of the human-automation literature and HRI literature yielded a set of human-, robot-, interaction-, and task- related variables critical in understanding robot autonomy. These variables include acceptance, situation awareness, trust, robot intelligence, reliability, transparency, methods of control and social interaction. Additionally, task variables were identified that influence a robot's capability of functioning within a complex environment.

Finally, the knowledge gained from the two large literature reviews was used to develop a framework of levels of robot autonomy in HRI. This framework may be used for several functions. First, the framework provides guidelines for determining and categorizing robot autonomy along a 10-point taxonomy. Second, the framework incorporates the human-, robot-, interaction-, and task- related variables identified to be important in HRI along the autonomy continuum. These variables can be used as evaluative criteria to determine if the robot autonomy is appropriate for a given task. The strength of this investigation and proposed framework is the emphasis on HRI, specifically on psychological variables and their interaction with autonomy. Development of a psychological framework, as proposed in this paper, not only holds promise to conceptualize and better understand the construct of autonomy, but also to account for human cognitive and behavioral responses (e.g., situation awareness, workload, acceptance) within the context of human-robot interaction.

As a result of this investigation, research avenues in need of further attention by the HRI community were identified. HRI research is needed to identify appropriate trade-offs in allocating tasks to either a human or a robot. Efforts to research and understand robot autonomy will support the development of service robots that can be successfully implemented and support effective human-robot interaction.

2. Introduction

Autonomy: from Greek *autos* ("self,") and *nomos* ("law")
"*I am putting myself to the fullest possible use…*" –HAL 9000 (2001: Space Odyssey)

The primary focus of this paper is robot autonomy. One might ask, why focus on levels of autonomy for human robot interaction (HRI)? Developing fully autonomous robots has been a goal of roboticists and other visionaries since the emergence of the field, both in product development and science fiction. However, a focus on robot autonomy has scientific importance, beyond the pop culture goal of creating a machine that demonstrates some level of artificial free will.

The definition of autonomy will be thoroughly discussed later; however, it is interesting to note that the term autonomy has been used in many definitions of robots and agents (see a table of definitions and a word cloud in Appendix A). Inspection of these definitions demonstrates the importance of autonomy from the perspective of language use. However, autonomy is important beyond its inclusion in robot/agent definitions.

There are several reasons why considering autonomy is important. First, determining appropriate autonomy in a machine (robotic or otherwise) is not an exact science. An important question is not "what can a robot do", but rather "what should a robot do, and to what extent." Second, robot autonomy influences human robot interaction. Properly implementing robotics has promise of increasing human performance, whereas, inappropriate or unreliable function allocation to a machine often results in detrimental consequences in human situation awareness, trust, or acceptance. A scientific base of empirical research can guide designers in identifying appropriate trade-offs to determine which functions and tasks to allocate to either a human or a robot. For these reasons, autonomy is a central factor determining the effectiveness of the human-machine system. *Therefore, we propose that to understand robot autonomy is essential to understand human-robot interaction*.

A number of Human-Robot Interaction (HRI) frameworks exist, and consider autonomy as an influential construct (e.g., Feil-Seifer, Skinner, & Mataric, 2007; Goodrich & Olsen, 2003; Goodrich & Schultz, 2007; Huang et al., 2004; Kahn et al., 2007; Murphy & Woods, 2009; Thrun, 2004; Yanco & Drury, 2004a). However, these frameworks primarily discuss autonomy at a general level and fail to explore the construct as a central component that influences the very nature in which humans will interact with robots. Although the definition of "framework" implies only a loose collection of concept and principles (Salthouse, 1991), there is a need to consider autonomy in a more in-depth manner. To date, the construct of autonomy has not been reviewed or investigated in relation to other variables (e.g., trust, workload, reliance) known to also play a role in successful interaction between a human and a robot.

With that said, the field of HRI largely lacks frameworks and conceptual models that organize empirical observations to theoretical concepts related to human-robot interaction. Why propose specifically a *psychological* framework on levels of autonomy for human-robot interaction? Psychology plays a critical role in HRI by focusing on the human-side of the interaction. Development of a psychological framework, as proposed in this paper, not only holds promise to conceptualize and better understand the construct of autonomy, but also to account for human cognitive and behavioral responses (e.g., situation awareness, workload, acceptance) within the context of human-robot interaction.

2.1. Scope and Understanding Human-Robot Interaction for the Current Investigation

The field of HRI is a junction between multiple disciplines, notably psychology, computer science, and engineering. The field of HRI has been described as "dedicated to understanding, designing, and evaluating robotic systems for use by or with humans" (Goodrich & Schultz, 2007, p. 204). Because of the breadth of literature included in HRI, it is important to define the scope included in this investigation.

I will first consider what is meant by the term "robot." There is no agreed upon definition of robot. Joseph Engelberger, the father of robotics, once famously remarked "I can't define a robot, but I know one when I see one." This well-known quote exemplifies the abstract notion of "robot". Even though most individuals of the general population have never interacted with a robot directly, most people have ideas or definitions of what a robot should be like (Ezer, 2008). The term "robot" derives from the Czech word "robota" which translates to forced labor. Within the research community, a robot has often been broadly described as a machine that can sense, think/plan, and act (e.g., Bekey, 2005). This definition may be criticized as being too broad, as many types of technology could potentially be described as such. A synthesis of varying definitions of the term "robot" (i.e., Bekey, 2005; Murphy, 2000; Russell & Norvig, 2003; Sheridan, 1992) produces a more complex description: *A robot is a physical computational agent. Robots are equipped with sensors for perceiving the environment, and usually contain effectors (also called actuators) that manipulate and exert physical forces on the environment. Robots can either be stationary (anchored at fixed locations) or mobile.*

For this investigation, a certain type of robot will be of focus: professional and personal service robotics. Service robots can be described as "systems that function as smart, programmable tools, that can sense, think, and act to benefit or enable humans or extend/enhance human productivity" (Engelhardt & Edwards, 1992, p. 315-316). More specifically, service robots are designed to reside in the home or professional setting to assist people with personal or professional goals (Agah, 2001; Thrun, 2004).

Although this class of robots is still broad, it was chosen as the focus of this investigation for a number of reasons. First, service robots of varying degrees of autonomy have been applied to a wide range of applications, such as domestic or personal assistance, healthcare nursing tasks, search and rescue, and education. Second, due to the range of service applications human-robot interaction will often be necessary, and service robots may be expected to interact with humans with limited or no formal training (Thrun, 2004). Feasibly, in the future multiple robots may be applied to service domains (e.g., teams of robots, with one human supervisor). This investigation will primarily focus on human-robot interaction between one robot and one human. Human-multiagent interaction is a complex, albeit separate, problem space with a focus on team-related constructs (for review, see Ezer 2007). Furthermore, the large literature base of human-automation interaction will be informative in this investigation, and those interactions generally involve one human and one automated system.

A parallel focus on human-automation interaction brings forth the questions *how is a robot similar/dissimilar to automation* and *why investigate both robotics and automation?* Automation is most often defined as a "device or systems that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator" (Parasuraman, Sheridan, & Wickens, 2000). In the Handbook of Automation (Nof, 2009), a robot is considered a subcategory of automation. However, we disagree. Although many robots in use have been little more than automated machines (e.g., industrial robots), as service robots increase in interactive capabilities a simplistic view of robots would provide little value for motivating the advancement of such collaborative agents. Capabilities such as mobility, environmental manipulation, and social interaction separate robots from automation in both function and physical form. Indeed, the definitions of robot and automation may be different, but the lines separating "automation" from "robotic" are also blurred. The goal here is not to redefine robot or automation, rather simply to depict that robots are a technology class of their own, separate but related to automation.

Automation researchers have a long history of studying and understanding human-related variables. Service robots, similar to automation, will be required to interact with humans. Therefore, from a psychological perspective, the most important common denominator between these two technologies is the "human-machine interaction." Unfortunately, research and development of automation and robots have, in large part, occurred separately as different fields. As we will demonstrate in this current paper, the literature from both fields can certainly inform one another.

2.2. Goals and Importance of Current Investigation

Autonomy has been defined, measured, and conceptualized in a variety of ways across several different disciplines (i.e., HRI, robotics, automation, and psychology). In the current investigation, the overarching goal not only is to parse the literature to better understand what autonomy is, but also to critique autonomy from a psychological perspective and identify variables that influence – and are influenced by – autonomy. To meet this overarching goal, the current investigation will integrate and synthesize the robotics and HRI literature, in addition to the human automation literature, an established literature base in the field of human factors. Specifically, the objectives of this paper are to:

1. refine the definition of autonomy and analyze how the construct has been conceptualized in automation, robotics, and HRI;
2. propose a process of determining level of robot autonomy, that may be used by developers in determining the appropriate level of robot autonomy for a given environment/task;
3. suggest a psychological framework within the context of HRI that identifies potential human, robot, and task variables related to autonomy; and
4. finally, identify areas of HRI that need further research.

3. Autonomy as a Construct that Influences Human-“Machine” Interaction

The first step in this investigation is to consider autonomy as a construct that influences human-machine interaction. Autonomy has been defined, measured, and considered in a variety of ways across several different disciplines (i.e., psychology, automation, robotics, and HRI). In this section, the first objective is to review how this term is defined in the current literature, and to propose a refined definition of autonomy for this investigation. The second objective is to assess how autonomy has been conceptualized in automation, robotics, and HRI. This assessment will include a review, critique, and synthesis of the models, frameworks, and taxonomies related to autonomy for each of the literature bodies. Such an assessment of autonomy is critical in identifying similarities and differences between these fields and to answer ‘*what exactly is autonomy?*’

3.1. The Construct of Autonomy: Terminology and Definitions

To effectively investigate autonomy and its role in HRI, it is critical to understand the characteristics that define the term *autonomy*. It is also critical to determine whether the term autonomy is synonymous across the fields of psychology, automation, and robotics, or if there are differing ways in which this term is used.

3.1.1. What is autonomy? Autonomy has been a construct of both philosophical and psychological interest for over 300 years. In the 18th century, autonomy was most famously considered by philosopher Immanual Kant as a moral action determined by a person’s free will (Kant, 1967). Early psychology behaviorists (e.g., Skinner, 1978) claimed that humans do not act out of free will, rather their behavior is in response to stimuli in the environment. However, in psychology, autonomy has been primarily discussed in relation to child development. In this literature, the term autonomy is discussed as a subjective construct involving self-control, governing, and free will. For instance, Piaget (1932) proposed that autonomy is the ability to self govern, and a critical component in a child's moral development. Erikson (1997) similarly defined autonomy as a child's development of a sense of self control (e.g., early childhood toilet training). Children who successfully develop autonomy feel secure and confident. Whereas, those children who do not develop autonomy may experience self-doubt and shame.

Autonomy as a construct representing free will only encompasses one way in which the term is used. The phenomenon of *psychological* autonomy (and the underlying variables) is different than the phenomenon of *artificial* autonomy that engineers would like to construct in machines and technology (Ziemke, 2008). For instance when the term autonomy is applied to technology, particularly automation, it is discussed in terms of autonomous function (e.g., performing aspects of a task without human intervention). Although the specific term "autonomy" is not commonly used in the automation literature, some models (Parasuraman, Sheridan, & Wickens, 2000; Sheridan & Verplank, 1978) describe higher levels of automation possessing "increased autonomy of computer over human action" (Parasuraman, Sheridan, & Wickens, 2000, p. 287). In this sense, autonomy represents trading control of authority between the human and automation. Parasuraman and colleagues provided the example of some types of automation exhibiting autonomy over decision making. That is, in this example the task of decision making is allocated to the automated system, giving that system the authority (i.e., autonomy) to perform that aspect of the task without human intervention.

How is autonomy addressed in the field of robotics? Robot autonomy has been discussed in the literature as both as a psychological construct as well as an engineering construct. In fact, the term is used to describe many different aspects of robotics, from the robot's ability to self-govern much like a human, to the level of human intervention. To help parse the use of this term, various definitions of robot autonomy are examined (

Table 1).

Table 1

Definitions of Autonomy Found in Robotics Literature

Definitions of Agent and Robot Autonomy	
"The robot should be able to **carry out its actions** and to **refine or modify** the task and **its own behavior** according to the current **goal** and execution **context of its task**."	Alami, Chatila, Fleury, Ghallab, & Ingrand, 1998, p. 316
"Autonomy refers to systems capable of **operating in the real-world environment without any form of external control** for extended periods of time."	Bekey, 2005, p.1
"An autonomous agent is a system situated within and a part of an environment that **sense that environment and acts on it**, over time, in pursuit of its **own agenda** and so as to effect what it senses in the future." "**Exercises control** over its own **actions.**"	Franklin & Graesser, 1997, p. 25.
"An Unmanned System's **own ability** of sensing, perceiving, analyzing, communicating, planning, decision making, and acting, **to achieve goals** as assigned by its human operator(s) through designed HRI" "The condition or quality of being **self-governing**."	Huang, 2004, p. 9
" "Function autonomously" indicates that the robot can operate, **self-contained**, under all reasonable conditions **without requiring recourse** to a human operator. Autonomy means that a robot can **adapt to change** in its **environment** (the lights get turned off) or itself (a part breaks) and continue to reach a **goal**."	Murphy, 2000, p. 4
"A rational agent should be autonomous – it should **learn** what it can **to compensate for partial or incorrect prior knowledge**."	Russell & Norvig, 2003, p.37
"Autonomy refers to a robot's **ability to accommodate variations in its environment**. Different robots exhibit different degrees of autonomy; the degree of autonomy is often measured by relating the degree at which the environment can be varied to the **mean time between failures**, and other factors indicative of robot performance."	Thrun, 2004, p.14
"Autonomy: agents operate **without the direct intervention of humans** or others, and have some kind of **control** over their **actions and internal states**."	Wooldridge & Jennings, 1995, p.116

Note: Emphasis (bold) added.

The muddled use of this term reflects the multi-disciplinary nature of the field. Some researchers in cognitive science, artificial intelligence, and computer science posited that autonomy is a characteristic of modeling a human-like or a socially interactive (Feil-Seifer, Skinner, & Mataric, 2007) machine. As depicted in Table 1, others in engineering and computer science defined the term in relation to objective and/or functional characteristics. There are several characteristics covered across the definitions listed above.

First, an overarching characteristic was that a robot demonstrating autonomy should refine or modify its own *operations*. Operation, throughout these definitions, referred to actions and behaviors (Alami et al., 1998), as well as internal states such as sensing, perceiving, and learning (e.g., Huang, 2004; Russell & Norvig, 2003; Wooldridge & Jennings, 1995).

Second, in many of the definitions, autonomy was described as performing those actions within the context of the *environment*. The environment poses a large set of unknown or unpredictable factors that may influence the robot's performance. The robot's ability to refine and modify its actions in response to unpredictable environmental stimuli (e.g., path planning, obstacle avoidance), has been suggested as a form of autonomy. For instance, it has been suggested (e.g., Thrun, 2004) that service robots require a high level of autonomy due to the unpredictable and changing environment of a home or healthcare setting.

Third, the concept of *goals* was included in a number of the definitions in Table 1. The term *goal* has not generally been used in the automation literature. However, the concept of creating *goal* maps roughly to the automation literature's term *information analysis* (Parasuraman, Sheridan, & Wickens, 2000; also referred to as *generating*; Endsley & Kaber, 1999). How and who determines a robot's goal (the robot or the human) has been inconsistently considered in the robotics literature. For example, the Huang (2004) definition explicitly stated that the human operator determines the goal of the system. Whereas, Franklin and Graesser (1997) explicitly stated that the robot should pursue its own agenda.

Finally, the concept of *control* was also a theme, albeit inconsistently used. For instance, a number of the definitions clearly stated that the robot should perform with no human intervention or external control (e.g., Bekey, 2005; Murphy, 2000; Wooldridge & Jennings, 1995). However if the human sets the robot goal, as suggested by Huang, then some level of human involvement must be needed. It is noted that no single definition of autonomy incorporates all of the important characteristics identified above. Therefore, it is crucial to clarify and redefine the term. Proposing a new definition ensures scientific process by highlighting the important characteristics of the term, as well as maintaining an inclusive meaning of the construct while assessing it from multiple disciplines.

3.1.2. Defining autonomy for this paper. In an attempt to clarify the term autonomy we will propose new definitions. First, a weak, or global, definition of autonomy is proposed as the following: *the extent to which a system can carry out its own processes and operations without external control.* This weak definition of autonomy can be used to denote autonomous capabilities of humans, automation, or robotics.

However, a stronger more specific definition can be given to agents (e.g., robots), based upon an attempt to integrate definitions and most prevalent characteristics related to robot autonomy provided in Table 1. Because autonomy will be discussed in this paper within the context of robotics as an assistive *technology*, the proposed definition of autonomy focuses on functional characteristics rather than the abstract notion of "free will". Autonomy, as related to robotic agents, is defined in this paper as:

The extent to which a robot can ***sense*** *the environment,* ***plan*** *based on that environment, and* ***act*** *upon that environment, with the intent of reaching some* ***goal*** *(either given to or created by the robot) with little or no external* ***control****.*

Note that both the weak and strong definition begin with the phrase “to the extent to which…” My choice in wording exemplifies that autonomy is not all or nothing. Autonomy exists on a continuum (from no autonomy to full autonomy). The proposed strong definition of autonomy attempts to integrate the current definitions of autonomy, and highlight the prevalent characteristics of autonomy (i.e., sense, plan/decide, act, goal, and control).

Now that a working definition of autonomy has been established, we will next explore the construct of autonomy and how it affects human-machine interaction by taking a multi-disciplinary approach. A review of how autonomy has been conceptualized within the fields of automation, robotics, and HRI will be discussed.

3.2. Autonomy and its Relevance to Automation

In this next section, autonomy is considered within the context of human-automation interaction. In particular, autonomy, as encompassed in levels of automation, is reviewed with regard to function allocation between the automated machine and the human.

3.2.1. Terminology. The term autonomy, as used in automation, generally refers to the amount of human intervention, the automation’s functionality, as well as function allocation (authority) between the human and machine. Terminology associated with automation autonomy is often unique to individual frameworks and taxonomies related to levels of automation. There are several terms commonly used in the literature:

Degree: Refers to the extent to which or how much of a task is being performed by the automation. For example, the degree of automation may range from the human performing the task (manual control) to the automation performs the entire task (full automation) and anything in between (combination of human and automated control).

Scale: Describes the stage of processing that system functions automate. For example, according to Parasuraman and colleagues (2000), automation can be applied to four classes of functions: information acquisition, information analysis, decision and action selection, and action implementation.

Type: Generally describes the system as a whole based on its purpose. For example, warning systems are considered a type of automation. Sometimes, the term type is synonymous with the scale or stage of processing (Parasuraman, Sheridan, & Wickens, 2000).

Levels: Commonly refers to the combination of degree and scale of automation. The individual levels are often used to categorize the assignment of functions to the human or automation or a combination of the two.

3.2.2. How autonomy has been conceptualized in automation. Various taxonomies, classification systems, and models related to levels of automation (LOA) have been proposed. Automation has been defined as full or partial replacement of a function previously carried out by a human (Parasuraman, Sheridan, & Wickens, 2000). This implies that automating tasks is not all or nothing. Rather, automation exists on a continuum, ranging from manual control to full automation. The earliest categorization scheme, which organizes automation along both degree and scale, was proposed by Sheridan and Verplank (1978). This 10-point scale categorized higher levels of automation as representing increased autonomy, and lower levels as decreased autonomy. The 10-point scale is listed in **Table 2**.

This taxonomy specified what information is communicated to the human (feedback) as well as allocation of function split between the human and automation. However, the scale used in this early taxonomy was limited. That is, the taxonomy specified a set of discernible points along the continuum of automation applied primarily to the *output* functions of decision and action selection. It failed to specifically address *input* functions related to information acquisition (i.e., sensing) or the processing of that information (i.e., formulating options or strategies).

Table 2
Sheridan and Verplank (1978) Levels of Decision Making Automation

Level of Automation	Allocation of Function
1	The computer offers no assistance; the human must take all decisions and actions.
2	The computer offers no assistance; the human must take all decisions and actions.
3	The computer offers a complete set of decision/action alternatives, or
4	Narrows the selection down to a few, or
5	Suggests one alternative
6	Executes that suggestion if the human operator approves, or
7	Allows the human a restricted time to veto before automatic execution, or
8	Executes automatically, then necessarily informs the human, and
9	Informs the human only if asked, or
10	Informs the human only if it, the computer, decides to

Endsley and Kaber (1999) later proposed a revised taxonomy. This taxonomy built on the work of Sheridan and Verplank, but with greater specificity on *input* functions such as how the automation acquires information and formulates options.

The strength of the Ensley and Kaber model is the detail used to describe each of the automation levels. The taxonomy is organized according to four generic functions which include: (1) *monitoring* – scanning displays; (2) *generating* – formulating options or strategies to meet goals; (3) *selecting* – deciding upon an option or strategy; and (4) *implementing* – acting out chosen option. The taxonomy specifying 10 levels of automation and each levels description is shown in **Table 3**.

Table 3

Endsley and Kaber (1999) Levels of Automation

Level of Automation	Description
1. Manual Control:	the human monitors, generates options, selects options (makes decisions) and physically carries out options.
2. Action Support:	the automation assists human with execution of selected action. The human does perform some control actions.
3. Batch Processing:	the human generates and selects options then they are turned over to automation to be carried out (e.g., cruise control in automobiles).
4. Shared Control:	both the human and the automation generate possible decision options. The human has control of selecting which options to implement; however, carrying out the options is a shared task.
5. Decision Support:	the automation generates decision options that the human can select. Once an option is selected the automation implements it.
6. Blended Decision Making:	the automation generates an option, selects it and executes it if they human consents. The human may approve of the option selected by the automation, select another or generate another option.
7. Rigid System:	the automation provides a set of options and the human has to select one of them. Once selected the automation carries out the function.
8. Automated Decision Making:	the automation selects and carries out an option. The human can have input in the alternatives generated by the automation.
9. Supervisory Control:	the automation generates options, selects and carries out a desired option. The human monitors the system and intervenes if needed (in which case the level of automation becomes Decision Support).
10. Full Automation:	the system carries out all actions.

Parasuraman, Sheridan, and Wickens (2000) proposed the most recent model for types and levels of automation (Figure 1). Similar to previous taxonomies, the authors stated that functions can be automated to differing degrees along a continuum of low to high (i.e., fully manual to fully automated). However, they argued that the previous taxonomies focused on output (e.g., decision and action), rather than input (e.g., sensing, gathering information). Therefore, this model proposed classes of functions, referred to as stages or types of automation, which cover both input and output functions. These types included: (1) *information acquisition*; (2) *information analysis*; (3) *decision and action selection*; and (4) *action implementation*.

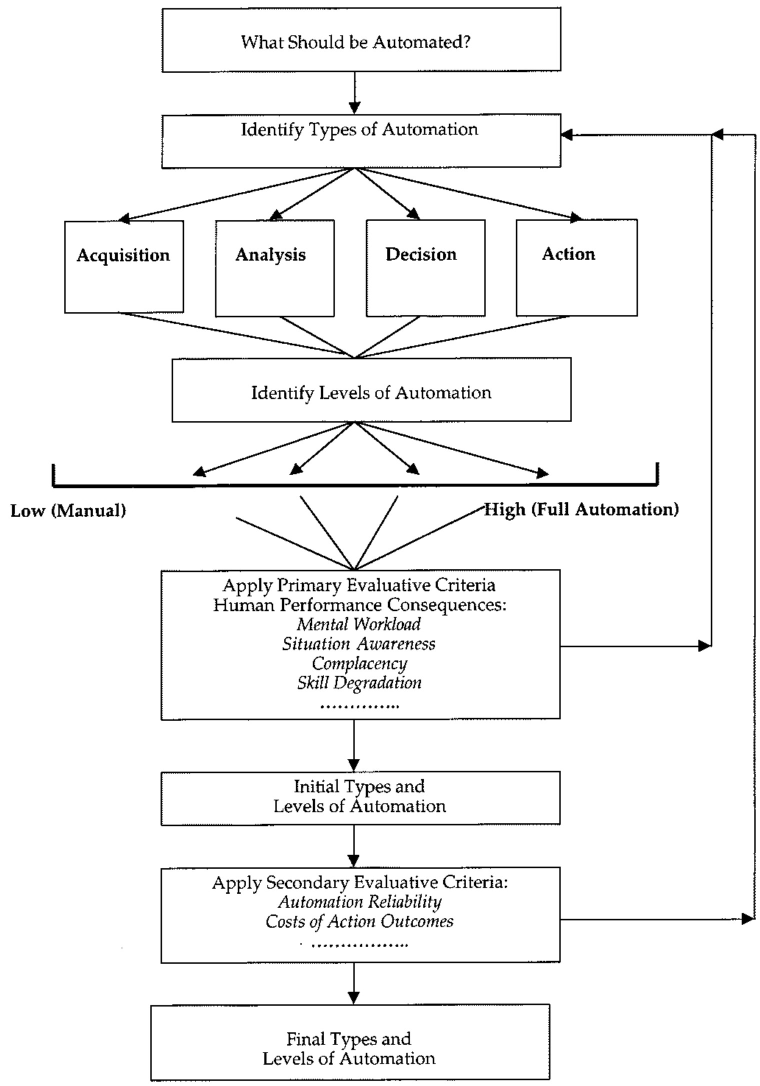


Figure 1. Flow chart showing application of the model of types and levels of automation (Parasurman, Sheridan, & Wickens, 2000).

Automation categorized under the *information acquisition* stage support processes related to sensing and registering input data. This stage of automation supports human sensory and perceptual processes, such as assisting humans with monitoring environmental factors. Automation in this stage may include systems that scan and observe the environment (e.g., radars, infrared goggles). At higher levels of information acquisition automation may organize sensory information (e.g., in air traffic control an automated system which prioritizes aircraft for handling). The *information analysis* stage refers to automation that performs tasks similar to human cognitive function, such as working memory. Automation in this stage may provide predictions, integration of multiple input values, or summarization of data to the user. Automation in the *information analysis* stage is different from automation in the *information acquisition* phase, in that the information is manipulated and assessed in some way.

Automation included in the *decision selection* stage selects from decision alternatives. For example, automation in this stage may provide navigational routes for aircraft to avoid inclement weather, or recommend diagnoses for medical doctors. Finally, *action implementation* automation refers to automation that actually executes the chosen action. In this stage, automation may complete all, or subparts, of a task. For example, action automation may include the automatic stapler in a photocopy machine, or autopilot in an aircraft.

The bottom of the flow chart (Figure 1) depicts primary and secondary evaluative criteria. These evaluative criteria were meant to provide a guide for determining a system's level of automation. In other words, the purpose of the Parasuraman and colleagues' model was to provide an objective basis for making the choice on to what extent a task should be automated. To do this, the authors proposed that an evaluation of the consequences of both the human operator and the automation. Therefore, first primary evaluative criteria are to be evaluated (e.g., workload, situation awareness), and then the level of automation is adjusted. Next the secondary criteria are evaluated (e.g., automation reliability, cost of action outcomes), and again the level of automation is adjusted. The authors proposed that this iterative process may be a starting point for determining the appropriate types of levels of automation should be implemented in a particular system.

3.2.3. Critique of autonomy and automation. Inspection of the various models and taxonomies for levels of automation reveal number of insights can be made. First, comparing and contrasting Sheridan and Verplank's (1978) and Endsley and Kaber's (1999) 10 levels of autonomy, the taxonomies seem to emphasize different stages and aspects of automation. As pointed out by other reviews (Durso et al., 2011), the Sheridan and Verplank taxonomy primarily makes distinctions on the higher levels of automation and focus on details related to *output* functions and stages (i.e., decision selection). Endsley and Kaber, however, make fine grained distinctions on the *input* functions with details on the human or automation generating of options and strategies.

Comparisons between the Endsley and Kaber (1999) model and the Parasuraman, Sheridan, and Wickens (2000) model suggests that the types/stages of automation roughly correspond, with *information acquisition* similar to *monitoring*; *information analysis* similar to *generating*; *decision and action selection* similar to *selecting*; and finally *action implementation* similar to *implementing* (Figure 2). Yet, the equivalence between these stages and the functions each comprise is not one to one. For example, Endsley and Kaber described monitoring as "scanning system displays to perceive system status" (p. 464). It was not explicitly stated how the information provided on those displays is actually acquired. Whereas, Parasuraman more clearly described their information acquisition stage as using sensors to collect and register input data. Furthermore, in Endsley and Kaber's taxonomy, monitoring is not explicitly stated in most of the level descriptions, creating further ambiguity as to how the stage is implemented.

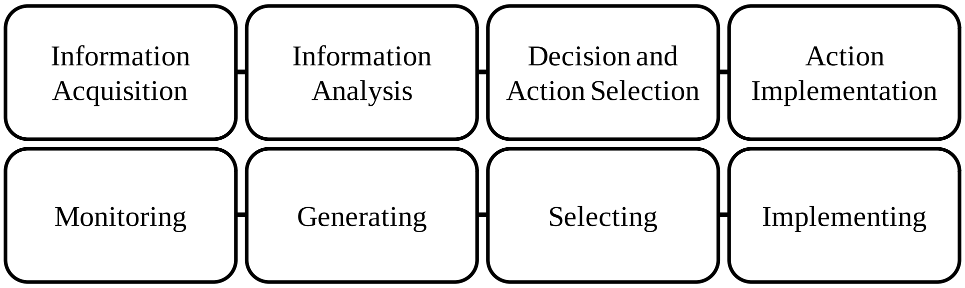


Figure 2. Comparison of Parasuraman, Sheridan, and Wickens (2000) and Endsley and Kaber (1999) four stage models.

Additionally, Endsley and Kabers' *generation* stage appears to comprise a mix of Parasurman, Sheridan, and Wickens' *information analysis* and *decision selection* stages. For example, both generating and information analysis involve algorithms applied to incoming data to predict and integrate input variables to augment human cognitive functions. However, Endlsey and Kaber stated that generation also involves generating possible actions, which more closely resembles Parasurman and colleagues' decision and action selection (also see Durso et al., 2011 for discussion).

Finally, in my critique the most recent model, Parasurman, Sheridan, and Wickens (2000), a number of strengths and weaknesses are evident. The model (Figure 1) highlights primary and secondary criterion for assessing the consequences of applying any given level of automation. The primary criterion in particular, looks at human-related variables affected by the automation. This is critical, because the model takes a human-automation interaction approach, an approach not nearly as explicitly highlighted in the previous models. On the other hand, the model can be criticized for its lack of specificity for intermediate levels of automation. What constitutes a medium level of automation? That is, the authors note that automation may fall on a range from low to high (i.e., fully manual to fully automated) with little detail as to what lays in between for each of the four stages. The model, while useful, is relatively open, making it difficult to make distinctions between detailed system characteristics along this continuum.

Whatever the similarities or differences, these models are useful. Each provides an organizational framework in which to categorize not only the purpose or function of the automation (e.g., stages), but also considers automation along a continuum of autonomy. These models are important to consider within the context of both robotics and HRI, because they can serve as a springboard for development of similar taxonomies and models specific to robot autonomy. In fact, these models, in particular the Sheridan and Verplank's taxonomy, have been suggested as appropriate to describe how autonomous a robot is (Goodrich & Schultz, 2007). However, it is important to consider the differences between automation and robotics, as previously outlined in this paper. Robots may serve different functions as automation; for example, some robots may play a social role; social ability is not a construct considered in the LOA models and taxonomies. From a human robot interaction emphasis, a complimentary way to think about how these taxonomies could relate to HRI is to consider the degree to which the human and robot interact, and to what extent each can act autonomously. The next sections address how autonomy has been applied to robotics, and how autonomy's conceptualization in robotics is similar or different from automation.

3.3. Autonomy and its Relevance to Robotics (Engineering / Computer Science)

In this next section, a literature review was conducted assessing how autonomy has been applied to robotics, and how autonomy's conceptualization in robotics is similar or different from automation.

3.3.1. Terminology. Similar to automation, autonomy and levels of robot autonomous capability is a widely considered construct. Autonomy, in the robotics literature, is also depicted as a continuum. The continuum is most often described as ranging from teleoperated to autonomous robots. However the terminology used within the robotics field differs from automation, and specific terminology relates to the paradigms and architectures which characterize an approach to programming autonomy into a robot. There are several common terms found in the literature:

Paradigm: Refers to a philosophy or set of assumptions that describe an approach to solve a class of problems. There may be a number of paradigms suited for solving a problem, and no one paradigm may be consider right. This can be thought of as analogous to solving for the roots of a quadratic function; one could solve the problem geometrically by graphing the function, or algebraically by factoring the function. Therefore, if the goal is to develop robot intelligence, programming autonomy would be one potential paradigm for reaching that goal.

Software architecture: Sometimes simply referred to as ‘architecture’ or ‘robot architecture’ describes the building blocks for programming a robot. Software architectures usually include the tools and languages for writing programs, and provide a principled way of organizing components of a system. Consider the following example: when designing a house, most houses share the same architectural components (e.g., bedrooms, bathrooms, kitchen, general living space). These components can be organized in different ways (i.e., architecture), even though all the designs follow the same ‘designing a house’ paradigm (Murphy, 2000).

SENSE: Describes a set or type of robot function that relates to taking in information from the environment (e.g., from sensors).

PLAN: Describes a set or type of robot function that relates to use sensory information obtained from the sense function and developing one or more tasks or goals for the robot to perform. Plan often is referenced in relation to programming robot behavior algorithms, and may be considered as a subset of the overall robot’s artificial intelligence. Sometimes, plan is referred to as *think*.

ACT: Refers to output motor actuator commands. In other words, act specifies motor behavior performed by the robot.

3.3.2. How autonomy has been conceptualized in robotics. Autonomy has been conceptualized in robotics with a variety of approaches housed in domains such as artificial intelligence (also referred to as AI robotics), control theory, and cognitive science. One common way in which robot autonomy has been approached is through the application of SENSE, PLAN, ACT primitives. The *sense-plan-act* model of decision making (Murphy, 2000), follows a sequential set of processes. First the robot uses sensors to collect data of its environment. Next, the robot uses that data to plan the directives needed to reach some goal. Finally, the robot acts to carry out a directive. After this sequence is complete, the robot repeats the process: SENSE → PLAN → ACT.

The *sense-plan-act* model may be criticized as a gross simplification of the processes required for a robot to perform at any level of autonomy. Similarly, Parasuraman and colleagues (2000) simplify of the components human information processing as applied to automation stages. However, the goal is not to mimic the complex theoretical structure of the human cognitive system. Rather, both approaches (i.e., automation stages and the SENSE, PLAN, ACT primitives) propose an autonomy structure that is useful in practice.

In fact, the stages of automation and the sense-plan-act model can be compared. As shown in Figure 3, my comparison depicts how the sense, plan, and act primitives map onto the Parasuraman and colleagues (2000) stages of automation. That is, the sense primitive corresponds to the information acquisition stage, with both terms relating to the sensing and registering of sensory data. The plan primitive corresponds with both the information analysis and decision and action selection stages. In this primitive, the sensory information collected in the sense phase, is used to develop directives/goals/decisions for the robot to carry out. Finally, the act primitive maps onto the action implementation phase of automation, with both terms referring to the actual execution of the action choice.

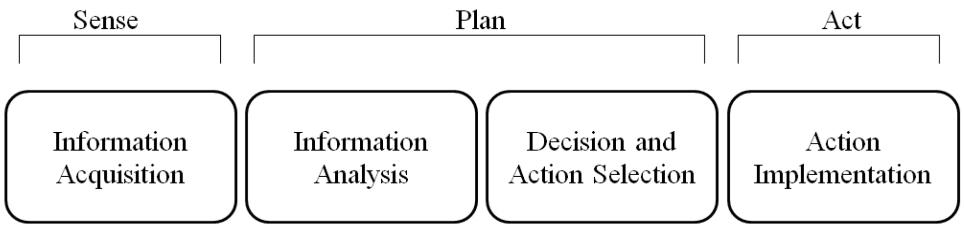


Figure 3. Comparison of SENSE, PLAN, ACT robot primitives and Parasuraman, Sheridan, and Wickens (2000) stages of automation.

The *sense-plan-act* model was popular in the 1960s, 1970s, and 1980s. However, in the late 1980's another approach to robot autonomy was introduced (Murphy, 2000). This approach, sometimes called behavior-based robotics (Arkin, 1998), is characterized by low level sensor-action loop. That is, the plan primitive is completely discarded. This approach reflects the behaviorist approach (e.g., Skinner, 1978) where sensory input directly results in behavioral output. This decision cycle occurs in the order of milliseconds (Russell & Norvig, 2003), and a robot could run multiple instances of SENSE-ACT couplings. Behavior-based robotics, while it revolutionized the field of robotics, was limited in programming robots to process information in a similar manner as human cognition.

Today, the most recent robot architectures use reactive techniques at low levels of robot functioning (SENSE→ACT) with deliberate techniques (PLAN) at the higher levels. Architectures that combine these techniques are usually called *hybrid architectures* (Russell & Norvig, 2003). Hybrid architectures, claimed to be a major breakthrough in robot autonomy (Goodrich & Schultz, 2007), simultaneously allow for a robot to perform reactive behaviors along with high-level cognitive reasoning about goals and plans. Under this architecture, a mixture of the *sense-plan-act* model and behavior-based robotics is applied, with PLAN being implemented at one step, while simultaneously SENSE→ACT is done at another. The three approaches to robot autonomy, and their relative sequencing of SENSE, PLAN, and ACT primitives is depicted in Figure 4.

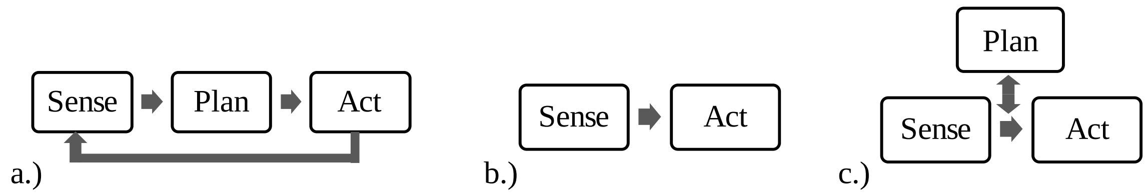


Figure 4. The three approaches to robot autonomy a.) *sense-plan-act* model; b.) behavior-based robotics; c.) hybrid architecture. (Figure adopted from Murphy, 2000).

3.3.3.   Critique of autonomy and robotics.   The primitives SENSE, PLAN, and ACT have been used in software architectures as a way to implement autonomy in robotics. These primitives map onto automation's four stage models. For example, comparing these primitives to the Parasuraman, Sheridan, and Wicken's (2000) levels of automation model, *sense = information acquisition, plan = information analysis* and *decision selection, act = action implementation* (a similar comparison can be made with Endsley & Kaber, 1999 terminology). Although this mapping is not one to one (e.g., *plan* encompasses two automation stages), it is useful to observe this relationship. The way in which autonomy is conceptualized for robots and automation can be considered as relatively equivalent.

However, unlike the automation models, the human's role is not considered in each of these robot software architectures. The software architecture models focus on programming robot autonomy. That is, this section provides a high level description of the 'behind the scenes' processes required for a robot to perform autonomously. These architectures are not meant to be models that specify how the robot's autonomous behavior will affect the world it interacts with (which may include humans). To further investigate the relationship between robot autonomy (as implemented by software architectures) and the humans a robot may interact with, the next section will investigate autonomy and its relevance to human-robot interaction.

3.4. Autonomy and its Relevance to Human-Robot Interaction

In this next section, autonomy is considered within the context of human-robot interaction. In particular, HRI was reviewed with emphasis on frameworks and models describing the role between the human and robot, and how autonomy influences that interaction.

3.4.1. Terminology. Autonomy, within the context of human robot interaction, is discussed with a large variety of terminology. Because of the multi-disciplinary nature of the field, some language is similar to the automation and robotics fields, and other language is unique to HRI alone. Two terms unique to HRI are *intervention* and *interaction*. The differences between these two terms are blurred, but a description of how they are generally used is described below. Note that these terms, in relation to autonomy, are often synonymous.

Intervention: sometimes refers to the amount of time the human is controlling the robot (e.g., Yanco & Drury, 2004a). This term sometimes also refers to the frequency of unplanned action or input by the human to help the robot complete task (e.g., Huang, Messina, Wade, English, Novak, & Albus, 2004).

Interaction: this term, specifically related to autonomy, has sometimes been used to describe general engagement between the human and robot. Another use of this term is a switch in human attention from a secondary task to the robot (usually to intervene in robot's performance in some way; e.g., Goodrich & Olsen, 2003). The difference between *interaction* and *intervention* is not always made clear, and the terms are sometimes used interchangeably.

The continuum of autonomy is often discussed as "levels" of autonomy, opposed to the automation literature, which often deconstructs the levels into degree and type. Some HRI researchers discuss this continuum with similar terms as automation (e.g., manual control/teleoperation to full autonomy; Anderson, 1996; Yanco & Drury, 2004a). Conversely, other HRI researchers describe the autonomy continuum from the perspective of the interaction between the human and robot (e.g., Goodrich & Schultz, 2007; Milgram, Rastogi, & Grodski, 1995). For example, higher autonomy levels sometimes referred to the robot as "teammate" or "peer," illustrating the *role* of the robot in relation to the human rather than its autonomous capability. Intermediate levels of robot autonomy are referenced as interaction strategies between the robot and human (e.g., "supervisory control" or "collaborative control" or "peer to peer collaboration"; Figure 5).

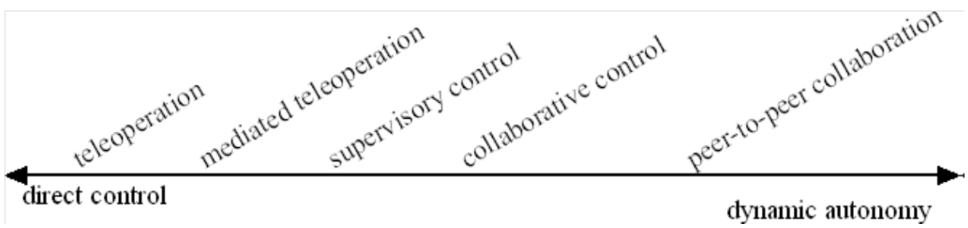


Figure 5. Continuum of levels of robot autonomy with an emphasis on human interaction (Goodrich & Schultz, 2007)

3.4.2. How autonomy has been conceptualized in human-robot interaction. As this section will review, autonomy has been included in a number of frameworks and HRI reviews. Interestingly, autonomy within an HRI context is a widely considered construct; however, the ideas surrounding how autonomy influences human-robot interaction are varied. Through analysis of the literature, it becomes clear that there are two schools of thought regarding how autonomy relates to human-robot interaction. The dichotomous viewpoints are: (1) higher robot autonomy involves *lower levels or less frequent HRI*; and (2) higher robot autonomy requires *higher levels or more sophisticated forms of HRI*.

The first viewpoint, that higher autonomy requires less HRI, has namely been proposed by Huang and colleagues (Huang, Messina, Wade, English, Novak, & Albus, 2004; Huang, Pavek, Albus, & Messina, 2005; Huang, Pavek, Novak, Albus & Messina, 2005; Huang, Pavek, Ragon, Jones, Messina, & Albus, 2007). The ongoing goal of this research group is to develop a framework for autonomy, and metrics used to measure robot autonomy. Although this framework is used primarily for robots used in military applications, the general framework has been cited as a basis for HRI autonomy classes (Yanco & Drury, 2004a). In this framework, metrics for autonomy are based on (1) task/mission complexity; (2) environmental difficulty; (3) human interaction/interface. The robot's autonomy is determined along certain levels along the three axes depicted in Figure 6. This conceptual model depicts the multi-dimensional nature of autonomy, but the authors have yet to determine exactly how to compute the overall autonomy level (i.e., average of scores along three axis).

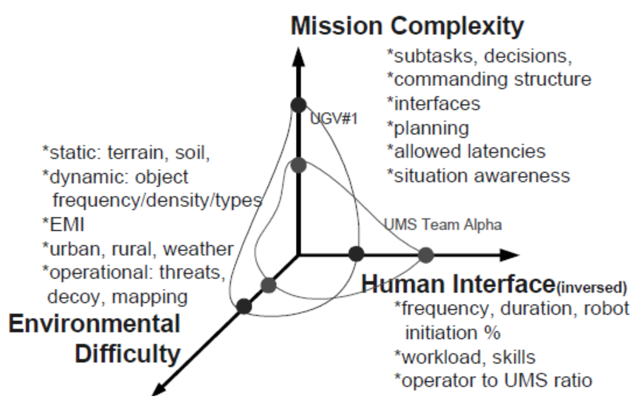


Figure 6. Autonomy Levels For Unmanned Systems (ALFUS) model of autonomy. The three axes represent three metrics for autonomy (Huang, Pavek, Albus, & Messina, 2005).

In relation to HRI, the Huang framework states that the relationship between the level of human robot interaction and the autonomy level of the robot "…is fairly linear for simple systems" (Huang et al., 2004, p. 5). The authors proposed a negative linear correlation between autonomy and HRI so that as the level of robot autonomy increases the HRI decreases. Interestingly, although this model states that HRI decreases with increased robot autonomy, the relationship between the human and robot approaches a team setting. HRI, according to this model, includes constructs such as human intervention (number of unplanned interactions), operators workload (as measured by NASA TLX), operator skill level, and the operator-robot ratio. This proposed linear relationship can be seen in Figure 7.

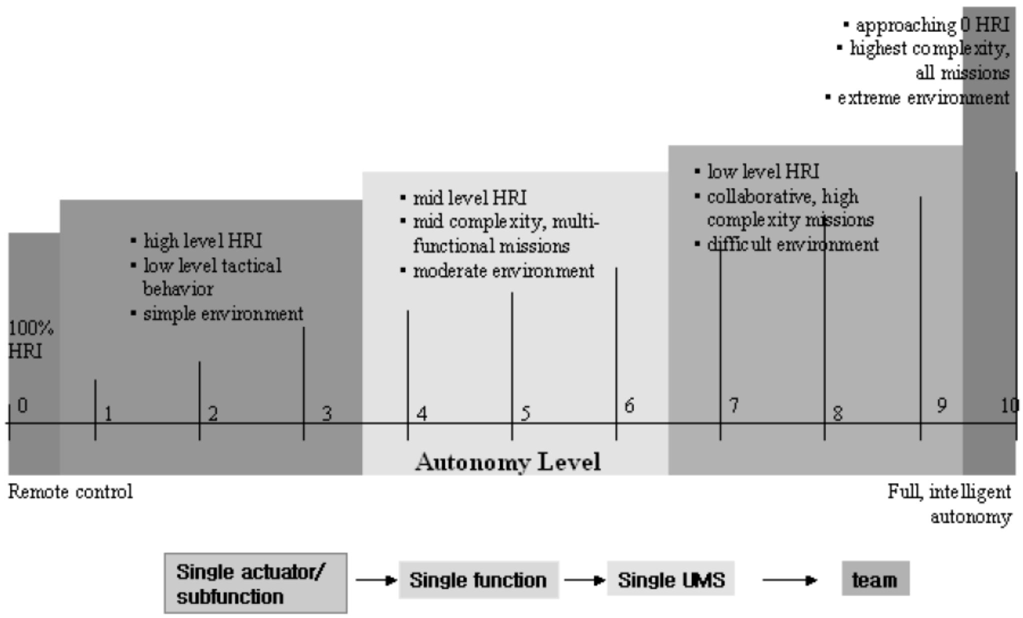


Figure 7. Autonomy Levels For Unmanned Systems (ALFUS) model of autonomy, depicting level of HRI along autonomy continuum (Huang, Pavek, Albus, & Messina, 2005).

Similar to Huang and colleagues' model, other researchers have proposed that higher robot autonomy requires less interaction or intervention (see quote below, Yanco & Drury, 2004a). Inversely, autonomy has been described as the amount that a person can neglect the robot. Neglect time (Goodrich & Olsen, 2003) refers to the measure of how the robot's task effectiveness (performance) declines over time when the robot is neglected by the user. Robots with higher autonomy can be neglected for a longer time period.

> "There is a continuum of robot control ranging from teleoperation to full autonomy: the level of human-robot interaction measured by the amount of intervention required varies along this continuum. Constant interaction is required at the teleoperation level, where a person is remotely controlling the

robot. **Less interaction is required as the robot has greater autonomy**" [emphasis added] (Yanco & Drury, 2004a, p. 2845).

The idea that higher autonomy reduces the frequency of human-robot interaction (e.g., Goodrich & Olsen, 2003; Huang et al., 2004; Yanco & Drury, 2004a) is a stark contrast to the way in which some other HRI researchers consider autonomy. A number of HRI frameworks and reviews propose that more robot autonomy requires more human-robot interaction (e.g., Feil-Seifer, Skinner, & Mataric, 2007; Goodrich & Schultz, 2007; Murphy & Woods, 2009; Thrun, 2004). The first of these reviews considers autonomy and HRI within the context of the science fiction guidelines for robotic behavior, Asimov's three laws of robotics. The three laws are (1) robot must not harm human; (2) robot must obey orders, except if conflicts with first law; (3) robot protects its own existence, as long as that does not conflict with law 1 or 2. Murphy and Woods revised and proposed an alternative set of laws based on what humans and robots can realistically accomplish in near future of HRI. Autonomy was a construct considered in two of the three laws.

The second law, "robot must obey orders", Murphy and Woods (2009) revised as "a robot must respond to humans as appropriate for their roles." Notice the word *respond.* The authors note that the notion of responsiveness (i.e., the capability to respond appropriately), may be considered different from the way in which autonomy is currently addressed in the literature. For a robot to be responsive, it may require a form of autonomy for engaging appropriately with others, where it will need to perceive and identify "different members, roles, and cues of a social environment". According to the authors, autonomy in the context of HRI requires the robot to perceive its social environment, and interact (respond) to that environment appropriately.

Additionally, Asimov's third law, "robot protects its own existence" was revised as "a robot must be endowed with sufficient situated autonomy to protect its own existence as long as such protection provides smooth transfer of control to other agents consistent with the first and second laws." Here, the authors discussed the importance of transfer of control between the robot and other agents, such as a human. Importantly, the revised third law suggested that a robot will need situated intelligence, and that more autonomy requires more sophisticated forms of coordinated activity between the human and robot. Coordination suggests the need for the human and robot to work together effectively.

The HRI perspective on the relationship between autonomy and human-robot coordinated activity is evident in other HRI frameworks. Thrun's (2004) framework of HRI defines three categories of robots, and each category requires a different level of autonomy as dictated by the environment the robot operates in. Industrial robots (e.g., manufacturing robots that assemble cars) have low autonomy because environment in which they operate is highly structured. Professional service robots (e.g., museum tour guides, search and rescue robots) and personal service robots (e.g., robotic walkers) mandate higher degrees of autonomy because they operate in a variable environment and interact in close proximity to people. For example, a personal service robot nursebot (http://www.cs.cmu.edu/~nursebot/) may require higher levels of autonomy to coordinate its behavior with patients and healthcare professionals. Thrun declares that autonomy is important in HRI, stating "human-robot interaction cannot be studied without consideration of a robot's degree of autonomy, because it is a determining factors with regards to the tasks a robot can perform, and the level at which the interaction takes place" (2004, p. 14).

Furthermore, autonomy has been proposed as a benchmark for developing *social interaction* in socially assistive robotics (Feil-Seifer, Skinner, & Mataric, 2007). The authors propose that autonomy serves two functions: to perform well in a desired task, and to be proactively social. However, the authors warn that the robot's autonomy should allow for social interaction only when appropriate (i.e., only when social interaction enhances performance). However, developing autonomous robots that engage in peer-to-peer collaboration with humans may be harder to achieve, than full autonomy with no social interaction (e.g., iRobot Roomba) (Goodrich & Schultz, 2007).

3.4.3. Critique of autonomy and human robot interaction. The frameworks and reviews outlined in this paper (Feil-Seifer, Skinner, & Mataric, 2007; Goodrich & Olsen, 2003; Goodrich & Schultz, 2007; Huang et al., 2004; Kahn et al., 2007; Murphy & Woods, 2009; Thrun, 2004; Yanco & Drury, 2004a) all mention autonomy as an important construct to consider within the context of HRI. These frameworks are useful in that they provide an overview of factors thought to be important in HRI. With consideration of these frameworks as a whole, there are a number of points to highlight.

First, a major problem with definitions exists. Although autonomy was a recognized HRI construct, the term lacks an explicit definition in some of these frameworks (Feil-Seifer, Skinner, & Mataric, 2007; Kahn et al., 2007; Murphy & Woods, 2009). Additionally, authors have been ambiguous concerning the precise meaning of the "I" in HRI. It seems intervention and interaction have been used synonymously, and the ambiguous use of these terms makes it unclear as to how autonomy should be measured. Conceivably interaction could be interpreted as a specific type of interaction (as suggested in Huang et al., 2004, but not clearly defined in the rest of Huang and colleagues' publications). However, do all forms of intervention/interaction constitute a reduction in robot autonomy? We do not think so. For example, a human may socially interact with a robot autonomously delivering refreshments at a social event. In this example, the robot may demonstrate high levels of autonomy, despite the human's interaction. The goal of some researchers (Goodrich & Olsen, 2003; Huang et al., 2004; Yanco & Drury, 2004a) for a robot to act autonomously with no HRI mirrors the human *out of the loop* phenomenon in the automation literature, which is known to cause performance problems (Endsley, 2006; Endsley & Kiris, 1995).

Second, Thrun (2004) clearly states that autonomy determines factors with regard to the tasks the robot can perform. This statement may be a bit misleading. We propose that autonomy can be thought of as influencing the *way* a task is carried out, not whether or not the task can be completed at all. For instance, a robot designed to mop floors could do so with no autonomy (i.e., teleoperated by a human), intermediate autonomy (i.e., the human and robot share aspects of the task), or fully autonomous (i.e., completes the mopping task by itself). Either way, the task is performed. However, the *way* the task is performed and the level to which the robot might interact with a human would be drastically different in these two examples. In fact, in these examples human interaction is present along the continuum of robot autonomy; however, the *nature* (not presence) of the interaction may change as a function of robot autonomy level.

Third, some of the frameworks made mention to the importance of the human and robot switching roles (Feil-Seifer, Skinner, & Mataric, 2007; Murphy & Woods, 2009; Thrun, 2004). Recall that in HRI, the autonomy continuum is often discussed in terms of "roles" that the human and robot fulfill (i.e., teleoperated to peer). However, intermediate levels of autonomy and subsequent roles that humans and robots then subsume are ill-defined. Furthermore, within the context of HRI, if a robot is expected to collaborate in a human-robot team, determining clear-cut allocation of functions may be particularly difficult. Teamwork is dynamic and complicated, and there are many factors that influence human-machine teams (Ezer, 2007).

3.5. Summary: Synthesis of Literatures and Critique of Autonomy as a Construct

Autonomy has been defined, measured, and considered in a variety of ways across several different disciplines: automation, robotics, and HRI. In an attempt to integrate the common characteristics included in the current definitions of autonomy found in the literature, the following strong definition has been proposed: *The extent to which a robot can* ***sense*** *the environment,* ***plan*** *based on that environment, and* ***act*** *upon that environment, with the intent of reaching some* ***goal*** *(either given to or created by the robot) with little or no external* ***control****.*

In addition to better understanding how autonomy is defined, the conceptualization of autonomy within the fields of automation, robotics and HRI has been investigated. This review yielded a number of overall insights. The fields of automation, robotics, and HRI considered autonomy as existing along a continuum. Although the terminology along this continuum varied slightly, the general principle was clear that autonomy may range from no autonomy (manual control) to a fully autonomous system. The *implementation* of autonomy within the fields of automation and robotics was somewhat consistent. Again, the terminology differed (see Figure 3), but both fields recognized that a system may exhibit autonomous capability in sensing (i.e., information acquisition), planning (i.e., information analysis and decision action selection) and behavior (i.e., action implementation).

Inconsistency in the conceptualization of autonomy was found in the HRI literature. Specifically, the notion surrounding *how* autonomy might affect the way in which a human and robot interact was varied, with some researchers believing higher levels of autonomy may reduce HRI, while others claiming that it will increase HRI. The empirical research surrounding HRI and autonomy (discussed in the next section) shed some light onto these dichotomous viewpoints. In short, the amount and nature of the human-robot intervention/interaction depends on the task at hand. Consider the following. Some HRI researchers have proposed that neglect time or the amount of human intervention determines autonomy level. However, this measure does not apply to all types of robots. What about a robot designed to be socially interactive and provide entertainment? Or a robot designed to generate speech and be conversational? There are many different types of robots that might require different measures and conceptualizations of autonomy. Autonomy, in relation to HRI, is useful only if it can support beneficial or successful interaction between the human and the robot.

Most importantly, the frameworks and reviews did not specify exactly *how* the interaction will change as a robot's autonomy may change. These are frameworks, not models, so the relationship between autonomy and other important constructs in HRI (e.g., safety, interfacing, teamwork, trust) is unclear. For instance, Feil-Seifer and colleagues (2007) relate autonomy to trust and allocation of functions. These are separate, albeit important, constructs that may interact with autonomy and influence HRI.

A model of autonomy and HRI is needed. As this section revealed, autonomy is an important construct related to HRI, and a multi-disciplinary approach to developing such a model is essential. Now that a definition of autonomy has been established, and inconsistencies in the literature identified, the rest of the current investigation will develop a framework based on human-, robot-, interaction-, and task-related variables related to autonomy, and finally the building blocks for a framework of autonomy and HRI will be proposed.

4. A Review of Variables Associated with Autonomy and HRI

Before developing a framework of autonomy in HRI, it is crucial to understand variables that influence – and are influenced by – robot autonomy. The human-, robot-, social interaction-, and task-related variables reviewed in this section will provide a foundation for a new framework to be introduced in Section 5. Finally, many of the variables included in this section have been researched with automated and robotics systems with fixed autonomy levels. For inclusiveness, a brief discussion of adjustable autonomy is included at the end of this section to highlight the idea of variable shifts in autonomy levels, and how that may impact HRI.

4.1. The Role of Human-Related Variables and Autonomy

In this section human-related variables of acceptance, situation awareness, trust, and mental workload are reviewed. These variables certainly influence one another (e.g., the relationship between workload and situational awareness; Tsang & Vidulich, 2006). Although some of these inter-variable relationships are discussed, the primary focus of this section is to consider each variable in relation to the continuum of autonomy levels.

Acceptance, situation awareness, trust, and mental workload are reviewed for a number of reasons. First, the four variables all have a strong basis in literature, with empirical research that deepens our scientific understanding of human performance; second, these variables have been investigated in HRI and have been established as important variables to consider when designing service robots; and third, these variables (in particular, situation awareness, trust, and workload) have been shown to be relevant to determine function allocation between humans and automation.

4.1.1. Acceptance. Given that the use of service robots may be expected to become a part of people's everyday lives, a critical issue that emerges is robot acceptance. Generally, acceptance has been described as a combination of attitudinal (i.e., users' positive evaluation or beliefs about the technology), intentional (i.e., users' plan to act a certain way with the technology), and behavioral (i.e., users' actions in using the product or technology) acceptance (Davis, 1989).

The widely recognized Technology Acceptance Model (TAM; Davis, 1989) proposed two main variables that affect acceptance: perceived usefulness and perceived ease of use. There is strong empirical support for TAM (Venkatesh & Davis, 2000; Venkatesh, Morris, Davis, & Davis, 2003), in part due to its ease of application to a variety of domains. The model's simplicity has evoked some criticism (Bagozzi, Davis, & Warshaw, 1992), and as a result more complex models have been developed (e.g., TPC, Goodhue & Thompson, 1995; UTAUT, Venkatesh, Morris, Davis & Davis, 2003). Various acceptance models differ in complexity and content; however, their overarching goal is to understand, explain, and model predictive variables that contribute to user acceptance. Technology acceptance models may provide general guidance for understanding acceptance, therefore there is a need to understand how and if these models will map onto robotics (see Ezer, Fisk, & Rogers, 2009).

Research investigating robot acceptance has focused in large part on user attitudes toward robots. Two widely recognized robot attitude scales, the Negative Attitude Towards Robots Scale (NARS; Nomura, Kanda, Suzuki, & Kato, 2004; Nomura, Suzuki, Kanda, & Kato, 2006a) and the Robot Anxiety Scale (RAS; Nomura, Suzuki, Kanda, & Kato, 2006b) have been used to gauge psychological reactions evoked in humans by robots. Use of these scales may delineate to what extent people feel unwilling to interact with a robot due to arousal of negative emotions or anxiety. The use of these scales have suggested that gender (Nomura, Kanda, & Suzuki, 2006), culture (Bartneck, Nomura, Kanda, Suzuki, & Kato, 2005; Bartneck, Suzuki, Kanda, & Nomura, 2007) and robot experience (Bartneck, Suzuki, Kanda, & Nomura, 2007) influence attitudes toward robots. However, anxiety and negative attitudes, as measured in NARS and RAS, may only capture a portion of influential factors on robot acceptance.

With regard to service robots, a number of reviews have been conducted with the goal of developing a comprehensive categorization of factors that influence user acceptance of service robots. Broadbent and colleagues (2009) conducted a review of factors found to influence human responses to healthcare robots. They categorized those factors into two categories: robot factors and person factors. Robot factors include the robot's appearance; humanness; facial dimensions and expressions; size; gender; personality; and ability to adapt to users' preferences and needs. Person factors include the user's age; needs; gender; technology/robot experience; cognitive ability; education; culture; role within society (e.g., job); and finally anxiety and attitudes towards robots.

In another recent literature review, Young and colleagues (2009) posited design guidelines for the development of home-based robots. In contrast to the aforementioned review, Young and colleagues used a social psychology approach to focus on robot acceptance within the context of socialization between home-based robots and humans. With basis in the social psychology literature, seven factors were identified as influential in people's acceptance of home robots: safety; accessibility and usability; practical benefits; fun; social pressures; status gains; and social intelligence. The authors also noted users' previous experiences and perceptions of media, personal social network, and robot design also critical.

To date, an empirical investigation of user acceptance as a function of the level of robot autonomy has not been conducted. Based on a review of technology acceptance models as well as the robot acceptance scales and reviews, we suggest that the key factors that influence and predict acceptance may vary along the autonomy continuum. For example, let us consider the variables perceived ease of use (Davis, 1989) and usability (Young et al., 2009). The nature of ease of use and usability in robots with low autonomy (e.g., teleoperation) may be dependent on the usability of input devices such as a joystick or remote to command a robot. Whereas, ease of use or usability for semi-autonomous robots may be dependent on more sophisticated control methods, such as physical manipulation. The use of control methods along the autonomy continuum will be discussed further in Section 4.2.4.

Appearance may be another factor that varies along the autonomy continuum (Broadbent et al., 2009; see also Section 4.3). For instance, users have expressed a preference for home-based mobile teleoperated robots to appear "soft" with "rounded edges" to match home decor (Beer & Takayama, 2011). Robots designed to perform jobs requiring social interaction (social intelligence also considered an acceptance factor; Young et al., 2009) such as semi-autonomous robotic museum tour guides (e.g., Faber et al., 2009), were preferred to have human-like appearances (Goetz, Kiesler, & Powers, 2003). Currently, it is unknown whether users will place equal/less/more consideration on the social intelligence of the robot compared to its robot capability/function. Most technology acceptance models do not include a social variable or construct (e.g., Davis, 1989; Goodhue & Thompson, 1995), with the exception of the UTAUT model (Venkatesh, Morris, Davis, & Davis, 2003) which only considers social norms focused on person-person social interaction, not person-technology social interaction.

In conclusion, acceptance is an important variable to consider with regard to predicting technology use (Davis, 1989), as well as HRI (Broadbent, Stafford, & MacDonald, 2009; Young, Hawkins, Sharlin, & Igarashi, 2009). Furthermore, designers should be mindful of users' acceptance, because radical technologies, such as personal robots, are not as readily accepted as incremental innovations (Dewar & Dutton, 1996; Green, Gavin, & Aiman-Smith, 1995). Despite the research community's acknowledgement that acceptance is an important construct to consider, further research is needed to understand and model the key variables that influence robot acceptance, how such variables interact, and finally how those variables related to acceptance may vary in predictive value over the autonomy continuum.

4.1.2. Situation Awareness and Mental Workload. Situation awareness and mental workload are two concepts that are intricately intertwined (see, Tsang & Vidulich, 2006). Additionally, much empirical evidence suggests that both constructs influence human performance changes as a function of LOA. The purpose of this section is to review the interactions between the variables situation awareness and mental workload, their relation to autonomy, and finally their potential role in HRI.

Situation awareness (SA) is a construct with a substantial body of research pointing to its utility and efficacy (for reviews, see Durso & Gronlund, 1999; Durso & Sethumadhavan, 2008; Endsley, 2006; Parasuraman, Sheridan, & Wickens, 2008). Situation awareness (SA) is defined as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley, 1995, p. 36). SA can be categorized into three levels (Endsley, 2006): level 1 relates to perceiving the status, attributes, and dynamics of relevant elements in the environment (e.g., a pilot perceives the terrain, system status, warning lights); level 2 is described as the comprehension and understanding of the significance of those elements perceived in level 1 (e.g., a pilot determines the meaning of a warning light and whether or not to abort); and finally level 3 SA is the operator's ability to project the future actions of the elements in the environment (e.g., a pilot perceives another aircraft and must determine whether the aircraft will violate safe separation in the near future).

As described by the three SA levels, SA is the diagnosis of the state of a dynamic world (Durso & Sethumadhavan, 2008). It is important to note that SA is a psychological construct, separate from human performance and behavior. That is, SA is not the decision of what action/behavior to take as a consequence of the diagnosis. Rather, SA moderates the *accuracy* of that choice. In other words, an individual could demonstrate low levels of SA, but still maintain high performance as long as the automation is working properly.

Therefore, in tasks where SA is critical, it can be maintained by teaching operators where and how to seek information, integrate that information, and predict its implications (Parasuraman, Sheridan, & Wickens, 2008). However, when automation is introduced, a balance between the automated assistance and the operators' task involvement needs to be met to promote optimal performance. Research suggests that the level of automation is more indicative of SA quality than subjecting operators to periods of passive manual control (as in empirical manipulations of adaptive automation, Kaber & Endsley, 2004).

The introduction of automation has the risk of lowering SA by putting operators *out of the loop*. Issues related to *out of the loop* performance and maintaining SA are especially prominent during automation failures. Poor SA and the operator being *out of the loop* may not be a problem when the automation is performing well. However, in the case of a failure or a situation the automation is not equipped to handle, the *out of the loop* operator may be unable to diagnose the problem and intervene in a timely manner (Endsley, 2006). In particular, poor SA during a failure in high LOA can be detrimental to performance (Endsley & Kaber, 1999; Endsley & Kiris, 1995).

Performance in using automated systems, particularly when recovering from system failure, has been shown to also rely on mental workload. Mental workload has been defined as "the relation between the function relating the mental resources demanded by a task and those resources available to be supplied by the human operator" (Parasuraman, Sheridan, & Wickens, 2008, pp. 145-146). One of the fundamental benefits of introducing automation to complex systems is to lessen the chance of human error by reducing operator mental workload. However, the relationship between workload and operator performance is complex. If automation is poorly designed, and engagement of the automation actually increases "cognitive overhead" (i.e., compared to completing the task manually), then automation may actually increase workload (Parasuraman & Riley, 1997). Kaber and Riley (1999) studied the effects of dual tasks while using automation. Their results suggested that when the operator was performing in a dual task (i.e., a dynamic control task with automation, and a monitoring task), a decrease in performance was found as workload increases.

The dependency of SA quality on level of automation is related, in part, to workload. There is much empirical support suggesting the criticality of balancing SA and workload, and an imbalance can lead to detrimental performance errors. In general, if the automation is designed properly, as LOA increases workload decreases. Conversely, as LOA decreases workload increases. However, low workload during high LOA has the potential to lead to boredom (Endsley & Kiris, 1995), particularly in monitoring tasks (e.g., air traffic control). On the other end of the spectrum, high workload during low LOA generally leads to low operator SA and decreased performance (Endsley & Kaber, 1999; Endsley & Kiris, 1995). Generally, it has been suggested that intermediate levels of automation may mitigate these problems. In fact, in some studies if the system operates with intermediate levels of automation, SA has been shown to increase in comparison to full automation (Endsley & Kaber, 1999; Kaber & Endsley, 2004; Kaber, Perry, Segall, McClernon, & Prinzel, 2006).

The rich empirical background of situation awareness in the automation literature can be informative to robotics. Although the automated systems discussed above have been primarily studied in the context of air traffic control and aviation, similar human-machine interactions may be expected in HRI. In fact, much of the work involving SA and robotics has been conducted in similarly dynamic service environments and tasks, such as search and rescue (e.g., Riley & Endsley, 2004), military operations (e.g., Kaber, Wright, & Sheik-Nainar, 2006), teleoperation of uninhabited vehicles (Scholtz, Antonishek, & Young, 2004), materials handling (e.g., Kaber, Onal, & Endsley, 2000), or assembly (e.g., Sellner, Heger, Hiatt, Simmons, & Singh, 2006).

The primary focus of most HRI research that investigated operator SA has been conducted with robots of low level autonomy, specifically teleoperation. For instance, Riley and Endsley (2004) investigated the use of a teleoperated robot in search and rescue training exercise. The case study demonstrated the difficulty in teleoperation in complex environments. In particular, operator SA was difficult to maintain, with limitations in the amount of information provided on a display, uncertainty about environmental obstacles perceived by robot camera, and issues with robot control. Similarly, Yanco and Drury (2004b) tested teleoperation of a robot for search and rescue, using four expert first responders as operators. They identified problems in maintaining SA related to a lack of feedback regarding camera angle (i.e., camera was often left off-center). The authors recommended feedback regarding the direction of the camera, availability of maps or GPS for identification of robot location, and the option for the robot to autonomously drive to reduce operator workload (similar findings found in Scholtz, Young, Drury, & Yanco, 2004).

In a study investigating multiple autonomy levels of a simulated robot teleoperation task, participants were instructed to use a simulated robot arm to handle and store containment vessels (Kaber, Onal, & Endsley, 2000). The simulation included five autonomy levels (mapping onto Endsley & Kaber, 1999 LOAs: action support, batch processing, decision support, and supervisory control, full automation). Mirroring findings in the automation literature, operator SA was lower for higher autonomy levels, compared to manual operation (action support). Additionally, higher autonomy levels yielded lower operator workload. Finally, during simulated robot failure the operator was forced to manually control the robot. Participants recovered from robot failure better when the system was functioning at lower or intermediate levels of autonomy compared to full autonomy. These findings are in line with other work suggesting that the robot operator needs time and proper feedback to regain SA when the robot interrupts the participants to "ask for help" (Sellner et al., 2006).

Although the HRI literature on SA is limited to the aforementioned operations, they highlight the importance of maintaining SA in a dynamic environment. Many other service applications (e.g., home and healthcare settings) are also dynamic, and challenges in maintaining the human's SA, using controls and displays, and responding to unpredictable variables (e.g., people, animals, other moving obstacles) should be expected to be present. However, more research is needed to investigate SA in home and healthcare settings. Furthermore, many of the studies discussed in this section tested applications with trained users, such as experts in search and rescue (Riley & Endsley, 2004; Yanco & Drury, 2004b) or involving participants with some experience with robots (Scholtz, Antonishek, & Young, 2004; Sellner, Heger, Hiatt, Simmons, & Singh, 2006). Future research investigating SA in home, office, or possibly healthcare settings will need to consider interaction involving uses with little to no formal training.

Most of the studies conducted to date involve a fixed robot autonomy level; if the autonomy level did change, the autonomy was only turned on or turned off (e.g., to simulate robot failures, Kaber, Onal, & Endsley, 2000; Kaber, Wright, & Sheik-Nainar, 2006). There are gaps in our understanding on how SA may be supported with adjustable autonomy involving fluctuation between autonomy levels along the entire autonomy continuum.

In large part, the HRI literature lacks studies investigating workload as either an independent or dependent variable. It may be speculated that the intrinsic relationship between workload and SA is likely to exist along the robot autonomy continuum. That is, high workload would likely decrease the human's SA about the state of the robot and its operational environment. Additionally, robot failures at high autonomy levels would likely negatively impact task performance. However, these relationships between workload, SA, and HRI need further empirical research.

Finally, we would like to suggest that the *nature* of the human's SA may change as a function of robot autonomy. Therefore, situation awareness at low levels of autonomy may primarily focus on where the robot is located, what obstacles lay ahead, or deciphering the sensor data the robot produces. As a robot approaches higher autonomy levels it may be perceived as a teammate or peer (Goodrich & Schultz, 2007; Milgram, Rastogi, & Grodski, 1995). SA associated with a robot peer may more closely resemble that of *shared SA*, where the human must know the robot's status and likewise the robot must know the human's status to the degree that they impact their own tasks and goals. Design principles for supporting SA in team operations (Endsley, Bolte, & Jones, 2003; Gorman, Cook, & Winner, 2006) may be applied to human-robot teams, and need to be empirically tested.

4.1.3. Trust. Trust in HRI can be informed by trust in automation (Hancock, Billings, Schaefer, Chen, Visser, & Parasuraman, 2011), which has been studied extensively (for reviews, Lee & Moray, 1992, 1994; Lee & See, 2004; Parasuraman, Sheridan, & Wickens, 2008). However, studies that directly investigate the effect of LOA on trust have not been conducted.

Rather, the studies included in the following review of the literature have exclusively focused on supervisory control and decision support automation. Nonetheless, a number of models and theories related to trust in automation (Cohen, Parasuraman, & Freeman, 1998; Dzindolet et al., 2003; Lee & See, 2004; Madhavan & Wiegmann, 2007), and preliminary models of trust in HRI have been proposed (Desai, Stubbs, Steinfeld, & Yanco, 2009; Hancock, Billings, & Schaefer, 2011). These models suggest that trust, in conjunction with many other factors, can predict automation/robot use.

Similar to situation awareness and mental workload, it is important to stress that trust is a *psychological* construct. In particular, trust has been described as a cognitive state or attitude, based on factors such as predictability or operator expectations, that usually influences behavior dependence on the automated system (Lee & See, 2004; Parasuraman & Riley, 1997; Parasuraman & Wickens, 2008). Trust is not a performance measure; it is measured subjectively.

In contrast, a behavioral outcome variable of trust is dependence. Dependence can be categorized as either reliance or compliance with an automated machine. Reliance occurs when the automated system does not require human action, and the human therefore does not intervene with the automated system; whereas compliance occurs when the automated machine requests human action, and the human therefore takes that action (Dixon & Wickens, 2006).

Although trust influences operator's reliant and compliant use of the system, these factors are not the same. Automation may not always be used in the way a designer intended if the operator inappropriately trusts the system. Specifically, inappropriate amounts of trust can lead to misuse of the system (in cases of high levels of trust and the operator over relies on the automation) or disuse of the system (in cases of very low levels of trust, and the operator rejects the automation's capabilities; Lee & See, 2004; Parasuraman & Riley, 1997).

A similar classification of errors can be applied to HRI. For example, although trust was not directly measured, researchers observed potential disuse and misuse during a Robot Rescue Competition (Scholtz, Young, Drury, & Yanco, 2004). Disuse may have occurred when a robot failed to autonomously navigate through an opening, suggesting an obstacle was in the way. The operator switched the autonomy mode to manual control and forcibly drove the robot through the opening, when in fact unbeknownst to the operator the opening was covered in Plexiglas (the environment was damaged and the Plexiglas broken). In the same competition, misuse may have occurred when an operator allowed a robot to autonomously navigate with the assumption that the robot's sensors would detect obstacles. In fact, unknown to the user, the robot collided with a number of obstacles, potentially damaging the robot and environment.

In the misuse example above, the robot demonstrated low levels of reliability (see Section 4.2.2 for review of reliability and autonomy). Both robots and automation are bound to make mistakes, particularly if performing in unknown, unstructured, and dynamic environments (see task/environment, Section 4.4). The automation literature largely supports that the reliability of a system is a predictor of human trust (e.g., Lee & See, 2004; Parasuraman & Riley, 1997; Parasuraman, Sheridan, & Wickens, 2008; Sanchez, 2006). In a multi-task simulation of an automated agricultural vehicle (Sanchez, 2006), the recency of errors was negatively related to both perceived reliability of and trust in the system. Similarly, participants' trust of a robot has been shown to be negatively impacted after initially experiencing low robot reliability (de Visser, Parasuraman, Freedy, Freedy, & Weltman, 2006). Other factors shown to influence trust in automation included the user's prior knowledge and understanding of system capabilities, the user's age, as well as the user's expectations of system performance (Sanchez, 2006).

Regarding trust in HRI, in a meta-analysis of factors affecting trust in robots 69 correlational and 47 experimental effect sizes were evaluated (Hancock, Billings, Schaefer, Chen, Visser, & Parasuraman, 2011). Their findings suggested that trust in robots was influenced by human-related variables (e.g., expertise, situation awareness, prior experiences), robot-related variables (e.g., reliability, adaptability, proximity, personality) and environmental variables (e.g., team collaboration, task characteristics). In particular, robot characteristics, particularly robot performance-based factors, were the largest current influence on perceived trust in HRI. Although this study is critical step toward identifying quantitative estimates of factors influencing trust in HRI, a caveat to consider is the substantially fewer empirical studies of trust in HRI compared to trust in automation. Additionally, the authors' definition of a robot was broad; some of the studies included in the meta-analysis appear to investigate automation rather than robots. More research, particularly investigating the influence of human- and environment- related factors on HRI trust, is needed (Hancock et al., 2011).

Although trust in automation may inform trust in robots, there are some important differences to consider. First, automation generally lacks physical embodiment (i.e., many automated systems are primarily software based). Many robots are physically mobile, look or behave like humans or animals, and physically interact with the world. Physical robot characteristics (e.g., size, weight, speed of motion) and their effects on trust need to be empirically evaluated.

Second, unlike most automated systems, some service robots are designed to be perceived as teammates or peers with social capabilities, rather than tools (e.g., Breazeal, 2005; Groom & Nass, 2007). Understanding how to develop trust in robots is an avenue of research critical for designing robots meant to be perceived as social partners.

In conclusion, there is a lack of empirical research investigating the relationship between level of autonomy and trust (Desai, Stubbs, Steinfeld, & Yanco, 2009). Furthermore, it is anticipated that as robots become increasingly advanced and perform complex tasks, the robot's autonomy will be required to adjust or adapt between levels (see "adjustable autonomy"; Section 4.5). In general, robotic and automated systems that operate under changing levels of autonomy (e.g., switching between intermediate levels) are not addressed in the trust literature. Many avenues of research need to be pursued to better understand the role of trust in HRI, how trust in robots is developed, and how misuse and disuse of robots can be mitigated.

4.2. The Role of Robot-Related Variables and Autonomy

In this section, robot-related variables associated with autonomy were reviewed and discussed within the context of automation and HRI. The variables of interest were intelligence/learning, reliability, and transparency/feedback. Similar to the sections above, the primary focus of reviewing robot-related variables was to consider each variable in relation to the continuum of autonomy levels. Intelligence/learning, reliability, and transparency/feedback were chosen due to each variable's relation to autonomy and large literature base in robotics and automation.

4.2.1. Intelligence and Learning. What is an intelligent robot? This may seem like a silly question. Aren't all robots intelligent? In short, the answer is 'no'. Aren't all autonomous robots intelligent? In short, the answer is 'to some degree'. There is no standard definition of human intelligence. Likewise, there is no standard definition of artificial intelligence. Historically, artificial intelligence has been researched and thought about in four different categories: reasoning, behavior, modeling humans, or rationality (Russell & Norvig, 2003).

However, in addition to these categories, *learning* is also considered a major component of intelligence (Bekey, 2005; Murphy, 2000; Russell & Norvig, 2003).

Robot learning is not synonymous with machine learning. Machine learning takes place in a computer. Whereas, 'robot learning' requires the computer to interact with the environment (Bekey, 2005). That is, the robot must use sensor data of the world around it and apply learning algorithms to that data. Not all robots have learning algorithms, but some common forms of learning include reinforcement learning, neural network learning, social learning, imitation, among many other methods.

Not all robots are intelligent, but robots that demonstrate higher levels of autonomy may be required to be more intelligent. According to Bekey (2005), robot intelligence may manifest as sensor processing (e.g., vision, hearing), reflex behavior (rapid SENSE-ACT primitive couplings), special purpose programs (e.g., navigation, localization), or cognitive functions (reasoning, learning, planning). Generally speaking, the more autonomous a robot is, the more sophisticated these components may be.

In the future, it is expected that most autonomous robots will be equipped with some ability to learn. This will be especially true as robots are moved from the laboratory to an operational environment, where the robot will have to react and adjust to unpredictable and dynamic environmental factors (Bekey, 2005; Russel & Norvig, 2003; Thrun, 2003).

4.2.2. Reliability. The effect of system reliability has a strong empirical basis in the automation literature, and those findings may provide insight into the relevance of robot reliability in HRI. Reliability refers to how well the system performs (note: reliability is different from *reliance* which is a human-related behavioral variable associated with trust, Section 4.1.3; it is also different from capability which is defined as the extent the task operations can be controlled by automation). No automated system or robot will be completely reliable. Reliability of automation and robots is generally expected to be less than perfect because of constraints in designing algorithms to account for every possible scenario found in the operational environment (Parasuraman & Riley, 1997). This would be true especially for service robots that operate in unpredictable and complex environments such as the home, hospital settings, or workplaces. Therefore, it is important to consider the consequences on human behavior when the system commits failures.

A number of automation studies indicate that higher LOA (e.g., action implementation) is beneficial in an automated task in comparison to lower LOAs, but only if reliability is 100% (Horrey & Wickens, 2001; Moray Inagaki, & Itoh, 2000; Rovira, Zinni, & Parasuraman, 2002). Degraded reliability (i.e., failures) at high LOA can be detrimental, where performance measures plummet after failure and the operator's "time to recovery" extend (Endsley & Kiris, 1995; Endsley & Kaber, 1999). The first failure an automated system commits is of particular importance, having the largest negative impact on trust and continued automation use (Lee & Moray, 1992).

How may reliability be measured, for both automation and robotics? Reliability should be measured against a standard, which is referred to as task criterion. For example, if a robot's task is to fetch items from the floor, the task criterion should be *fetching items*. The task criterion can be set to any degree of specificity, referred to as the threshold (e.g., fetching items of a pre-specified size range). The task criterion and associated threshold provide a standard for which reliability can be assessed.

The task criterion and threshold may differ along the autonomy continuum. For example, let us consider the four stages of automation: stage 1 information acquisition; stage 2 information analysis; stage 3 decision and action selection; and stage 4 action implementation (Parasuraman, Sheridan, & Wikens, 2000). At early stages of automation, diagnostic automation stages 1 and 2, reliability can be modeled by signal detection theory, where the automation threshold corresponds to signal detection theory response bias (β). Consider a stage 2 automated system, a warning alert system, the only errors the system could make is either a miss (alert is silent when it should not be) or a false alarm (alert goes off when it should not) (Parasuraman & Wickens, 2008). Determining the threshold in later stage automation (i.e., decision making or action implementation automation), and anticipating the various types of errors the automated system may commit, is far more complex.

Given that every system, automation and robotics, will make errors, a logical question to ask is *how do you determine the appropriate threshold?* Addressing this question proves to be a balancing act between designing with the assumption the machine will sometimes fail and consideration for automation errors on human performance. For early stage automation, oftentimes designers often set the threshold low, assuming that a miss is generally more costly than a false alarm (this is often the case in high fidelity simulations and in real systems; Parasuraman & Wickens, 2008). However if the false alarm rate is high, this can lead to a "cry-wolf" effect, and may lead to the human ignoring the alarm (i.e., lack of compliance). This is not to say that robots or automated systems should not perform in risky tasks, rather *if* a system is to assist in a critical task, the "burden of proof should be on the designer" to ensure that their design will not hinder the human's performance (Parasuraman, Sheridan, & Wickens, 2000, p. 292).

In summary, reliability is a variable to consider when determining the appropriate level of robot autonomy for a task. If the automated system or robot is unreliable, then it is recommended that the level of autonomy is reassessed and adjusted to support optimal performance (Parasuraman, Sheridan, & Wickens, 2000).

4.2.3. Transparency and Feedback. Developers should design the robot in a way that allows the user to observe the system and understand what the system is doing. Automated tasks where an operator can form a mental model, is often referred to as being transparent; whereas, a task where the operator lacks information about the automated reasoning or logic would be opaque or unpredictable. Unpredictability may increase with higher levels of autonomy as a result of the human not being in control directly and immediately (Miller & Parasuraman, 2007).

In the automation literature, transparency has been suggested to be of particular importance in two ways: first, transparency is critical in user recovery from automation failure, and second, transparency has been suggested to contribute to user recognition of the automation LOA mode (e.g., recognizing whether a plane is in autopilot or manual control). If proper feedback during failure is not provided in *any* LOA then performance can be negatively affected. For instance, Lorenz and colleagues (2001) demonstrated that failure intermediate levels of automation resulted in the worse performance. This was attributed to the fact that important diagnostic information that was available in low LOA and high LOA was not available in the intermediate level.

Using feedback to promote transparency is important even when the automation is functioning properly, particularly to depict the LOA and aspects of the task that are automated. In fact, in an observational and interview study, a major problem in using automated healthcare devices was the participants' failure to demonstrate an understanding of what aspects of the task were manual, and what aspects of the task were automated (Dinka, Nyce, & Timpka, 2006). Interfaces should be designed to support the human's understanding of the automation's capabilities and limitations during manual performance, automated performance, and importantly during the transition between these states (Kaber, Riley, Tan, & Endsley, 2001). Information which promotes transparency should ideally be available before the transition in LOA occurs.

Transparency research in the HRI literature has demonstrated the complexity of designing for appropriate feedback. For example, in a study using a simulated robot in a mine-disposal task, changes in robot autonomy (e.g., from semi-autonomous to manual control) were communicated to the human via visual and auditory cueing, such as icons or tones (Kaber, Wright, & Sheik-Nainar, 2005). In terms of transparency, the cuing feedback improved performance but it did not eliminate all SA problems.

In a case study with a remote robot collecting data for signs of life in a desert, the robot autonomy level was experimentally manipulated, and its effects of operator common ground (i.e., communication between science team), and mental model quality was qualitatively assessed (Stubbs, Wettergreen, & Hinds, 2007). The robot's autonomy levels were: low (record data, detect some failure conditions, and detect obstacles); medium (sense nearby obstacles, develop basic navigation plans, and act on plans with minimal human intervention); high (sense, plan, and deploy instruments with little to no human intervention). The results suggested that poor mental model development for all three autonomy levels could be attributed to different transparency problems. In the low autonomy condition, the operators did not understand the robot's perceptive capabilities, and feedback about errors or instrument failures from robot was missing. In the medium autonomy condition, major issues involved inadequate feedback from the robot, where operators had difficulty in understanding what the robot was doing. In the high autonomy condition, the operators' confusions was not based on what the robot was doing, rather there was a lack of feedback related to why the robot made particular decisions. Therefore for all levels of autonomy, poor transparency as a result of inadequate feedback was primary constraint. Higher autonomy did not lead to error-free interaction, and the authors suggest that a shared mental model is more complex for interaction with systems of increasing autonomy.

Designing feedback displays to support the relationship between transparency and mental models has been proposed as a major design component of HRI (Goodrich & Olsen, 2003). However, the relationship between transparency and mental models has not been thoroughly evaluated and is inconsistently considered. For instance, Goodrich and Olsen stated, "Interacting with the world requires a mental model, and interacting with the robot requires a separate mental model. If the robot is transparent to the user, then only one mental model is required" (p. 3946). The exact meaning of this phrase is unclear. Are the authors proposing that transparency entails a lesser need for the operator to develop a mental model of the robot? Whatever the authors' message, we stress that a mental model of the robot's capabilities and limitations is always important, and transparency increase the *accuracy* of humans' mental model of the robot.

Goodrich and Olsen advise that to promote transparency, information supporting the humans' understanding of both the robot and the environment in which the robot operates should be presented in a display. Additionally, Goodrich and Olsen's simplistic recommendation of presenting the operator with information about the robot and environment should be taken with some caution. Indeed, appropriate display design will help users understand the systems' functional capabilities and limitations (for reviews, Bennett, Nagy, & Flach, 2006; Lee, 2006). However, much consideration is needed in determining *how much, when*, and *what type of* feedback is most beneficial for any given task and any given robot autonomy level.

An additional consideration is the medium in which a robot could present feedback to a human. Traditionally for many automated systems, transparency is achieved by using visual feedback (i.e., on a graphical interface) or auditory feedback (i.e., warning alarms). However, we would like to highlight that robots can be transparent in many modalities. For instance, feedback concerning the function and autonomy level of the robot could be communicated using verbal feedback, emotive/social feedback, or clarity/ intent of robot movement and gestures (e.g., Bruce, Nourbakhsh, & Simmons, 2002; see also Section 4.3 for the role of social interaction in autonomy). At this time, it is unclear what form or method of feedback promotes optimal human-robot interaction as a function of robot autonomy.

4.2.4. Interfacing / Methods of Control. Given the review and conceptualization of autonomy thus far, there may seem to be a contradiction between autonomy, the extent to which a robot can carry out its own processes, and control, which implies some method of human intervention. Certainly, the nature of control and autonomy are interlinked.

Coinciding with levels of robot autonomy, levels of control can also be considered (Bekey, 2005). On the *robot* side of the interaction, control is often discussed in terms of control architectures. An example of "low level" control would be algorithms that ensure the robot motors are working properly, its legs are moving in a stable manner, or the motors controlling the robot's wheels do not begin to oscillate (Bekey, 2005). Low level control processes that function in parallel are known as behavior-based control architectures, discussed in Section 3.3.2 as SENSE-ACT primitive couplings (Arkin, 1998). The next "level" up may include capabilities such as obstacle avoidance during navigation, or following (i.e., a form robot navigation where the robot follows a moving target, such as a human). "High level" control may include functions related to goal planning (Bekey, 2005).

On the *human* side of the interaction, control or "method of control" refers to the way in which the human may intervene and provide input. The appropriate application of a control method is dependent on two factors: (1) the robot autonomy, and (2) the task. Although users have indicated a preference to command robots using voice control (Ezer, 2008; Kahn, 1998), a variety of control methods may be appropriate for varying levels of robot autonomy.

Service robots that are low on the autonomy continuum (i.e., teleoperated) most often receive human input from interfaces such as two-handed and one-handed controllers, computer mouse, or keyboard (e.g., Duran et al., 2009; Michaud et al., 2010; Takayama, Marder-Eppstein, Harris, & Beer, 2011). Whereas, semi-autonomous service robots may receive human input from a variety of shared control methods, such as demonstration (Billard, Calinon, Ruediger, & Schaal, 2008; Nicolescu, & Mataric, 2003), direct physical interfaces (Chen & Kemp, 2011), gesture recognition (Bremner et al.,2009; Charles et al., 2009; Gielniak, & Thomaz, 2011), laser pointers (Nguyen, Jain, Anderson, & Kemp, 2008; Kemp et al., 2008), or voice command (Asyal et al., 2011; Ceballos, Gomez, Prieto, & Redarce, 2009; Hyun, Gyeongho, & Youngjin, 2007).

The proper match between the level of robot autonomy and the method of control is essential. Performance measures, particularly for levels of robot autonomy where the human is implementing action using teleoperation, are highly dependent on the method of control. For example, performance decrements in a simulated teleoperated robot performing a materials handling task were attributed to operator difficulty in using a SpaceBall© motion input control device (Kaber, Onal, & Endsley, 2000). Similarly, in a Kaber et al. (2006) study, high operator mental workload in information analysis and decision making automation was attributed to visual and attentional demands of the user interface.

4.3. The Role of Interaction Variables and Autonomy

Although empirical research that directly manipulates and compares various levels of robot autonomy and social interaction have not been conducted, the role of social interaction was worth reviewing at a more general level. A rich literature base in social robotics exists (for review see Fong, Nourbakhsh, & Dautenhahn, 2003). However for this section, a specific review of how social interaction may potentially intersect with autonomy was conducted.

4.3.1. Social characteristics, social effectiveness, and appearance. Understanding social interaction in HRI, and its relation to autonomy, is a challenging task. For one, robots have been a topic of science fiction, media, and film for decades. In fact, robots are one of the few technologies in which design has been modeled in part by science fiction portrayals of autonomous systems (Brooks, 2003). Even though most individuals of the general population have never interacted with a robot directly, most people have ideas or definitions of what a robot should be like (Ezer, 2008; Khan, 1998). If users have preconceived notions of how robots should behave, then it becomes all the more important to better understand how to create a match between user expectations and the robot's actual autonomy. According to Breazeal (2003), when designing robots, the emphasis should not be whether people will develop a social model to understand robots. Rather, it is more important that the robot adhere to the social models the humans expect. What social models do people hold for robots? And do those social models change as a function of robot autonomy?

It is accepted in the research community that people treat technologies as social actors (Nass, Fogg, & Moon, 1996; Nass & Moon, 2000; Nass, Moon, Fogg, Reeves, 1995; Nass, Steuer, Henriksen, & Dryer, 1994), particularly robots (Breazeal, 2005). Social capability has been categorized into classes of social robots (Breazeal, 2003; Fong, Nourbakhsh, & Dautenhahn, 2003): socially evocative, social interface, socially receptive, sociable, socially situated, socially embedded, and socially intelligent. The differentiations between these classes is beyond the scope of this paper, but it is interesting to note that these classes can be thought of as a continuum (ranging from *socially evocative* where the robot relies on human tendency to anthropomorphize, to *socially intelligent* where the robot shows aspects of human style social intelligence, based on models of human cognition and social competence). Social classes higher on this continuum require greater amounts of autonomy to support the complexity and effectiveness of the human-robot interaction.

How do you measure "social effectiveness"? Of course it is difficult to determine the most appropriate metric for measuring social effectiveness. A variety of metrics have been proposed (Steinfeld et al., 2006) and applied to measure social effectiveness via interaction characteristics (e.g., interaction style, or social context), persuasiveness (i.e., robot is used to change the behavior, feelings, or attitudes of humans), trust, engagement (sometimes measured as duration), and compliance. Appropriate measures of social effectiveness may vary along the autonomy continuum. For instance, when a robot is teleoperated, social interaction may not exist between the robot and human, per se. Rather, the robot may be designed to facilitate social communication *between people* (i.e., the operator and a remotely located individual). In this case, "successful social interaction" may be assessed by the quality of *remote presence* (the feeling of the operator actually being present in the robot's remote location). Proper measures of "social effectiveness" may be dictated by the quality of the system's video and audio input/output, as well as communication capabilities, such as lag time/delay, jitter, or bandwidth (Steinfeld et al., 2006). Whereas, social interaction with intermediate or fully autonomous robots is appropriately assessed by the social characteristics of the robot itself (e.g., appearance, emotion, personality; Breazeal, 2003; Steinfeld et al., 2006).

The last aspect of social robots we want to discuss is appearance. Not all appearances elicit a social response. For example, if a robot is "functionally" designed, the appearance is intended to reflect the task the robot is to perform (e.g., the iRobot gutter cleaner). However, robots may also have a "biologically inspired" design, and have a physical embodiment that more closely resembles humans or animals. According to Fong and colleagues (2003), a robot that is high on the autonomy continuum and expected to be perceived as a peer, may require an appearance that projects "human-ness". Although, if a robot becomes too human-like it may be perceived as strange in appearance (uncanny valley; Mori, 1970). Other research suggests that the robots' appearance should project appropriate "robot-ness" so the user does not develop detrimentally false expectations of the robot capability (Duffy, 2003). In other words, the appearance should match the autonomy level of the robot. However, what aspects of a robot appearance affect the human's perception of autonomy, and what physical features constitute "human-ness" or "robot-ness" is not well understood.

In summary, there are many open research questions regarding social interaction and robot autonomy. Automation, unlike robots, is not necessarily designed to elicit social responses from humans. As the lines between automation and robots blur (with the development of machine intelligence and learning), *some* forms of automation may illicit a social response from operators. In these instances, the literature on social robotics can inform automation design.

4.4. The Role of Task Variables and Autonomy

In this section, task-related variables expected to relate to robot autonomy were reviewed and discussed. The variables of interest included task criticality, task accountability, and the environment. These variables were considered to the role task criticality and accountability play in function allocation between the robot and human. Finally, the environment was considered due to its critical influence on robot capability.

4.4.1. Task Criticality and Accountability. Service robot applied to home, healthcare, the workplace, and many other applications may be partaking in tasks of high criticality. Error, on the part of the robot, could result in consequences both minor and severe. In the event of a robot error, the human interacting with the robot will manage the consequences of that error.

In general, the automation literature suggests that as the consequences of failure increases, operators adjust their behavior to manage automation errors more effectively. For example, in a flight simulation study with trained aircrew (Mosier, Skitka, Dunbar, & McDonnell, 2001), automation errors were more often detected and corrected for critical aspects of flight (e.g., altitude capture), compared to errors that occurred in less critical functions (e.g., communications). Similarly, in an automated counting task (i.e., circles flashed on a screen, then the circles disappeared and the automation provided an estimate of the number of circles; Ezer, Fisk, & Rogers, 2008), participants could either accept the automation's estimate or recount the circles and type their own estimate. Cost of errors was manipulated with a scoring system. Participants (both younger and older adults) decreased their reliance on the automation as the cost of error (criticality) increased.

Task characteristics and consequences of error can be influenced by automation or robot autonomy level. As mentioned in Section 4.2.2, in many cases failures or errors at early stages of automation may not be as critical as errors at later stages of automation. One rational is that it may be risky to program a machine to have high autonomy in a task that requires decision support, particularly if the decision outcome involves lethality or human safety (Parasuraman, Sheridan, & Wickens, 2000; Parasuraman & Wickens, 2008). For example, unreliability in a robot that autonomously navigates may results in either false alarms or misses of obstacles. In this example, the criticality of errors is substantially less than errors conducted by a robot that autonomously determines what medication a patient should take. In this example, robot failure may result in critical, if not lethal, consequences. Criticality of errors related to medication management may explain the findings in a recent questionnaire, where older adults expressed a preference for humans, rather than robots, to decide which medication they should take (opposed to other less-critical related tasks such as medication delivery; Mitzner et al., 2011).

Criticality of robot error was investigated in an assessment of failure logs from 15 mobile robots (Carlson, Murphy, & Nelson, 2004). Data were compiled for over three years, and a total of 171 reported errors were logged. In this report, criticality was measured by "impact" or the amount of robot downtime after an error. The amount of downtime varied widely between robots and error types, but interestingly the data indicated that over the course of 3 years, the overall impact decreased, probably due to operator and technician behavior changes. That is, the humans interacting with the robot learned to identify common problems, troubleshoot, or order problematic parts in advance. Although "robot downtime" may not be a proper criticality measurement for all types of errors, these findings point to the importance of variables such as operator training or robot experience may have on management of critical errors.

The likelihood of the human detecting robot failure is not only dependent on the criticality of the task. Responsibility for the success of task completion also influences how well a human detects errors. In human teams, accountability is split among team members. As robots become more autonomous and are perceived as peers or teammates, it is possible that the distribution of accountability may be perceived to be split between the robot and human. However, a robot cannot be held accountable for detecting its own errors; this responsibility must fall onto a human.

Responsibility of identifying automation errors was investigated in a study by Skitka and colleagues (2000). Participants were either given non-accountable instructions (i.e., they told that they were not evaluated by performance), or one of four accountable groups (i.e., participants were told they were evaluated by performance and had to justify their actions in meeting overall performance goals). The results suggested those operators who were held accountable for system performance conducted fewer omission (failure to detect automation misses) and commission (failure to detect automation false alarms) errors.

Robot autonomy has been shown to play a role in participants' accountability of tasks errors. In a study investigating perceived robot autonomy and accountability (Kim & Hinds, 2006), a robot assisted participants in a Lego assembly task. When the robot was perceived as more autonomous, participants reported less self-blame (accountability) for task errors. These findings suggest, if a robot is perceived as autonomous, responsibility of consequences may be misplaced and the human operator may feel less accountable for errors.

In highly critical applications, such as healthcare, accountability of a healthcare robots' reliability will fall on human professionals and staff. Due to the possible potential of server consequences of failure, healthcare professionals and staff have expressed concern about (1) the potential for robot errors (Broadbent et al., 2011; Tiwari, Warren, Day, & MacDonald, 2009), and (2) who may be accountable for those errors (Tiwari, Warren, Day, & MacDonald, 2009). Therefore, care should be taken in determining which tasks a robot shall perform autonomously, as well as in designing the system so human supervisors held accountable for the robot can easily diagnose and alleviate consequences of error.

4.4.2. Environment. Service robot designed for assistive functions (e.g., home or healthcare applications), surveillance, or first responders (e.g., search and rescue) will be required to operate in unknown, unstructured, and dynamic environments. Functioning in such difficult environments will certainly influence the functional requirements of the robot. In fact, in an in depth assessment of common mobile robot errors, robots employed in the field (opposed to research robots) failed more often by a factor of 10, probably due to the demands of the field environments (Carlson, Murphy, & Nelson, 2004). As Desai and colleagues stated, “this lack of environmental constraints makes designing automation to cover all possible circumstances the robot might encounter very difficult. As a result, these types of robotics systems are likely to have lower reliability than other automated systems” (2009, p. 4).

Therefore, the robot’s capability to function in a dynamic environment is highly dependent on environmental factors (e.g., lighting, reflectivity of surfaces, glare) that influence the robot sensors to perceive the world around it. Higher levels of robot autonomy may be required for a service robot to function in unstructured ever-changing environments (Thrun, 2004). That is, the robot must have the autonomy to make changes in goals and actions based on the sensor data of the dynamic environment. However, not all aspects of the environment can be anticipated, so for many complex tasks the presence of a human supervisor may be required (Desai, Stubbs, Steinfeld, & Yanco, 2009).

4.5. Adjustable Autonomy

It is feasible to assume that for any given task, the autonomy requirements for the system to function may vary depending on the context of the task. For example, imagine a service robot that is teleoperated by the user to navigate a home environment in a find and fetch task. If the home is cluttered, and the user mental workload increases the robot could assist with obstacle avoidance. Whereas, if the user’s workload is low, then the robot might reduce its autonomy to allow for manual control. The idea that machine autonomy can dynamically change during operation has been suggested as a potential solution to alleviate operator out-of-the-loop performance problems, loss of situational awareness, and high mental workload (Kaber, Riley, Tan, & Endsley, 2001; Miller & Parasurman 2007).

In the automation literature, Scerbo (2001) outlined two forms of dynamic function allocation, with the difference between the two was a matter of authority. The first, adaptable automation, described a system where the user initiates changes among presentation modes or functionality (e.g., Miller & Parasuraman, 2007). The second, adaptive automation, included systems where both the user and the system can initiate changes in system states or modes (e.g., Inagaki, 2003; Kaber, Riley, Tan, & Endsley, 2001; Kaber, Wright, & Sheik-Nainar, 2005; Scerbo, Freeman & Mikulka, 2003). In either of these forms, triggering mechanisms for shifting among levels of automation may include environmental events, operator performance, or physiological assessment (for review, Miller & Parasurman 2007).

In HRI, the concept of robot autonomy shifting between levels has been dubbed adjustable autonomy (Bradshaw, Feltovich, Jung, Kulkarni, Taysom, & Uszok, 2004; Flacone & Castelfranchi, 2001; Maheswaran, Tambe, Varakantham, & Myers, 2004; Scerri, Pynadath, & Tambe, 2002) or sliding scale autonomy (Desai, 2007; Desai & Yanco, 2005; Sellner, Heger, Hiatt, Simmons, & Singh, 2006). Maheswaran and colleagues (2004) proposed two types of adjustable autonomy. The first was referred to as user-based adjustable autonomy, where high criticality of problem solving should result in human control. The second was referred to as agent-based adjustable autonomy, where the agent chooses to switch control to user, based on the system's own limitations in utility (i.e., agent unable to perform certain task) or uncertainty (i.e., shortfalls in agent reasoning or strategy/decision selection).

Although the idea of adjustable autonomy levels for robots is not new, the challenge is designing (1) optimal triggers for switching robot autonomy levels for a given task, and (2) appropriate feedback to indicate when a switch has been made so the user can attend to newly allocated functions while maintaining situation awareness.

4.6. Summary of Variables Associated with Autonomy and HRI

The purpose of this review was to develop a collection of concepts and principles thought to be important in understanding robot autonomy in HRI. Literature in automation and HRI provides a vast amount of research pertinent in identifying human-, robot-, interaction-, and task-related variables influenced by robot autonomy. These variables include acceptance, situation awareness, trust, robot intelligence, reliability, transparency, methods of control and social interaction. Additionally, task variables were identified that influence a robots capability of functioning within a complex environment, and therefore impact allocation of functions between a robot and human.

A theme present in much of this review is that the role of autonomy in HRI is complex. Assigning a robot with an appropriate level of autonomy is important because a service robot changes human behavior. Appropriate autonomy levels may increase task performance and decrease human workload. However inappropriate autonomy levels, with limitation in robot reliability, can decrease SA, negatively impact trust, or increase workload.

HRI is a relatively young field with substantial, albeit exciting, gaps in our understanding of causal relationships between variables and concepts. In large part, many of the variables reviewed in this section lack research where levels of robot autonomy are used as an independent variable. This is probably due to the lack of consensus in identification of intermediate levels of robot autonomy. However, the knowledge gleaned from the identified variables in both Section

3 and Section 4 can be applied toward the development of a framework. The goal of a framework is to organize variables related to autonomy in a potentially explanatory manner, and give structure to future investigations.

5. Toward a Framework of Levels of Robot Autonomy and HRI

In this section, we provide a framework for examining levels of robot autonomy and its effect on human robot interaction. The framework can be used as an organizing flow chart, consisting of several stages (Figure 8). Stages 1-3 serve as a guideline to determine robot autonomy and will be discussed in Section 5.1. Section 5.2 will describe Stage 4 of the framework. This stage categorizes robot autonomy using a proposed 10-point taxonomy, Finally, stage 5 (Section 5.3) broadly suggests the implications of the robot autonomy on HRI (i.e., human variables, robot variables, and interaction variables that were identified in Section 4), and a conceptual model of the framework is presented.

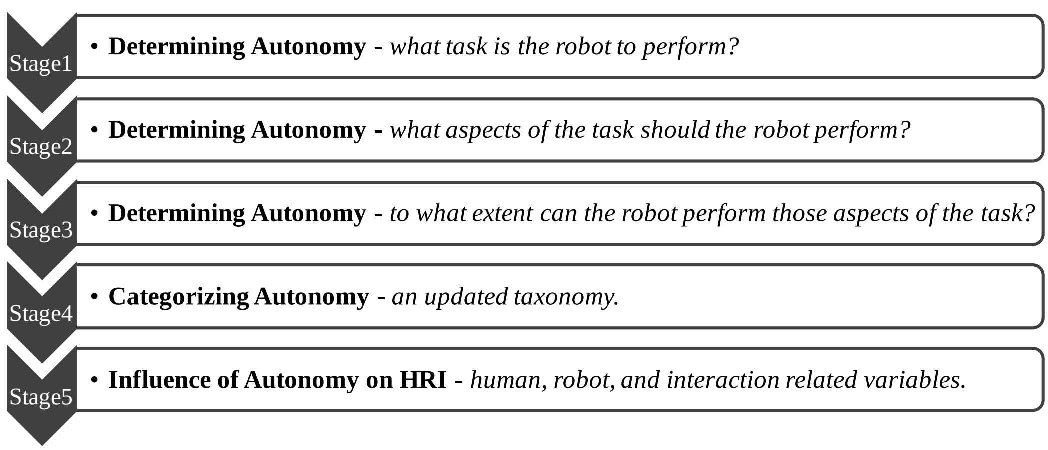


Figure 8. Organizing flow chart to determining robot autonomy and autonomy's effects on HRI. Steps to be included in more detail in the framework.

5.1. Determining Robot Autonomy

In this section, a review of guidelines for determining and measuring robot autonomy is presented. Specifically, the proposed guidelines in this section focus on human-robot interaction, with an emphasis on function allocation between a robot and a human.

5.1.1. Determining autonomy as a function of the task and environment. The role of task/environment is largely missing or lacks emphasis in previous models of LOA (Endsley & Kaber, 1999; Parasuraman, Sheridan, & Wickens, 2000). However, in the review of how autonomy has been conceptualized in both robotics and HRI, the role of the environment is evident (Sections 3.3 and 3.4). Furthermore, in my proposed definition of autonomy (Section 3.1.2), the environment is a critical characteristic. Consideration of the task/environment is particularly important for robotics, compared to automation, for a reason. A robot more so than automation is embodied, that is it is situated within an environment and usually expected to perform tasks by physically manipulating that environment. A robot's capability to sense, plan, and act within its environment is what determines autonomy. Therefore, in this framework, the first determining question to ask is:

"*What task is the robot to perform?*"

A researcher should not ask "is this robot autonomous"; rather the important consideration is "can this robot complete the given task at some level of autonomy". For instance, the iRobot Roomba is often considered an autonomous robot. The robot is capable of navigating and vacuuming floors autonomously. However, if the task *vacuuming* is broadened to consider other subtasks (i.e., picking up objects from floor, cleaning filters, emptying dirt bin/bag) then the Roomba may be considered semi-autonomous because it only completes a portion of those subtasks. Likewise, if the environment is changed (e.g., vacuuming stairs opposed to flat surfaces), the Roomba's autonomy could be categorized differently, as it is currently incapable of vacuuming stairs.

Therefore, specifying the context of the task/environment is critical in determining the task-specific level of robot autonomy. Specific task-related variables that influence autonomy are the task criticality, task accountability, and the environment. These variables are discussed in detail in Section 4.4, and will be incorporated in the conceptual model of the framework.

5.1.2. An objective basis for measuring autonomy. Once the task and environmental demands are determined, the next determining question is:

"What aspects of the task should the robot perform?"

Each task, no matter how simple or complex, can be abstractly broken down into primitives: SENSE, PLAN, and ACT. Let us consider robots equipped with assisted teleoperation features (e.g., Takayama et al., 2011). In this example, a teleoperated robot demonstrates low levels of autonomy by assisting the human operator in obstacle avoidance. Usually, this feature is programmed into the robot architecture using behavior-based SENSE-ACT couplings, where the robot is assisting with the aspects of the task by detecting obstacles (SENSE), then adjusting its behavior to avoid collision (ACT). The human remains, in large part, in charge of path planning and navigational goals (PLAN). However, a robot that navigates semi-autonomously (e.g., Few et al., 2008) may require a human may specify the high level goal of navigating to a specified location. Once the high level goal is given, the robot then autonomously navigates to that location. Here, the robot demonstrates a high level of autonomy in sensing the environment (SENSE), relatively high level autonomy in PLAN (except the human provided the high level goal), and a high level of autonomy in physically implementing the plan (ACT).

As the two examples suggest, autonomy can vary along any of the SENSE, PLAN, and ACT primitives, which relates to the next determining question:

"To what extent can the robot perform those aspects of the task?"

Each of the sense, plan, and act primitives could be allocated to either the human or the robot (or both). Similar to Parasuraman, Sheridan, and Wickens (2000) stages of automation, a robot can vary in autonomy level (from low to high) along the three primitives (see Figure 9).

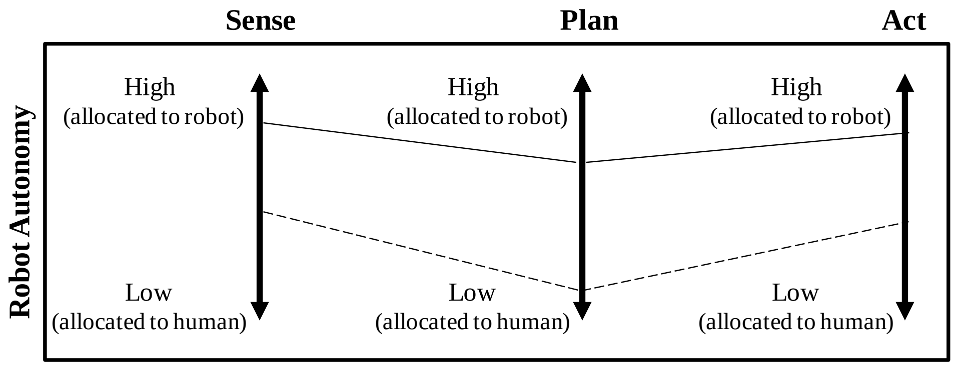


Figure 9. Levels of autonomy across the robot primitives sense, plan, and act. Two examples are given: assisted teleoperation (dotted line) and semi-autonomous navigation (solid line). Model modified from Parasuraman, Sheridan, and Wickens, 2000.

As depicted in Figure 9 the level of autonomy may vary from low to high for each of the robot primitives. Determining the robot autonomy prompts a clarification of how to *measure* the extent or degree to which a robot can perform each of those aspects (SENSE, PLAN, ACT) of the task. In the automation literature, level of autonomy is most often indentified by function allocation. Consider the Endsley and Kaber's (1999) model, the level of automation is specified in their taxonomy based on the allocation of function to either the human or automation. For instance, in their automation level Automated Decision Making: the automation selects and carries out an option; the human can have input in the alternatives generated by the automation.

In HRI the allocation of function has been commonly measured by amount of human intervention (Yanco & Drury, 2004a). Specifically, human intervention is measured by the percentage of time a task is completed on its own, and intervention is measured by the percentage of time the human must control the robot (Yanco & Drury, 2004a). The two measures, autonomy and intervention, must sum to 100%. For example, a teleoperated robot has autonomy=0%, and intervention=100%. A fully autonomous robot has autonomy=100%, and intervention=0%. In between these two anchor points, lies a continuum of shared control. For example, a medication management robot may select a medication, and handoff the medication to a human, but the human might be responsible for high level directional (navigation) commands. Here, autonomy=75% and intervention=25%. Similarly, autonomy has been measured as human neglect time (Olsen & Goodrich, 2003). In this metric, autonomy is measured by the amount of time that the robot makes progress toward a goal before dropping below effective threshold (see reliability threshold, Section 4.2.2) or requiring user instruction.

Although this idea of measuring the time of intervention and neglect is useful, it has limitations. Amount of *time* between human interventions may vary as a result of other factors, such as the inappropriate levels of trust (i.e., misuse and disuse), social interaction, task complexity, robot capability (e.g., robot speed of movement), usability of the interface/control method, and response time of the user. Therefore, if interaction *time* is used as a quantitative measure, care should be taken when generalizing those findings to other robot systems or tasks. We propose that a supplemental metric may be a qualitative measure of intervention *level* (i.e., subjective rating of the amount human intervention), or a general quantitative measure focused on subtask completion, rather than time (i.e., number of subtasks completed by robot divided by the number of total subtasks required to meet a goal). Each metric is not without tradeoffs, but still may provide some general indication as to what the robots degree of autonomy may be.

Intervention is defined as the human performing some aspect of the task. As we have discussed earlier, *intervention* and *interaction* are not necessarily interchangeable terms. Intervention is a type of interaction specific to task sharing. Interaction may include other factors not necessarily specific to the intervention of task completion, such as verbal communication, gestures, or emotion expression. Some autonomous service robots could work in isolation, requiring little interaction of any kind (e.g., an autonomous pool cleaning robot); whereas, other robots working autonomously in a social setting may require a high level of interaction (e.g., an autonomous robot serving drinks at a social event). Finally, the measure of autonomy, as discussed in this section, is specifically applicable to service robots, which perform tasks. Neglect time may not be an appropriate measure of autonomy for robots designed specifically for entertainment, for example. Other types or classes of robots may require different evaluative criteria for determining autonomy, beyond the scope of this paper

5.2. Categorizing Levels of Robot Autonomy (LORA) for HRI: A Taxonomy

Now that guidelines for determining robot autonomy have been outlined, the next stage is categorizing the robot's autonomy along a continuum. A lack of specification of intermediate autonomy levels is a limitation in previous HRI frameworks (e.g., Huang, Pavek, Albus, & Messina, 2005; Yanco & Drury, 2004a). Therefore, in Table 4, we propose a taxonomy in which the robot autonomy can be categorized into "levels of robot autonomy" (LORA).

The taxonomy has a basis in HRI by specifying each LORA from the perspective of the interaction between the human and robot, and the roles each play. That is, for each proposed LORA, we suggest the (1) function allocation between robot/human for each of the SENSE, PLAN, ACT primitives, (2) provide a proposed description of each LORA, and (3) provide support with examples of service robots from the HRI literature. The literature in Table 4 includes a mix of empirical studies involving robots and simulations, as well as published robot autonomy architectures. Autonomy is a continuum with blurred borders between the proposed levels. The levels should not be treated as exact descriptors of a robot's autonomy. Rather, the levels should be treated as an approximation of a robot's autonomy level along the continuum.

Table 4
Proposed Taxonomy of Levels of Robot Autonomy for HRI

Level of Robot Autonomy (LORA)	Function Allocation *Sense*	*Plan*	*Act*	Description	Examples from HRI Literature
1. Manual Teleoperation	H	H	H	The human performs all aspects of task including sensing the environment and monitoring the system, generating plans/options/goals, and implementation.	"Manual Teleoperation" Milgram, 1995 "Tele Mode" Baker & Yanco, 2004; Bruemmer et al., 2005; Desai & Yanco, 2005
2. Action Support	H/R	H	H/R	The robot assists the human with action implementation. However, sensing and planning is allocated to the human. For example, a human may teleoperate a robot, but the human may choose to prompt the robot to assist with some aspects of a task (e.g., gripping objects).	"Action Support" Kaber et al., 2000
3. Assisted Teleoperation	H/R	H	H/R	The human assist with all aspects of the task. However, the robot senses the environment and chooses to intervene with task. For example, if the user navigates the robot too close to an obstacle, the robot will automatically steer to avoid collision.	"Assisted Teleoperation" Takayama et al., 2011 "Safe Mode" Baker & Yanco, 2004; Bruemmer et al., 2005; Desai & Yanco, 2005
4. Batch Processing	H/R	H	R	Both the human and robot monitor/sense the environment. The human, however, determines the goals and plans of the task. The robot then implements task.	"Batch Processing" Kaber et al., 2000
5. Decision Support	H/R	H/R	R	Both the human and robot sense the environment and generate a task plan. However, the human chooses the task plan and commands robot to implement action.	"Decision Support" Kaber et al., 2000
6. Shared Control with Human Initiative	H/R	H/R	R	The robot autonomously senses the environment, develops plans/goals, and implements actions. However, the human monitors the robot's progress, and may intervene and influence the robot with new goals/plans if the robot is having difficulty.	"Shared Mode" Baker & Yanco, 2004; Bruemmer et al., 2005; Desai & Yanco, 2005 "Mixed Initiative" Sellner et al., 2006 "Control Sharing" Tarn et al., 1995
7. Shared Control with Robot Initiative	H/R	H/R	R	Robot performs all aspects of the task (sense, plan, act). If the robot encounters difficulty, it can prompt the human for assistance in setting new goals/plans.	"System-Initiative" Sellner et al., 2006 "Fixed-Subtask Mixed-Initiative" Hearst 1999
8. Supervisory Control	H/R	R	R	Robot performs all aspects of task, but the human continuously monitors the robot. The human has over-ride capability and may set a new goal/plan. In this case the autonomy would shift to shared control or decision support.	"Supervisory Control" Kaber et al., 2000
9. Executive Control	R	(H)/R	R	The human may give an abstract high level goal (e.g., navigate to environment to specified location). The robot autonomously senses environment, sets plan, and implements action.	"Seamless Autonomy" Few et al., 2008 "Autonomous mode" Baker & Yanco, 2004; Bruemmer et al., 2005; Desai & Yanco, 2005
10. Full Autonomy	R	R	R	Robot performs all aspects of a task autonomously without human intervening with sensing, planning, or implementing action.	

**Note*: H = Human, R = Robot

5.3. A Framework of Robot Autonomy: The Influence of Autonomy on HRI

Finally, the last stage of the framework is to evaluate the influence of the robot's autonomy on the interaction between the robot and human. This portion of the framework is informed by the literature review outlined in Section 4. As that review suggests, the influence of autonomy on human-, robot-, and interaction- related variables is complex.

Evaluation of the robot's autonomy level on these variables can be used as evaluative criteria to determine if the autonomy level of the robot is appropriate for optimal human-robot interaction. In this way, the framework can be used as a tool for guiding appropriate autonomy levels that support optimal human-robot interaction. For example, there is much empirical research suggesting that SA decreases if an operator is *out of the loop*. In the case of a system failure the *out of the loop* operator may be unable to diagnose the problem and intervene in a timely manner (Endsley, 2006). If evaluation of the operator indicates that their SA is suboptimal, then the framework could provide guidelines to reevaluate what aspects of the task should be allocated to the robot/human to keep the operator *in the loop*, and the robot's autonomy could then be reconsidered and adjusted along the continuum to support optimal operator involvement and SA.

The entire framework (stages 1-5) is depicted as a conceptual model in

Figure ***10***. If read from top to bottom, the model depicts the guideline stages. By no means should this model be treated as a method. Rather the framework and taxonomy should be treated as guidelines to determine autonomy, categorize the LORA along a qualitative scale, and consider which HRI variables may be influenced by the LORA.

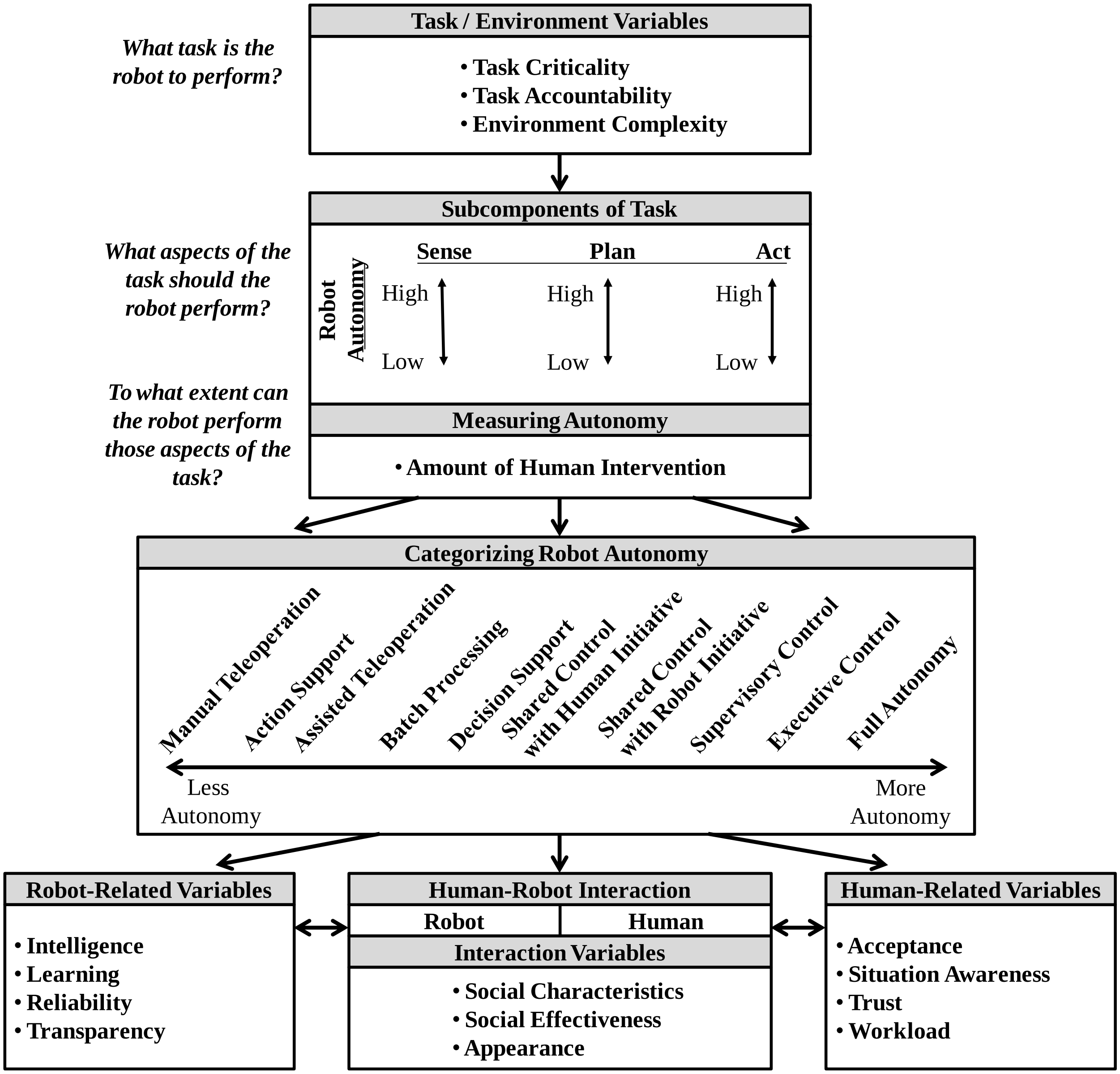


Figure 10. A framework of levels of robot autonomy for HRI. This framework can serve as a flow chart suggesting task and environmental influences on robot autonomy, guidelines for determining/measuring autonomy, a taxonomy for categorizing autonomy, and finally HRI variables that may be influenced by robot autonomy.

6. Conclusion

Levels of autonomy, ranging from teleoperation to fully autonomous systems, influence the nature of human-robot interaction. The purpose of this investigation was to investigate robot autonomy within the context of HRI. To do this, a number of evaluative steps were taken. It was first important to redefine the term *autonomy* and consider how the construct has been conceptualized within the fields of automation, robotics, and HRI. Next, a systematic review of the literature was conducted. This review revealed numerous human, robot, task, and interaction variables that are expected to influence and are influenced by autonomy. The knowledge gained from these literature reviews contributed to the development of a framework that should serve as a roadmap for categorizing LORA and evaluating the effects of robot autonomy on HRI.

Robot design can be guided by the framework proposed in this investigation. The framework provides a guide for appropriate selection of robot autonomy. This is important because the implementation of a service robot does not only supplement a task, but changes human activity by imposing new demands on the human. For this reason, the framework also has scientific importance, beyond the use as a tool for guiding function allocation. As such, the framework conceptualizes the autonomy along a continuum, and also identifies HRI variables that need to be evaluated as a function of robot autonomy. These variables include acceptance, SA, trust, robot intelligence, reliability, transparency, methods of control, and social interaction.

In large part, many of the variables included in the framework require further research to better understand autonomy's complex role in HRI. HRI is a young field with substantial, albeit exciting, gaps in our understanding. Therefore, the proposed framework does not provide indication of causal relationships between variables and concepts. As the field of HRI develops, empirical research can be causally mapped to theoretical concepts and theories. Gaps in research avenues related to the variables included in the framework have already been discussed throughout the paper (in particular, see Section 4). We will summarize a few of those gaps in brief, as well as propose additional research avenues that we suggest are in need of further attention in the HRI community:

- *SA and Workload* – although SA is becoming an increasingly common HRI metric, workload is not as often measured; both variables need further research as a function of robot autonomy.
- *Trust* – how trust changes over the autonomy continuum is largely not well understood in both automation and robots; most research has investigated only a few automation stages.
- *Learning* – a robot may increase autonomous capability via learning algorithms; but little is known about the development of human trust as the robot learns new tasks over time.
- *Modes of Feedback and Transparency* – robots can provide feedback of its state and autonomy mode with voice communication, gestures, and emotion; these constructs and their application to robots along the autonomy continuum need further investigation.
- *Safety* – service robots will function in close approximation with people; safety measures such as specifying robot workspace, collision avoidance techniques, or emergency override would likely vary along the autonomy continuum and need further research.
- *Perceived Robot Autonomy* –Little is known how a mismatch between perceived autonomy and actual autonomy can impact human performance.

In summary, further HRI research is needed to continue to identify appropriate trade-offs in allocating tasks to either a human or a robot. Implementing service robots has the potential to improve the quality of life for many people. But robot design will only be successful with consideration as to how the robots' autonomy will impact the human-robot interaction.

7. References

Agah, A. (2001). Human interactions with intelligent systems: Research taxonomy. *Computers and Electrical Engineering, 27*, 71-107.

Alami, R., Chatila, R., Fleury, S., Ghallab, M., & Ingrand, F. (1998). An architecture for autonomy. *International Journal of Robotics Research, 17*(4), 315-337.

Anderson, R. J. (1996). Autonomous, teleoperated, and shared control of robot systems. *Proceedings of the IEEE International Conference on Robotics and Automation*, 2025-2032.

Arkin, R. C. (1998). *Behavior Based Robotics*. Boston: MIT Press.

Asyali, M. H., Yilmaz, M., Tokmakci, M., Sedef, K., Aksebzeci, B. H., & Mittal, R. (2011). Design and implementation of a voice-controlled prosthetic hand. *Turkish Journal of Electrical Engineering and Computer Sciences, 19(1)*, 33-46.

Bagozzi, R. P., Davis, F. D., & Warshaw, P. R. (1992). Development and test of a theory of technological learning and usage. *Human Relations, 45*(7), 659-686.

Baker, M., & Yanco, H. A. (2004). Autonomy mode suggestions for improving human-robot interaction. *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, 3*, 2948-2953.

Bartneck, C., Nomura, T., Kanda, T., Suzuki, T., & Kennsuke, K. (2005). A cross-cultural study on attitudes towards robots. *Proceedings of the HCI International*, Las Vegas, NV, United States.

Bartneck, C., Suzuki, T., Kanda, T., & Nomura, T. (2007). The Influence of People's Culture and Prior Experiences with Aibo on their Attitude Towards Robots. *AI & Society – The Journal of Human-Centered Systems, 21*(1-2), 217-230.

Beer, J. M., & Takayama, L. (2011). Mobile remote presence systems for older adults: Acceptance, benefits, and concerns. *Proceedings of Human-Robot Interaction Conference: HRI 2011*, Lausanne, CH, 19-26.

Bekey, G. A. (2005). *Autonomous Robots: From Biological Inspiration to Implementation and Control*. Cambridge, MA: The MIT Press.

Bennett, K. B., Nagy, A. L., & Flach, J. M. (2006). Visual displays. In G. Savendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd ed., pp. 1191-1221). New York: Wiley.

Billard, A., Calinon, S., Ruediger, D., & Schaal, S. (2008). Robot programming by demonstration. In Bruno Siciliano & Oussama Khatib (Eds.) *Handbook of Robotics* (pp.1-24). Berlin: Springer.

Bradshaw, J. M., Feltovich, P., Jung, J., Kulkarni, S., Taysom, W., & Uszok, A. (2004). Dimensions of adjustable autonomy and mixed-initiative interaction. In M. Nickles, M. Rovatsos, and G. Weiss (Eds.) *Agents and Computational Autonomy : Potential, Risks, Solutions, Lecture Notes in Computer Science*, 2969.

Breazeal, C. (2003). Emotion and Sociable Humanoid Robots. *International Journal of Human Computer Interaction, 59*, 119-115.

Breazeal, C. (2005). Socially intelligent robots. *Interactions, 12*(2), 19-22.

Bremner, P., Pipe, A., Melhuish, C., Fraser, M., & Subramanian, S. (2009). Conversational gestures in human-robot interaction. *The International Conference on Systems, Man and Cybernetics (SMC)*, 1645-1649. San Antonio, TX.

Broadbent, E., Stafford, R. & MacDonald, B. (2009). Acceptance of healthcare robots for the older population: Review and future directions. *International Journal of Social Robotics, 1*(4), 319-330.

Broadbent, E., Tamagawa, R., Patiences, A., Knock, B., Kerse, N., Day, K., & Macdonald, B. A. (2011). Attitudes towards health-care robots in a retirement village. *Australasian Journal on Aging*, DOI= http://dx.doi.org/10.1111/j.1741-6612.2011.00551.x.

Brooks, R. A. (2002). It's 2001 already. *Flesh and Machines: How Robots will Change Us* (63-98). New York: Vintage Books, Random House Inc.

Bruce, A., Nourbakshsh, I., & Simmons, R. (2002). The role of expressiveness and attention in human-robot interaction. *Proceedings of the IEEE International Conference on Robotics and Automation*, 4138-4142.

Bruemmer, D. J., Few, D. A., Boring, R. L., Marble, J. L., Walton, M. C., & Nielsen, C. W. Shared understanding for collaborative control. *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics, 35*(4), 494-504.

Carlson, J., Murphy, R. R., & Nelson, A. (2004). Follow-up analysis of mobile robot failures. *Proceedings of the IEEE International Conference on Robotics and Automation*, 4987-4994.

Ceballos, A., Gomez, J., Prieto, F., & Redarce, T. (2009). Robot command interface using an audio-visual speech recognition system. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications 14th Iberoamerican Conference on Pattern Recognition (CIARP 2009)*, Berlin, Germany, 869-876.

Charles, L., Qixin, C., Zhu Xiao, X., & Zhang, Z. (2009). Gesture recognition system based on acceleration data for Robocup referees. *The 5th International Conference on Natural Computation, (ICNC '09*), *2*, 149-153. Washington, D.C.: IEEE.

Chen, T. L. & Kemp, C. C. A direct physical interface for navigation and positioning of a robotic nursing assistant. *Advanced Robotics, 25*, 605-627.

Cohen, M. S., Parasuraman, R., & Freeman, J. T. (1998). Trust in decision aids: A model and its training implications. *Proceedings of the International Command and Control Research and Technology Symposium*, 1-37. Monterey, CA.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

De Visser, E., Parasuraman, R., Freedy, A., Freedy, E., & Weltman, G. (2006). A comprehensive methodology for assessing human-robot team performance for use in training and simulation. *Proceedings of the Human Factors and Ergonomics Society*, 2639-2643.

Desai, M. (2007). *Sliding Scale Autonomy and Trust in Human-Robot Interaction.* Master of Science Thesis, University of Massachusetts Lowell.

Desai, M., & Yanco, H. A. (2005). Blending human and robot inputs for sliding scale autonomy. *The IEEE International Workshop on Robot and Human Interactive Communication*, 537-542.

Desai, M., Stubbs, K., Steinfeld, A., & Yanco, H. (2009). Creating trustworthy robots: Lessons and inspirations from automated systems. *Proceedings of the AISB Convention, New Frontiers in Human-Robot Interaction.*

Dewar, R. D., & Dutton, J. E. (1986). The adoption of radical and incremental innovations – an empirical- analysis. *Management Science, 32*(11), 1422-1433.

Dinka, D., Nyce, J. M., & Timpka, T. (2006). The need for transparency and rationale in automated systems. *Interacting with Computers, 18*(5), 1070-1083.

Dixon, S. R. & Wickens, C. D. (2006). Automation reliability in unmanned aerial vehicle control: A reliance-compliance model of automation dependence in high workload. *Human Factors, 48*(3), 474-486.

Duffy, B. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems, 42*, 177-190.

Duran, L., Fernandez-Carmona, M., Urdiales, C., Peula, J., & Sandoval, F. (2009). Conventional joystick vs. Wiimote; for holonomic wheelchair control. *Bio-Inspired Systems: Computational and Ambient Intelligence, Proceedings of the International Work-Conference on Artificial Neural networks (IWANN),* 1153-1160.

Durso, F. T., & Gronlund, S. D. (1999). Situation awareness. In F. T. Durso (Ed.), *Handbook of Applied Cognition* (pp. 283-314). New York, NY US: John Wiley & Sons Ltd.

Durso, F. T., & Sethumadhavan, A. (2008). Situation awareness: Understanding dynamic environments. *Human Factors, 50*(3), 442-448.

Durso, F. T., Feigh, K., Fischer, U., Morrow, D., & Mosier, K. (2011). *Automation in the cockpit: Toward a human-automation relationship taxonomy*. Unpublished manuscript, Georgia Institute of Technology, San Francisco State, and University of Illinois.

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A. Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Computer Studies*, *58*, 697-718.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*, 32-64.

Endsley, M. R. (2006). Situation Awareness. In G. Savendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd ed., pp. 528-542). New York: Wiley.

Endsley, M. R., Bolte, B., & Jones, D. G. (2003). *Designing for Situation Awareness: An Approach to Human-Centered Design*. London: Taylor & Francis.

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics, 42*(3), 462-492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*(2), 381-394.

Engelhardt, K. G., & Edwards, R. A. (1992). Human-robot integration for service robotics. In M. Rahimi & W. Karwowski (Eds.), *Human-Robot Interaction* (pp. 315-346). Washington, DC: Taylor & Francis, Ltd.

Erikson, E. H. (1997). *The Life Cycle Completed*. New York, N.Y.: Norton.

Ezer, N. (2007). *Collaborative human-multiagent teams: Towards a model of team effectiveness*. Unpublished manuscript, Georgia Institute of Technology.

Ezer, N. (2008). Is a robot an appliance, teammate, or friend? Age-related differences in expectations of and attitudes towards personal home-based robots. *Unpublished Dissertation*. Georgia Institute of Technology, Atlanta, GA.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2008). Age-related differences in reliance behavior attributable costs within a human-decision aid system. *Human Factors, 50*(6), 853-863.

Ezer, N., Fisk, A. D., & Rogers, W. A. (2009). Attitudinal and intentional acceptance of domestic robots by younger and older adults. *Lecture notes in computer science, 5615*, 39-48.

Faber, F., Bennewitz, M., Eppner, C., Gorog, A., Gonsior, C., Joho, D., Schreiber, M., & Behnke, S. (2009). The humanoid museum tour guide Robotinho. *Paper presented at the Robot and Human Interactive Communication* (RO-MAN), 891-896.

Falcone, R., & Castelfranchi, C. (2001). The human in the loop of a delegated agent: The theory of adjustable social autonomy. *IEEE Transactions on Systems Man and Cybernetics Part a-Systems and Humans, 31*(5), 406-418.

Feil-Seifer, D., Skinner, K., & Mataric, M. J. (2007). Benchmarks for evaluating socially assistive robotics. *Interaction Studies*, *8*(3), 423-439.

Few, D., Smart, W. D., Bruemmer, D., & Neilsen, C. (2008). "Seamless autonomy": Removing autonomy level stratifications. *Proceedings of the Conference on Human System Interactions,* 446-451.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems, 42*, 143-166.

Franklin, S., & Graesser, A. (1997). Is it an agent, or just a program? A taxonomy for autonomous agents. *Proceedings of the Third International Workshop on Agent Theories, Architectures, and Languages, Intelligent Agents III*, 21-35.

Gielniak, M. & Thomaz, A. (2011). Spatiotemporal correspondence as a metric for human-like robot motion. *Proceedings of the Human-Robot Interaction Conference (HRI).* Lausanne, Switzerland, 77-84.

Goetz, J., Kiesler, S., & Powers, A. (2003). Matching robot appearance and behavior to tasks to improve human-robot cooperation. *Proceedings of the 2003 IEEE International Workshop on Robot and Human Interaction Communication,* 55-60. Millbrae, CA.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.

Goodrich, M. A., & Olsen, D. R. (2003). Seven principles of efficient human robot interaction. *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, 1-5*, 3943-3948.

Goodrich, M. A., & Schultz, A. C. (2007). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction, 1*(3), 203-275.

Gorman, J. C., Cook, N. J., & Winner, J. L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics, 49*(12-13), 1312-1325.

Green, S. G., Gavin, M. B., & Aimansmith, L. (1995). Assessing a multidimensional measure of radical technological innovation. *IEEE Transactions on Engineering Management, 42*(3), 203-214.

Groom, V., & Nass, C. (2007). Can robots be teammates? *Benchmarks in human-robot teams. Psychological Benchmarks of Human-Robot Interaction: Special issue of Interaction Studies, 8*(3), 483-500.

Hancock, P. A., Billings, D. R., & Schaefer, K. E. (2011). Can you trust your robot? *Ergonomics in Design, 19*(3), 24-29.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors, 53*(5), 517-527.

Hearst, M. A. (1999). Mixed-initiative interaction: Trends and controversies. *IEEE Intelligent Systems*, 14-23.

Horrey, W. J., & Wickens, C. D. (2001). Supporting situation assessment through attention guidance: A cost-benefit and depth-of-processing. *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomic Society*. Santa Monica, CA: Human Factors & Ergonomics Society.

Huang, H.-M, (2004). Autonomy levels for unmanned systems (ALFUS) framework volume I: Terminology version 1.1. *Proceedings of the National Institute of Standards and Technology (NISTSP)*, Gaithersburg, MD.

Huang, H.-M, Messina, E. R., Wade, R. L., English, R. W, Novak, B., & Albus, J. S. (2004). Autonomy measures for robots. *Proceedings of the International Mechanical Engineering Congress (IMECE)*, 1-7.

Huang, H.-M, Pavek, K, Albus, J., & Messina, E. (2005). Autonomy levels for unmanned systems (ALFUS) framework: An update. *Proceedings of the SPIE Defense and Security Symposium*, *5804*, 439-448.

Huang, H.-M, Pavek, K., Novak, B., Albus, J. S., & Messina, E. (2005). A framework for autonomy levels for unmanned systems (ALFUS). *Proceedings of the AUVSI's Unmanned Systems North America*, 849-863, Baltimore, Maryland.

Huang, H.-M, Pavek, K., Ragon, M., Jones, J., Messina, E., & Albus, J. (2007). Characterizing unmanned system autonomy: Contextual autonomous capability and level of autonomy analyses. *The Unmanned Systems Technology IX, The International Society for Optical Engineering*, 6561.

Hyun, J., Gyeongho, K., & Youngjin, P. (2007). An application of speech/speaker recognition system for human-robot interaction. *International Conference on Control, Automation and Systems (ICCAS)* 1915-1918. Piscataway, NJ.

Inagaki, T. (2003). Adaptive automation: Sharing and trading of control. In E. Hollnagel (Ed.), *Handbook of Cognitive Task Design* (pp. 221–245). Mahwah, NJ: Erlbaum.

Kaber, D. B. and Riley, J. (1998). Adaptive automation of a dynamic control task based on workload assessment through a secondary monitoring task. In M. Scerbo and M. Mouloua (Eds.), *Automation Technology and Human Performance: Current Research and Trends* (pp. 129-133). Lawrence Erlbaum Associates: Mahwah, NJ.

Kaber, D. B., & Endsley, M. R. (2004). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomic Science*, *5*(2), 113-153.

Kaber, D. B., Onal, E., & Endsley, M. R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors and Ergonomics in Manufacturing, 10*(4), 409-430.

Kaber, D. B., Onal, E., & Endsley, M.R. (2000). Design of automation for telerobots and the effect on performance, operator situation awareness, and subjective workload. *Human Factors & Ergonomics in Manufacturing*, *10*(4), 409-430.

Kaber, D. B., Perry, C. M., Segall, N., McClernon, C. K., & Prinzell, L. J. (2006). Situation awareness implications of adaptive automation for information processing in air traffic control task. *International Journal of Industrial Ergonomics, 36*, 447-462.

Kaber, D. B., Riley, J. M., Tan, K.-W., & Endsley, M. R. (2001). On the design of adaptive automation for complex systems. *International Journal of Cognitive Ergonomics*, *5*(1), 37-57.

Kaber, D. B., Wright, M. C., & Sheik-Nainar, M. A. (2006). Investigation of multi-modal interface features for adaptive automation of a human-robot system. *International Journal of Human-Computer Studies, 64*, 527-540.

Kahn, P. H., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., & Miller, J. (2007). What is a human? Toward psychological benchmarks in the field of human-robot interaction. *Interaction Studies, 8*(3), 363-390.

Kant, I. (1967). *Kant: Philosophical Correspondence, 1795-99.* (A. Zweig, Ed.). Chicago: University of Chicago Press.

Kemp, C., Anderson, C., Nguyen, H., Trevor, A. & Xu, Z. (2008). A point-and-click interface for the real world: Laser designation of objects for mobile manipulation. *Proceedings of Human-Robot Interaction (HRI)*, 241-248.

Khan, Z. (1998). Attitude towards intelligent service robots. *Numerical Analysis and Computer Science Tech.. Rep. (TRITA-NA-P9821)*. Stockholm Sweden: Royal Institute of Technology.

Kim, T., & Hinds, P. (2006). Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN),* 80-85.

Lee, J. D. (2006). Human factors and ergonomics in automation design. In G. Savendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd ed., pp. 1570-1596). New York: Wiley.

Lee, J. D., & Moray, N. (1992). Trust, control strategies, and allocation of function in human-machine systems. *Ergonomics, 35*, 1243-1270.

Lee, J. D., & Moray, N. (1994). Trust, self-confidence and operator's adaptation to automation. *International Journal of Human Computer Studies, 40*, 153-184.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*, 50-80.

Lorenz, B., Di Nocera, F., Rottger, S., & Parasuraman, R. (2001). The effects of level of automation on the out-of-the-loop unfamiliarity in a complex fault-management task during simulated spaceflight operations. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44-48.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human-human and human-automation trust: An integrative review. *Theoretical Issues in Ergonomics, 8*(4), 277-301.

Maheswaran, R. T., Tambe, M., Varakantham, P., & Myers, K. (2004). Adjustable autonomy challenges in personal assistant agents: A position paper. In M. Nickles, M. Rovatsos, and G. Weiss (Eds.) *Agents and Computational Autonomy: Potential, Risks, Solutions, Lecture Notes in Computer Science*, *2969*, 187-194.

Michaud, F., Boissy, P., Labonte, D., Briere, S., Perreault, K., Corriveau, H., Grant, A., Lauria, M., Cloutier, R., Roux, M.-A., Iannuzzi, D., Royer, M.-P., Ferland, F., Pomerleau, F., & Letourneau, D. (2010). Exploratory design and evaluation of a homecare teleassistive mobile robotic system. *Mechatronics, 20*(7), 751-766.

Milgram, P., Rastogi, A., & Grodski, J. J. (1995). Telerobotic control using augmented reality. *IEEE International Workshop on Robot and Human Communication,* 21-29.

Miller, C.A., & Parasuraman, R. (2007). Designing for flexible interaction between humans and automation: Delegation interfaces for supervisory control. *Human Factors*, *49*(1), 57-75.

Mitzner, T. L., Smarr, C. A., Beer, J.M., Chen, T. L., Springman, J.M., Prakash, A., Kemp, C. C., & Rogers, W. A. (2011). *Older adults' acceptance of assistive robots* (HFA-TR-1105). Atlanta, GA: Georgia Institute of Technology, School of Psychology, Human Factors and Aging Laboratory.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive atomation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*, 44-58.

Mori, M. (2005). Bukimi no tani. The uncanny valley (K. F. MacDorman & T. Minato, Trans.). *Proceedings of the Humanoid-2005 Workshop: Views of the Uncanny Valley*, Tsukuba, Japan. (Reprinted from *Energy, 7*(4), 33-35, 1970).

Mosier, K. L., Skitka, L. J., Dunbar, M., & McDonnell, L. (2001). Aircrews and automation bias: The advantages of teamwork? *The International Journal of Aviation Psychology, 11*(1), 1-14.

Murphy, R. R. (2000). From teleoperation to autonomy. *Introduction to AI Robotics* (pp. 13-40). Cambridge, MA: The MIT Press.

Murphy, R. R. (2000). Robotic paradigms. *Introduction to AI Robotics* (pp. 1-13). Cambridge, MA: The MIT Press.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81-103.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies, 45*(6), 669-678.

Nass, C., Moon, Y., Fogg, B. J., & Reeves, B. (1995). Can computer personalities be human personalities? *International Journal of Human-Computer Studies, 43*(2), 223-239.

Nass, C., Steuer, J., Henriksen, L., & Dryer, D. C. (1994). Machines, social attributions, and ethopoeia: Performance assessments of computers subsequent to 'self-' or 'other-' evaluations. *International Journal of Human-Computer Studies, 40*(3), 543-559.

Nguyen, H., Jain, A., Anderson, C., & Kemp, C. (2008). A clickable world: Behavior selection through pointing and context for mobile manipulation. *IEEE/RJS International Conference on Intelligent Robots and Systems (IROS)*, 787-793.

Nicolescu, M., & Mataric, M. (2003). Natural methods for robot task learning: Instructive demonstrations, generalization and practice. *The 2nd International Joint Conference on Autonomous Agents and Multi-Agent Systems*, *2*, 241-248. Melbourne, Australia.

Nof, S. Y. (2009). Automation: What it means to us around the world. In S. Nof (Ed.), *Handbook of Automation* (pp. 13-52). Berlin, Germany: Springer-Verlag.

Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. *AI & Society, 20*(2), 138-150.

Nomura, T., Kanda, T., Suzuki, T., & Kato., K. (2004). Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots. *Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication (RO-MAN)*, (pp. 35-40).

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006a). Measurement of negative attitudes toward robots. *Interaction Studies, 7*(3), 437-454.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006b). Measurement of anxiety toward robots. *The 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN06)*, 372-377. Hatfield, UK.

Olsen, D. R., & Goodrich, M. A. (2003). Metrics for evaluating human-robot interactions. *Proceedings of NIST Performance Metrics for Intelligent Systems Workshop*.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors, 39*(2), 230-253.

Parasuraman, R., & Wickens, C. D. (2008). Humans: Still vital after all these years of automation. *Human Factors, 50*(3), 511-520.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems Man and Cybernetics Part A: Systems and Humans, 30*(3), 286-297.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2008). Situation awareness, mental workload, and trust in automation: Viable, empirically supported cognitive engineering constructs. *Journal of Cognitive Engineering and Decision Making, 2*(2), 140-160.

Piaget, J. (1932). *The Moral Judgment of a Child*. Glencoe, IL: The Free Press.

Poncela, A., Urdiales, C., Perez, E. J., & Sandoval, F. (2009). A new efficiency weighted strategy for continuous human/robot cooperation in navigation. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans, 39*(3), 486-500.

Riley, J. M., & Endsley, M. R. (2004). The hunt for situation awareness: Human-robot interaction in search and rescue. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting,* 693-697.

Rovira, E., Zinni, M., & Parasuraman, R. (2002). Effects of information and decision automation on multi-task performance. *Proceedings of the Human Factors & Ergonomics Society 46th Annual Meeting*. Santa Monica, CA, 327-331.

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd ed.). Upper Saddle River, NJ: Pearson Education, Inc.

Salthouse, T. A. (1991). The need for, and requirements of, theories of cognitive aging. In T. Salthouse (Ed.), *Theoretical Perspective on Cognitive Aging* (pp. 1-31). Hillsdale: Lawrence Erlbaum Associates, Inc.

Sanchez, J. (2006). *Factors that affect trust and reliance on an automated aid*. Ph.D. Dissertation, Georgia Institute of Technology.

Scerbo, M. W., Freeman, F. G., & Mikulka, P. J. (2003). A brain-based system for adaptive automation. *Theoretical Issues in Ergonomic Science*, *4*(1-2), 200-219.

Scerbo, M.W. (2001). Adaptive automation. In W. Karwowski (Ed.), *International encyclopedia of ergonomics and human factors* (pp. 1077-1079). London: Taylor and Francis, Inc.

Scerri, P., Pynadath, D. V., & Tambe, M. (2002). Towards adjustable autonomy for the real world. *Journal of Artificial Intelligence Research*, *17*, 171-228.

Scholtz, J., Antonishek, B., & Young, J. (2004). Evaluation of a human-robot interface: Development of a situational awareness methodology. *Proceedings of the International Conference on System Sciences*, 1-9.

Scholtz, J., Young, J., Drury, J. L., & Yanco, H. A. (2004). Evaluation of human-robot interaction awareness in search and rescue. *IEEE International Conference on Robotics and Automation (ICRA)*, *3*, 2327-2332.

Sellner, B., Heger, F. W., Hiatt, L. M., Simmons, R., & Singh, S. (2006). Coordinated multiagent teams and sliding autonomy for large-scale assembly. *Proceedings of the IEE – Special Issue on Multi-Robot Systems, 94*(7), 1425-1444.

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of underseaa teleoperators (Man-Machine Systems Laboratory Report). Cambridge: MIT.

Skinner, B. F. (1978). *Reflection on Behaviorism and Society*. Englewood Cliffs, N.J.: Prentice-Hall.

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bais. *International Journal of Human-Computer Studies, 52*, 701-717.

Steinfeld, A., Fong, T., Kaber, D., Lewis, M., Scholtz, J., Schultz, A., & Goodrich, M. (2006). Common metrics for human-robot interaction. *Proceedings of Human-Robot Interaction Conference*, 33-40.

Stubbs, K., Hinds, P. J., & Wettergreen, D. (2007). Autonomy and common ground in human-robot interaction. *IEEE Intelligent Systems, 22*(2), 42-50.

Takayama, L., Marder-Eppstein, E., Harris, H., & Beer, J. M. (2011). Assisted driving of a mobile remote presence system: System design and controlled user evaluation. *Proceedings of the International Conference on Robotics and Automation: ICRA 2011*, 1883-1889. Shanghai, CN.

Tarn, T.-J., Zi, N., Guo, C. & Bejczy, A. K. (1995). Function-based control sharing for robotics systems. *Proceedings on IEEE International Conference on Intelligent Robots and Systems, 3*, 1-6.

Thrun, S. (2004). Toward a framework for human-robot interaction. *Human-Computer Interaction, 19*(1-2), 9-24.

Tiwari, P., Warren, J., Day, K. J., & MacDonald, B. (2009). Some non-technology implications for wider application of robots to assist older people. *Proceedings of the Conference and Exhibition of Health Informatics, New Zealand*.

Tsang, P. S. & Vidulich, M. A. (2006). Mental workload and situation awareness. In G. Savendy (Ed.), *Handbook of Human Factors and Ergonomics* (3rd ed., pp. 243-268). New York: Wiley.

Urdiales, C., Poncela, A., Sanchez-Tato, I., Galluppi, F., Olivetti, M., & Sandoval, F. (2007). Efficiency based reactive shared control for collaborative human/robot navigation. *Proceedings from the IEEE/RSJ International Conference on Intelligent Robots and Systems* (IROS), 3586-3591.

Venkatesh, V., & Davis, F. D. (2000). A theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science, 46*(2), 186-204.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425-478.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *Knowledge Engineering Review, 10,* 115-152.

Yanco, H. & Drury, J. (2004a). Classifying human-robot interaction: an updated taxonomy. *IEEE International Conference on Systems, Man and Cybernetics, 3*, 2841-2846.

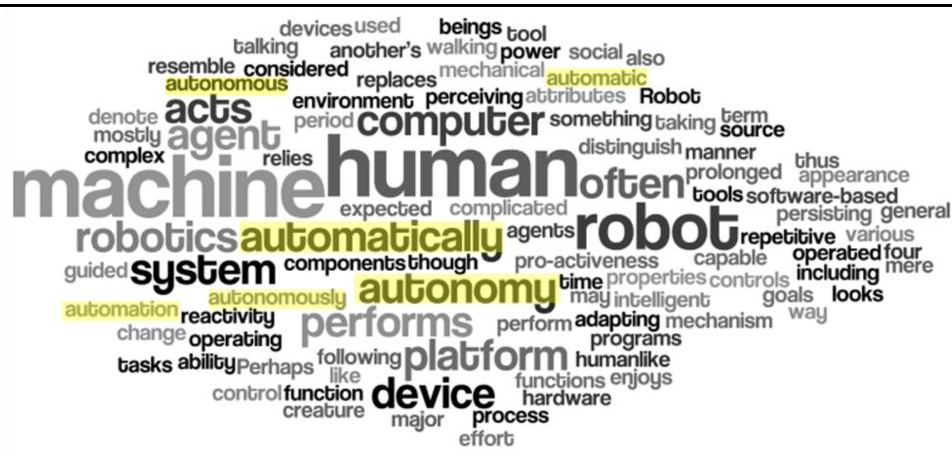
Yanco, H. & Drury, J. (2004b). "Where am I?" Acquiring situation awareness using a remote robot platform. *Proceedings of the IEEE Conference on Systems, Man, and Cybernetics*, *3,* 2835-2840.

Young, J., Hawkins, R., Sharlin, E., and Igarashi, T. (2009). Toward acceptable domestic robots: Applying insights from social psychology. *International Journal of Social Robotics*, *1*(1), 95-108.

Ziemke (2008). On the role of emotion in biological and robotic autonomy. *BioSystems*, *91*, 401-408.

8. Appendix A

Definition of Robot/Agent	Source
Robot, any automatically operated machine that replaces human effort, though it may not resemble human beings in appearance or perform functions in a humanlike manner.	(Encyclopedia Brittanica, http://www.britannica.com/EBchecked/topic/505818/robot)
"As in automation, robotics also relies on four major components, including a platform, autonomy, process, and power source, but in robotics, a robot is often considered a machine, thus the platform is mostly a machine, a tool or device, or a system of tools and devices."	(Nof, 2009, p.19)
"An agent is something that acts. But computer agents are expected to have other attributes that distinguish them from mere "programs," such as operating under autonomous control, perceiving their environment, persisting over a prolonged time period, adapting to change, and being capable to taking on another's goals."	(Russell & Norvig, 2003, p. 4)
"A robot is a machine that looks like a human being and performs various complex acts (as walking or talking) of a human being; a device that automatically performs complicated often repetitive tasks; a mechanism guided by automatic controls."	(Merrium-Webster Dictionary, http://www.merriam-webster.com/dictionary/robot)
"An intelligent robot is a mechanical creature which can function autonomously"	(Murphy, 2000, p. 3)
"Perhaps the most general way in which the term agent is used is to denote a hardware or software-based computer system that enjoys the following properties:" autonomy, social ability, reactivity, and pro-activeness.	(Wooldridge & Jennings, 1995)



Word cloud of robot/agent definitions; autonomy depicted as a major component of language used in definitions.