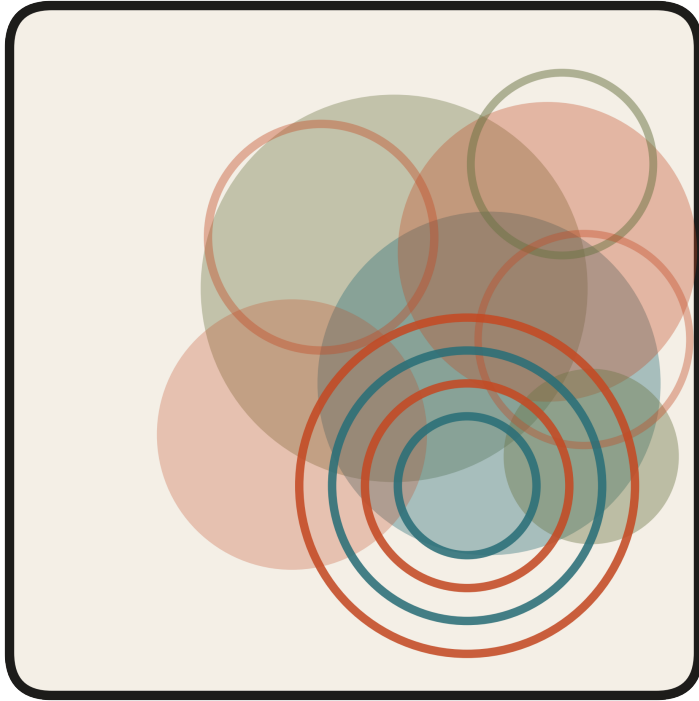




A Handbook of Human-AI
and Human-Automation Interaction

Working Draft

Shane T. Mueller

© 2026 Shane T. Mueller and the chapter authors.

This work is licensed under a [Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International License \(CC BY-NC-ND 4.0\)](https://creativecommons.org/licenses/by-nc-nd/4.0/). You may copy and redistribute the material for non-commercial purposes provided appropriate credit is given. You may not remix, transform, or build upon the material.

How to cite this book.

Mueller, S. T. (Ed.). (2026). *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

ISBN: [TBD]

Edition: Draft manuscript, October 5, 2026.

About the cover. The artwork on the title page is an edition mark created for this handbook, not a publisher-supplied jacket design. It was generated with `tools/make_cover.py`: concentric rings and overlapping translucent disks in the handbook palette (teal, rust, olive on warm paper). The mark is meant to suggest human–machine coupling without depicting a humanoid robot or chat interface. It was designed by asking Cursor/Grok to create an image inspired by two other book covers: the classic text “Human judgments of optimality” by Shelly and Bryan (1964), and the fictional book featured in the Amazon Prime series *Patriot: The Integral Principles of the Structural Dynamics of Flow*, by Leslie Claret.

Contact: [TBD: editor email or project URL]

Printed in the United States of America.

How to Cite This Book

Mueller, S. T. (Ed.). (2026). *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Preface

This handbook is a practical guide to *human-centered* artificial intelligence and automation. It is written primarily for human factors engineers, cognitive systems engineers, HCI and UX practitioners, and students who need working literacy in AI and automation without becoming machine-learning specialists, and desire entry points into the large and diverse literature on AI and automation.

Each chapter began as a research and writing project in Human-AI Interaction (HF 5430) at Michigan Technological University course in Spring 2024, taught by Professor Mueller. For each chapter, small teams of students were given basic readings and topic areas, and they eventually produced draft chapters covering the topics. Once the class was completed, Dr. Mueller spent time during a sabbatical revising, editing, augmenting, checking, and improving every section of each chapter.

Shane T. Mueller
Houghton, Michigan

Acknowledgments

About the Contributors

Shane T. Mueller is a Professor of Psychology in the School of Health, Human, and Biological Systems at Michigan Technological University. His research involve human-AI interaction, computational modeling of behavior, computer-based testing and evaluation, and naturalistic approaches to studying complex human behavior.

Yifan Wang is a Ph.D. candidate and Graduate Research Assistant in Civil Engineering at Michigan Technological University. His research focuses on artificial intelligence, robotics, and human factors in construction, with particular emphasis on human–AI interaction and human–robot collaboration.

Editor: Shane T. Mueller, Michigan Technological University.

Contents

How to Cite This Book	iii
Preface	v
Acknowledgments	vii
About the Contributors	ix
1 Demystifying Automation, Artificial Intelligence, and Intelligent Software Tools and Applications	1
1.1 Introduction	1
1.2 General Principles of Intelligent Software Tools	3
1.3 The purpose and function of intelligent systems	7
1.4 Architecture: Representations and Algorithms	17
1.5 Applications of Intelligent Systems	29
1.6 Summary and Discussion	35
2 Human-AI Interaction Research: An Overview	37
2.1 Introduction	37
2.2 Schools of thought for designing AI and automated systems for human users	38
2.3 Evaluating AI and automation for human use	46
2.4 Summary	51
3 Stages of Technology Adoption and Development for Human-Centered AI Tools and Methods	53
3.1 Introduction	53
3.2 Technology Acceptance Model (TAM) and AI-specific revisions	54

- 3.3 Product Development Lifecycles 56
- 3.4 Human Readiness Levels 57
- 3.5 Threat-based stage analysis 58
- 3.6 Critical Periods 58
- 3.7 Understanding needs of users/supporters and other stakeholders 61
- 3.8 Understanding user expectations 65
- 3.9 Cost-effective, efficient, and phase-appropriate evaluation . . 67
- 3.10 Identifying capabilities and limitations of the system and the users 70
- 3.11 Understanding the Work Domain and Sociotechnical System . 71
- 3.12 Enable Users to Learn a System 72
- 3.13 Summary 73
- 4 Human-like AI: The Turing Test, Artificial General Intelligence, and Anthropomorphism of AI 75**
 - 4.1 Introduction 75
 - 4.2 The Turing test 77
 - 4.3 Chatbots: from ELIZA to ChatGPT 80
 - 4.4 Moravec’s Paradox 83
 - 4.5 Artificial General Intelligence 83
 - 4.6 Anthropomorphism of AI 86
- 5 Trust in Automation: Definitions, Models, and Related Concepts 97**
 - 5.1 Introduction 97
 - 5.2 Definitions of Trust and Trustworthiness 98
 - 5.3 Human–Human (Interpersonal) vs. Human–Machine Trust . . 103
 - 5.4 Theories and Models of Trust in Automation 106
 - 5.5 Contributors to trust and trustworthiness 116
 - 5.6 The flavors of trust 117
 - 5.7 Trust in “Artificial Intelligence” 121
 - 5.8 Conclusions 128
- 6 Measuring Trust in Automation and AI: Key Points, Methods, and Scale Applications 131**
 - 6.1 Introduction 132
 - 6.2 Methods for measuring trust in automation 132
 - 6.3 Contexts of trust measurement 133
 - 6.4 Measures and metrics for characterizing trust in automation . 135
 - 6.5 Questionnaires and Scales Used to Measure Trust in Automation145
- 7 How We Understand AI and Automation: Mental Models and Transparency 155**
 - 7.1 Introduction: Are internal models *necessary*? The Good Regulator Theorem 156

7.2	Mental Models: Definitions and Uses	157
7.3	Kinds of Mental Models	157
7.4	Mental Models across Disciplines	162
7.5	Eliciting and Measuring Mental Models	169
7.6	Transparency, Observability, and Explainability	173
7.7	Practical Application of Mental Models	175
7.8	Conclusions	178
Glossary		181
Bibliography		201
Index		233

Demystifying Automation, Artificial Intelligence, and Intelligent Software Tools and Applications

Christopher Cischke, Blade Frisch, Erin Matas, Shane T. Mueller, Lauren Sprague, Katie Ulinski, Samantha Walker, Joanne Wang, Yifan (Jeff) Wang, and Brandon Woolman

Michigan Technological University

Abstract

Algorithms implementing automation, artificial intelligence (AI) and other intelligent software tools are not a unitary thing. They involve a number of different concepts, purposes, algorithms, and applications. Understanding the basics of their operation can be important for improving human usability, but human factors researchers and practitioners and cognitive systems engineers interested in designing and evaluating usable AI and automation often have limited experience in such systems. The goal of this chapter is to demystify some of the common concepts in AI, automation, and intelligent software. To do this, we examine general principles of operation, representations, learning and computation algorithms, and conclude with an examination of some of the more common applications of AI and automation systems.

1.1 Introduction

When we discuss AI, Machine Learning (ML), automation, optimization, and other intelligent software tools, we often group them all together into a single general construct when asking how humans interact or understand them. But

in reality, AI, ML, optimization, and automation represent dozens of different approaches applied to many different problems. Although these different approaches are well understood by researchers implementing AI or ML, they often represent black boxes to users and to cognitive systems engineers and UX/UI professionals whose job it is to think about how humans will interact with the system. Consequently, one of the goals of this chapter is to help demystify many of the algorithms and approaches used to implement automation and artificial intelligence, and give some concrete examples to help novices understand the algorithms. Although a deep understanding of particular computations may not always be necessary, having a good mental model of the process can help users understand the boundary conditions of the system: what it is good at, and where it is not appropriate. Understanding these boundaries helps users know where to trust the system and where to anticipate that it may fail. Furthermore, it helps users understand the errors these systems make. Without being able to interpret why an error is made, a naive user may attribute it to the system being bad or broken. When systematic errors can be anticipated, this helps encapsulate and contextualize the problems, provide greater understanding of the boundary conditions, and develop workarounds.

Consequently, this chapter will discuss many aspects of the algorithms underlying AI and automation, with the goal of providing some basic understanding of the landscape of this domain. It will be especially helpful for human factors researchers and practitioners who may not have a strong background in the technical side of AI and automation.

We will organize this chapter by thinking about AI and automation at several levels of abstraction, which might all be considered part of the **system architecture**. We will start with a few very **general principles** underlying many kinds of systems. Following that, we will examine some of the common knowledge and information **representation** approaches used across such systems (e.g., how information or data is modeled; for example as a markov network, a tree, a vector). Then, we will discuss some of the distinct **inference/optimization** approaches used (the way in which training data is transformed into the representation, learned from data, or parameters are optimized). Then, we will discuss different classes of **computational algorithms** (the computational process by which input is transformed into output; e.g., via a convolutional neural network). Finally, we will describe a number of **application domains** and the systems used in particular domains. It may be true that many individual systems involve a variety of different approaches along each of these dimensions, and many applications of AI use multiple combined systems to solve the problem. Later chapters take up human-centered methods (Chapter 2), lifecycle methods (Chapter 3), anthropomorphism/AGI (Chapter 4), trust (Chapters 5–6), mental models/transparency (Chapter 7), explainability (Chapter 8), ethics/fairness (Chapter 9), user support (Chapter 10), large language models and generative AI (Chapter 11), and human–machine

teaming (Chapter 12). This chapter focuses on technical demystification.

Although users of such intelligent software tools may not need to know detailed aspects of the computations underlying the system, oftentimes it can make a difference for a human factors researcher. For example, understanding aspects of the architecture of the system may help identify aspects of the system that could be used to improve usability. Understanding inference algorithms may help determine whether certain adaptive systems are practical in terms of time or cost. When thinking about automated systems, a useful distinction is training versus deployment: many systems estimate parameters or generate rules during training, then process data and create output when deployed, and although the algorithms for training/creation need to be coupled with the system deployed algorithm, they are often distinct, sometimes general learning or optimization algorithms can be used across different classes of algorithms, and the choice can impact both costs and performance.

1.2 General Principles of Intelligent Software Tools

In this section, we will describe and define some of the most important high-level approaches to AI, automation, and other intelligent software systems. In many cases, these concepts are very general and have no agreed definition, but it is useful to understand how these concepts are used by various communities.

1.2.1 Automation

Automation has been defined as *“The process of structuring a task, environment, or system to reduce labor and decision-making time through offloading labor and decision-making to a human-created system”* (Goldberg, 2011). While often described via circular definitions (automation is often defined as simply, “to automate”), automation is largely understood as processes that reduce human labor by passing that labor and activity to a machine, algorithm, or other devised mechanism. Automation covers a wide range of technology developed over hundreds of years, and includes simple mechanical systems (a parking meter automates the job of a meter reader; a stop light automates the job of a traffic cop, an industrial machine may automate the job previously performed by a worker, an autohelm controller automates the job of keeping a boat on course).

Human factors researchers have classified how much and what kind of work is automated. Levels-of-automation scales (Sheridan and Verplank, 1978; Parasuraman et al., 2000) distinguish stages such as information acquisition, analysis, decision selection, and action implementation, each automatable to different degrees. The International Society of Automotive Engineers (SAE On-Road Automated Vehicle Standards Committee, 2014) has followed suit to

create SAE J3016, a taxonomy that specifies six levels of automation—from completely human drivers to completely autonomous.

To automate something is to reduce human involvement in a task. It can be as simple as the natural proceduralization of a task to reduce decision-making labor or as complex as training an algorithm or building a machine to do the task for you. Creating an automated system generally takes time and effort but reduces time and effort later on. Additionally, not all tasks/environments/systems can or should be automated. Automation is useful within task domains that are repetitive, time-consuming, or error-prone, but may not be appropriate in situations that heavily rely on in-the-moment decision making.

Some technologies that are considered or appear to be automation may not meet our definition, because they are designed to shift the labor burden rather than reduce it. For example, grocery store self-checkouts and self-service gas station pumps are technologies that replace a job previously held by a human, but their intent is to make it possible to shift the human labor safely or economically to the buyer to reduce costs. Similarly, IKEA furniture, tools, and other housewares sold with 'assembly required' replace a factory worker and reduce shipping costs, but simply shift work from one human to another, or perhaps from a machine in a factory to a human in their living room. In contrast, an ATM automates the work of the bank teller (counting money keeping records) but does not materially shift the work burden of the customer, and might be considered true automation.

1.2.2 Artificial Intelligence

Artificial Intelligence (AI) is a blanket term for many different approaches in many different communities. Wang (2019) provided one of the most exhaustive and careful arguments understanding and defining AI. Wang argued that hundreds of forms of human intelligence have been defined (see Monett and Lewis, 2018), and that AI can be understood as an abstraction of these human intelligences. Moreover, human intelligence can be understood in general as a capacity to adapt to an environment with limited information. Wang argues that AI typically involves one of five kinds of approaches to simulate or creating human-like intelligence: Structure (simulating a brain), Behavior (simulating intelligent behavior), Capability (ability to solve problems effectively), Function (possessing of commonly-understood general cognitive functions such as memory, learning, planning, perception, etc.), and Principle (capable of fundamental principles such as rationality). Although this approach is not universally embraced (see Monett et al., 2020), many approaches to AI fit into one or another of these categories.

1.2.3 Machine Learning

Machine learning is a broad term to describe a variety of computational algorithms that are typically involved in producing behaviors or representations (classifications, labels, categories, control, generative AI systems) based on data, rather than relying on hard-coded or programmed rules (El Naqa and Murphy, 2015). An important distinction in machine learning is among *supervised learning* (predicting from labeled examples), *unsupervised learning* (identifying structure in unlabeled data), and *reinforcement learning* (policies via reward from interaction; see Section 4.4). Many modern systems combine these (e.g., pretraining then fine-tuning). The notion of 'learning' is quite general as well, and although may involve specific *learning algorithms* such as reinforcement learning or TD-learning (see Sutton and Barto, 2018), it may also involve other means of optimization (e.g., as in support vector machines, genetic algorithms, or Bayesian inference), or simply maintaining a database of past incidents (as in Case-based reasoning systems and K-nearest-neighbor classifiers). Also, machine learning systems can be contrasted with rule-based expert systems in which rules are typically identified via subject-matter experts using knowledge-elicitation approaches, although rules can also be generated via algorithms.

1.2.4 Optimization

Optimization is a fundamental concept relevant to AI, ML, and other intelligent software systems. Its origins come from calculus, which can be used to identify minima and maxima on a multivariate function. Generalizing this, many intelligent software tools involve optimizing some cost or loss function: finding a set of input values (parameters) that lead to the best outcome. Within AI, ML, and automation, optimization can be defined as the systematic and iterative process of improving the accuracy of a machine learning model or algorithm's output through testing and refinement. It can also be used to describe the process by which a system/task/environment is iteratively improved through the usage of machine learning and/or AI algorithms (Surianarayanan et al., 2023).

Optimization is used in two ways in intelligent software tools. First, problems can be framed in terms of a set of input parameters or choices (states) within a space of options, and each possible state can be associated with a cost or benefit. This use of optimization attempts to find the optimal (or at least good) parameter state that minimizes cost or maximizes benefit, with cost and benefit defined very flexibly.

The basic framework of optimization in AI and automation is that the effectiveness of behavior produced by an intelligent system can be incorporated into a cost function relevant to the particular domain and algorithm. Cost may

incorporate time, efficiency, monetary costs, or other heuristic values.

For example, solving supply chain, shipping, and warehousing problems involve finding a way to minimize both costs and delivery time by determining the best way to store and ship goods to customers. Here, cost might be related to shipping and storage costs, whereas the input parameters would be the way in which a product is shipped, when it is sent and delivered, and from which source it originates. Game-playing and route-finding AI agents also frequently work from an optimization framework. For example, an AI chess player searches future game states that might result from a particular move to identify moves that minimize potential losses or maximize potential benefits.

Another way optimization is used is to describe inference processes in machine learning—the way that parameters are determined in a way that minimizes cost, error or “loss” functions. Here, the goal is to find parameter settings that optimize a system for a given set of data, so that future behavior will be efficient or effective. For a simple linear regression model, optimization is typically done via least-squares, but for more complex models other techniques are used. This optimization is often thought of as ‘learning’, as it allows a model to incorporate information from many cases in order to make future decisions. For example, an image classifier built with a neural network undergoes training, and typically uses a backpropagation and a learning algorithm to set weights in the network that minimize error rates in classification. Here, rather than finding the optimal solution to a particular problem, the optimization attempts to find optimal settings that will solve future problems effectively. In all such cases, there is a risk of overfitting: identifying parameters that optimize the cost for the training set, to the detriment of cases outside the training set.

1.2.5 Humans in the loop: From Cybernetics to Cognitive Systems Engineering

These different high-level constructs all need to be understood in the context of the human work they are augmenting, replacing, displacing, or improving. The scientific study of this has occurred under many labels, often considered a specialization of Human Factors Engineering and Psychology. One of the earliest terms was Wiener’s (1948) concept of cybernetics, or as his book’s subtitle states “*Control and Communication in the Animal and the Machine.*” The subfield of Cognitive Systems Engineering (CSE) (Woods and Roth, 1988) emerged from Cybernetics to deal largely with how to evaluate and design automated systems. Here, the perspective is shifted from a human operating a system to a human being *part of* the system. Although human-computer interaction (HCI) originally focused on ‘dumb’ computer systems, it is increasingly involving interaction with intelligent systems. Other subdisciplines such as human-centered AI (Shneiderman, 2022) are also popular, as well as many other approaches we will cover in other chapters (HRI: Human-Robot

Interaction; coactive design; mixed-initiative interaction, social robotics, just to name a few).

It is probably true that all automation shifts human work and creates new problems while it solves others. For example, Bainbridge (1983) described “ironies of automation”, including that more automation can leave humans with harder residual tasks (such as monitoring and exception handling) that are poorly supported, and that the times when human control is needed will require skill (manual or other) that are no longer as well developed. Endsley (2023) argued that Bainbridge’s warnings are still relevant in the age of artificial intelligence. Others have identified automation-induced complacency and automation bias (Wickens et al., 2015), and shifting bottlenecks to different kinds of tasks such as supervision and control (Woods, 1996), as similar specific ways that automation does not simply replace human work; it transforms it—often adding monitoring, coordination, and recovery demands.

1.3 The purpose and function of intelligent systems

Now that we are familiar with some of the broader terms, an important next step is to understand the general functional classes of systems. In many cases, from a human factors perspective, knowing the purpose and function of a system is more important than understanding the underlying logic, because it can help identify the goals of the user. There are many grey areas and overlaps among systems, and this is not comprehensive, but a brief overview includes:

- **Numerical prediction or estimation.** Regression analysis is a classic case of this—given some things you know about an object, can you predict a numerical value? This is how Carvana can give you a price for your car (incorporating its model, age, mileage, condition, and features), but also useful as a building block for many other applications.
- **Categorical classification and detection.** Given information about a case, can you identify its class? Face recognition and other machine vision algorithms are a prototypical example, but the breadth and use of classification algorithms are vast, from automated driving (detecting other cars, locations, people) to marketing, business, target detection, cancer diagnosis, etc.
- **Control and regulation.** Perhaps the most widely used (and simplest) intelligent systems, these sense a state while controlling inputs and attempt to manage or regulate the system for some purpose. Includes thermostats, cruise control, factory automation, autopilots.
- **Rank, retrieve, and recommend.** Systems that attempt to produce information based on input. Including search engines, music and movie recommendation systems, dating services, shopping applications.

- **Cluster, compress, and collate.** Generally 'unsupervised' systems that attempt to build a model from data that extract important regularities and similarities, including PCA and cluster analysis. Used in many classification systems to build features, cluster analysis for marketing, self-organizing maps for summarizing data patterns.
- **Decide/Plan/Act.** Extensions of control systems that have a representation of a problem space (often a network) and can reason about costs and benefits of different moves within the problem space, they reason to achieve a goal that is either carried out or recommended. Applications include GPS systems, logistics/supply chain management, self-driving vehicles and video game AI.
- **Constraint/schedule satisfaction.** Systems that attempt to achieve optimal mappings of resources based on constraints. E.g., airline crew assignment, supply-chain management, med school student matching.
- **Generate/Synthesize Content.** Algorithms that create new information—from predictive text entry, to algorithmic music, to image, video, and text generation.

These basic categories of functions map onto overlapping sets of algorithms, each with different strengths and weaknesses. It is important to understand that AI and automation is not a single algorithm with a single function. More details are provided about each of these in the following subsections:

1.3.1 Regression and numerical prediction models

Regression models are often not even considered AI or automation as they are fairly simple and well-studied. However, they do form important components in many other more complex systems, and there are many extensions (e.g., generalized additive models, non-linear and non-parametric regression, auto-regression) that are quite sophisticated, and all such systems are incredibly useful in many contexts. Although they are often used to make inferences about data in experimental contexts, they are also useful as predictive models and are considered machine learning models, insofar as they are 'trained' with a data set and can then be used to predict new cases. Furthermore, models with a variety of architectures far outside standard regression are useful—consider models that predict the temperature or amount of rain expected over the next 10 days, the price of a house or vehicle, or the expected value of a farmer's crop in the future when it is harvested, or the expected re-entry point and speed of a spaceship.

These models typically make predictions based on past cases. New cases outside their experience will necessarily be extrapolations that are more likely to make errors. This illustrates how, for humans using the system, its *competence envelope* needs to be understood, so that the user can know when to trust it, and when to not trust it.

1.3.2 Classification and classifier systems

One of the simplest classification algorithms is just a transformed regression model known as logistic regression. Here, instead of modeling a numerical value directly, you model the *probability* that the outcome is one of two cases (often through negation; such as 'has disease' versus 'does not have disease'). Many different approaches can be used, but the ability to place a label on input so that different courses of action can be taken is very powerful. For example, classifiers are used to identify individuals in photos, to detect copyrighted content used on youtube, to identify defects in manufacturing processes, to perform content filtering of images, text, and speech on social media, to determine diagnoses based on symptoms or radiology, and many more.

Classification typically requires large datasets of known cases that can be used to train an algorithm to distinguish new cases. Once trained, classification systems can then be distributed and used for almost any purpose. But just as with numerical prediction models, it is important to understand from a user perspective that the further the new cases are from the training, the less reliable the system is likely to be, and so its competence envelope depends on the training set.

Furthermore, many classifiers pick up on correlated features in the training set that are not easily detected by humans but are only incidental to the category. For example, most published and labeled pictures of birds are already selected by the photographer to be easy to detect, and they normally have the bird in its natural habitat. Thus, a bird that can be found in winter feeding on specific frozen berries will typically be photographed in those conditions because they are difficult to locate or identify otherwise. Consequently, a classifier is likely to include snow and berries as important features of its model, be poor at identifying the bird outside that context, and mis-identify other birds in snow as that bird.

1.3.3 Control and Regulation

Controller systems are some of the oldest and simplest automations. They require only a sensor and some type of actuator to influence the output. Even pre-mechanization involves simple controls—you add fuel to a fire when it gets too cool; you rein in a horse when you detect it is going too fast. Mechanical controllers were critical in developing the first clocks (controlling how quickly a gear rotated), but are most common in simple thermostats that detect temperature and close a circuit, often based on the expansion of two different kinds of metal. Modern controllers are much more highly advanced, including antilock brakes, automated driving systems and robotics.

A famous result in cybernetics is the *good regulator* theorem of Conant and Ashby (1970). Their argument, in a paper titled "Every good regulator of a

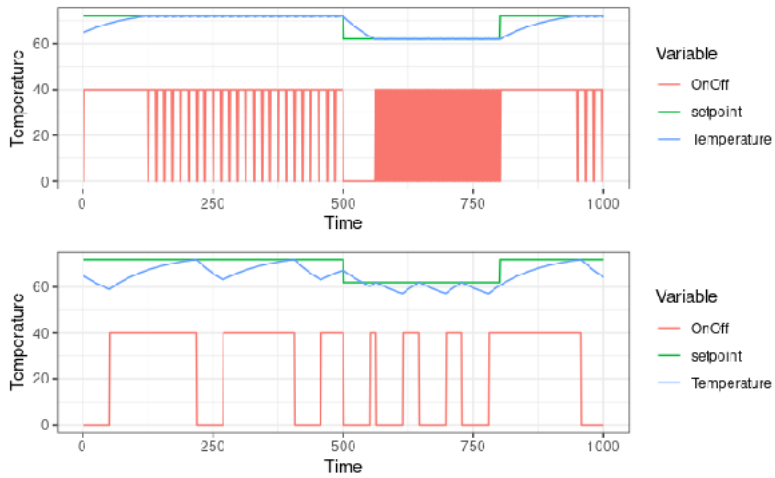


Figure 1.1: Simulated temperature controller in which the desired temperature (green) is set to 70 for 500 minutes, then reduced to 60 for 200 time units, then increased back to 70. The actual simulated temperature is shown in blue, and the heater state (on/off) is shown in red. Top panel shows a simple thermostat in which any deviation from the set point turns on the heat source, which leads to constant cycling of the heater. Bottom panel shows a more complex algorithm in which the heater will not turn on for a fixed period of time after it shuts off, preventing cycling that can be inefficient or harmful to the heating system, at the cost of temperature drops.

system must be a model of that system", states that an effective controller of a system must embed within it the critical elements of the system it controls. This suggests an important role for "digital twins" and environment models in modern control systems, to allow forecasting and predicting the result of changes prior to making them.

A reasonable corollary of this theorem is that if a system can be modeled in simple terms, it permits a simple controller that is likely to be easy to understand. The on-off mechanical thermostat used in heating and cooling systems has a simplistic but inaccurate model of the heater. The thermostat turns the heater on when the temperature it senses is too low, and off when it reaches the desired temperature. To a first approximation, this models the basic heating system (it is either too hot or too cold), and this level of fidelity is satisfactory in many cases. However, these systems do not create a stable temperature, but creates rise-and-fall hysteresis cycles as the temperature moves to the critical trigger point and the heater turns off, then falling to a point low enough to trigger the heater again (see Figure below)

The problem that emerges is that this simple model is not sufficient for many systems. One example is consumer-level espresso machines, which

typically have a thermostat controlling the temperature of the water going through the coffee grounds. As the pump is activated, the hot water in the boiler is replaced by cooler water, and eventually the temperature falls below the optimal level, at which point the thermostat triggers when more heat is needed. But this is too late, because by the time the heater starts again, the water is too cool and the result is a sour cup of coffee. Some machines are sold with or can be modified to use other controllers that model the state of the water temperature more exactly. These can cost hundreds of dollars more, but will maintain the temperature better because they detect and respond to small changes in the temperature.

These kinds of controllers are called PID (Proportional-Integral-Derivative), and are useful devices that can be used to regulate a variety of processes in industrial settings and consumer products. The PID model of the system is more sophisticated than the on-off thermostatic model. As the name implies, these systems are tuned with three separate processes: proportional control regulates how close the environment is to the set condition, integral controls change the intensity of the system's function to reach the set condition and derivative controls adjust for over-accommodation of the integral controls (Johnson, 2005). Thus, PID controllers can be good regulators—they have a more accurate model of the system they are controlling, and can better forecast future states of the system and take earlier action.

Another example is the cruise control on a vehicle. Suppose that the cruise control was a thermostat-style controller. Whenever the vehicle was travelling too slow, it would need to activate the engine full-throttle until it reached the desired speed. Then it would deactivate, coasting until it went below the specified speed. Even with a band of indifference of several MPH or KPH, it would be uncomfortable and dangerous. Instead, it controls the throttle level to maintain a set speed, and if the vehicle speed differs from the set speed, can approach it smoothly.

1.3.4 Rank/Retrieve/Recommender Systems

A number of systems work to identify, rank, and retrieve relevant content. Although the underlying systems can differ greatly, they all have a similar goal, and are critical to sensemaking in information-rich environments.

Recommender systems. Recommender systems attempt to identify recommendations about content based typically on a prior model of the user. This model could be demographic data, past purchases or preferences, or other information collected about the user that can help provide future predictions of behavior (Vultureanu-Albiși and Bădică, 2021). It can be described as a filter application that hides or displays information based on user activity and demographics. An important limitation of this type of system is that it is heavily dependent on user input (Jannach et al., 2010). In theory, a recommender

system can tell you books or film you might find interesting, but it will only work well if you have provided ratings of books and films you have read or watched, or answer surveys about what types of properties of films/books you like..

A common architecture for recommender systems as *collaborative filters*. Here, rather than trying to build a model of the user based on the kinds of films they like, you can identify other people who liked similar content, and make recommendations about what those people also liked. This uses the power of large user databases to make informed decisions, but they can also be very blunt tools—the reason you like a specific film may be because of the director or star; another person may like the genre or writer, and so recommendations are likely to be based on reasons that don't appeal to you.

Information Retrieval and Search. Traditionally, information search and retrieval worked by creating an index of content-to-documents. As a simple example, one might look at all the keywords used for articles in a scientific journal. Over years of publication, there may be hundreds or thousands of keywords, stored efficiently in a single database or document. Then, a search engine merely needs to identify documents that match combined keywords along with rules like OR and AND to identify relevant articles. Of course, keywords are notoriously limited, and so other indices can be made (authors, words in the title, abstracts, dates, etc.), and efficient databases can easily retrieve content.

Initial web search engines worked in similar ways. They would index all web content, and create index tables mapping words to documents. Of course, some words are irrelevant (of, the, and, when, etc.), but by representing each document as the words embedded in it (a bag-of-words model), it could be easily indexed and retrieved. However, early search-engine-optimization recognized this, and many web pages would embed the entire dictionary of words on their page to increase traffic or to sabotage search engines. The important goal turned out not to be identifying pages containing content, but finding the most relevant pages—ranking them by how important they were. One strategy that led to the success of Google (see the PageRank algorithm discussed later) was to essentially use collaborative filtering—content is important if other sites link to it. This creates its own SEO problems, and modern search engines use a complex set of algorithms to provide relevant content, but they all work by attempting to rank and retrieve relevant information.

1.3.5 Cluster, compress, and collate

A variety of 'unsupervised' learning algorithms attempt to build a useful model of a data set based on its own inter-relations. Unsupervised systems take unlabelled cases and apply algorithms to map these cases onto meaningful dimensions or categories. Thus, they are akin to regression and classification models—but without the answer key.

Clustering algorithms. Clustering involves the grouping of similar objects or data using a machine-learning algorithm (Halkidi et al., 2001). Many forms of clustering models exist, each looking at the data in a specific way. Generally, clustering is a form of pattern recognition, in which the machine-learning model looks for similarities between the data points and quickly sorts them into manageable clusters of data for the researcher (Rodriguez et al., 2019). Clustering is an efficient method of evaluating data for trends and outliers, making it very useful in areas such as data science, statistics, engineering, etc. (Ezugwu et al., 2022).

Dimension Reduction and scaling. Other systems involve scaling numerical feature data into fewer dimensions that often characterize important aspects of the data. Some examples of these include eigen decomposition, principal components analysis, singular value decomposition, factor analysis, multidimensional scaling, and other self-organizing maps.

Many intelligent algorithms are centered on vector-based representations of data. That is, cases are represented by a vector of numerical values. Any data-rich domain will fit naturally into this paradigm; for example a network of sensors will produce a series of samples from each sensor that represent the state of the system at a given point in time. Similarly, data about a person in a database may be represented by the information a company has—their age, education level, zip code, etc. Data about a word may be represented by the words it appears near in documents, or other properties. In many cases, the raw vector data is extremely highly dimensional, and better inferences can be made by reducing the dimensions of the data via certain matrix arithmetic operations.

Dimensionality reduction techniques capture essential information in data while getting rid of redundant or less informative features to extract meaningful data, facilitating simplified computational tasks. Two related approaches include Principal Components Analysis (PCA) and Singular Value Decomposition (SVD).

PCA and SVD are used for dimensionality reduction in multivariate data and are a widely-used technique in data analysis and machine learning (Baruah, 2023; Chengwang, 2010). Its objective is to transform high-dimensional data into a lower-dimensional representation, capturing the most important information. The distinction between the two approaches involves primarily whether the input data are a case-by-case matrix (such as a correlation, similarity, confusion, or co-occurrence matrix) which can be reduced via PCA, or a case by context matrix (i.e., a term by document matrix), in which case SVD is used. Depending on the data set and complexity and goals, dimensions may be reduced from thousands or millions (a document x word matrix) to hundreds or even just a few. Sometimes the major dimensions (components) are interpretable, but generally this won't be true unless specific rotations or additional analyses are performed.

1.3.6 Decide/Plan/Act

This describes a very broad set of systems that involve agent-based or 'agentic' systems, decision systems, planning systems and the like. Historically, the General Problem Solver (GPS; Ernst and Newell, 1967) was perhaps one of the first comprehensive general AI systems, developed as a way of moving within a problem space, making progress toward goals by creating subgoals and using operators to move within that space. However, game-based agents are often of this nature. Holland (1999) highlighted Samuel's (1959) checker-playing robot as an emergent system that learns to play checkers based on simple pattern recognition, and more advanced systems that play Chess and Go are more complex systems that take an action based on a well-defined game state, but these often implicitly or explicitly identify series of moves—a plan—to accomplish the goal. More specific planning systems exist to generate routes within a map, and these are now embedded in all our phones and most new vehicles. Similar systems help create low-observable flight plans for stealth aircraft, identify the cheapest way to deliver a product from a warehouse to your home, and (with modern LLMs), carry out tasks such debugging computer code or making automated purchases.

1.3.7 Constraint Satisfaction

When an AI searches through the set of possible solutions to a problem, sometimes the sequence of steps used to solve the problem are important, such as GPS routing. In other cases, only the solution state is important. These types of problems are referred to as Constraint Satisfaction Problems (CSPs). Generally, we think of these problems as having *variables* that take on values from a particular *domain* and are specified with *constraints* that specify allowable solutions (Jones, 2008). Finding the best seating chart for a Thanksgiving dinner to minimize the chance for family drama is a CSP. Each person (the variables) can be placed in a seat at the table (the domain) near some people and far from others (the constraints). If there are enough constraints, most random solutions will violate one or more constraints, and so the goal of a CSP algorithm is to find permissible outcomes that satisfy all constraints.

A brute-force method called *Generate and Test* can be used to solve CSPs. It begins by generating a solution, whether randomly or guided by heuristics, and then testing to see if the constraints are satisfied. More sophisticated algorithms include backtracking and forward checking. Backtracking starts with the most constrained variable, tries adding another variable, and possibly backtracking when a constraint is found to be violated. Forward checking preemptively eliminates invalid partial solutions to reduce the search space (and correspondingly, the time to a solution).

The essence of CSPs is that the problem involves specifying the constraints, and coming up with a solution that obeys the constraints or minimizes the violations.

1.3.8 Generative AI

Generative artificial intelligence is the broader term for machine learning models that receive input (such as a prompt) from a user and create an output based on the training data the model has access to (Brynjolfsson et al., 2023). Although these systems are gaining wide usage with modern systems like ChatGPT, Shannon (1951) was perhaps one of the first to propose them in terms of a way to generate 'fake' language text using what is called a Markov model. His algorithm was simple. Take a book, open it to a page, and pick a word, and write it down. Then find that word again somewhere else in the book, and write down the word that follows it. Continue until you get a word at the end of a sentence. For example, in this chapter, I used this method to generate the following 'sentence':

A VARIETY OF IDENTIFYING AND INTERPRETING QUERIES OR OUTPUT LAYER BY OPTIMIZING SOME PURPOSE.

Notice that this is not grammatically correct or semantically meaningful at the high level, but its themes are recognizable and relevant to the current chapter. We could easily create 'fake French' text that would fool someone who doesn't speak French. If we want a better approximation, we just need to use pairs of words or larger chains (what are called *n*-grams). Now, we would be cutting out small pieces of language from real text, and pasting them together only when they match pairs. Or we can do it at the level of a letter. Here are some example 'sentences' made by reading a dictionary of words, and using just the letter tri-grams to generate text.

THE THEM FROW ITS DIS HATINTE WERCHAT DON ANCE ACRIVE
THE NOTHATORS MONLY SMER ING RIT IMPUT BLE ING THOSED
BUSTRARSO GAR CAUSE SED WAY KED WITS DIS URTUS THE
CALLED SON INALLIED HITY WIDE AVE TOR OFFIS VALL HEACHE
THE ITHEREKS ANDAT THE ITOUBEFFS ANE BITS TAGE USIS MEND
MINGED WIT STRATIS HIC ING CON EDISTAIR THED ANDLYSIS ALP
GUITED SENTIN THE LEGINGIS WHE THE KARD ITTLY JUSTURE
BETED

SIGE CERY LOSED ROMEEDS SED SULD THEY FUNITAT WHE AND

Notice that many words are real English words, and the non-words resemble English—and are likely to fool a non-English reader. They even often appear to include orthographic rules of English. Consider MINGED, ITOUBEFFS, JUSTURE, NOTHATERS.

This is a very simple example of a generative model. If we want to approximate text better and better, we just need to use higher-order *N*-grams.

This is simple—training involves making a matrix the size of your token set (e.g., 27 with letters $\{a-z\}$ + space), and dimensions the same as your desired N , and fill it all with 0s. Then, read your text corpus, and for every N -gram, find that location in the matrix and increment its value. So, for a 3-gram, if you see the word BUTTER, you would increment $matrix[2,21,20]$ (BUT), $matrix[21,20,20]$ (UTT), $matrix[20,20,5]$ (TTE), $matrix[20,5,18]$ (TER), and $matrix[5,18,27]$ (ER_). To generate text, we simply start by finding the distribution of letters that begin words (maybe $\{_ _ _ \}$), which includes 27 values, and sample proportionately. If we get the letter C, we then find the row of the matrix representing our current context ($\{_C _ \}$) and sample again., which might generate an L ($\{_CL\}$), then an A (CLA), then an R (LAR). We can repeat this until some stopping point.

But there are several problems. First, the storage required to store N -grams goes up as a power of the number of steps. For text, 1-grams take 27 numbers, 2-grams take 27^2 , 3-grams take 27^3 , and n -grams take 27^N . This is very costly and inefficient, and quickly scales beyond what is possible on computers. Of course, even if you had all text ever generated, most n -grams would eventually never occur, so your matrix is mostly empty. You might be able to deal with this by clever data structures that only maintain values greater than 1 (a so-called 'sparse' matrix), but even if then, with an N of say 20, you would essentially be encoding most sentence strings uniquely. When you generated text you would really just be extracting text from your data sources directly.

Modern generative algorithms using recurrent neural networks solve the first problem, specifically by using recurrent networks that feed back into themselves so that they learn the sequence of tokens that represent language, and the second problem because their representations are inherently noisy and their learning process has the effect of generalizing. In fact, although 20-deep matrices are probably impossible with our markov approach, modern algorithms now can use at least a million tokens as the context to prompt subsequent tokens. The context includes the recent history, specific instructions, and the 'prompt' given by the user. Conceptually, this is the same as how the CLA prompt generated R. This is essentially how modern LLMs work, although there are many other details that are needed to give them the success they have.

The output of a generative AI model can vary but includes text, images, video, audio, lines of code, etc. (García-Peñalvo and Vázquez-Ingelmo, 2023). Generative AI has become a large topic of discussion in media recently, as generative AI models using Large Language Models (LLMs) such as ChatGPT (text generator), Copilot, Cursor, or Claude Code (computer code generator), and Midjourney (art generator) have risen in popularity (AbuMusab, 2023). Many find the use of generative AI models to be beneficial to users, as it may make work more efficient, but many question if the implementation of these

models will impact the need for human jobs in certain fields, such as writing and media (Zewe, 2023). Both in the workforce and in education, the use of a good prompt can create a high-quality output that resembles human work, leading to many debates of whether this technology should be embedded into everyday work (AbuMusab, 2023).

1.3.9 Summary and Takehome for Human Factors Researchers

We identified 8 general classes of intelligent systems which cover many (but perhaps not all) systems that provide intelligent support for users. From the perspective of a human-centered researcher, basic literacy about the distinctions among these is important, because you may be working with researchers or developers who have solutions from one perspective or another. Understanding these also presents possibilities—if you are involved in designing a system, you may not recognize how a classifier or recommender may improve usability until you understand the roles of these systems and see examples of how they get used. It is also important to recognize that the underlying algorithms are not mutually- exclusive; modern LLM systems can achieve many of the functions listed here if prompted properly; planning systems may also use constraint-satisfaction; the same model that classifies can be re-framed as a generative model. However, it is also important to understand that some of these systems are very capable and efficient, and may provide a better solution than another more complex one. Certainly, one could embed a chatbot in a thermostat, creating an observe-and-decide cycle where it determines when to turn on a furnace, but this would be costly and may be error prone in comparison to mechanical thermostat.

1.4 Architecture: Representations and Algorithms

Many different approaches have been used to implement AI and automated software. One consideration is the underlying way in which knowledge is represented. This ultimately can impact the transparency, efficiency, and traceability of the system, and developers familiar with one kind of representation can often predict weaknesses and problems that will arise from using that architecture. Closely linked, but distinct, are the algorithms (computational procedures) used to either build the system or achieve a solution. These representations and algorithms sometimes are closely linked (back-propagation and activation in neural networks distributed representation), but sometimes interchange (a constraint-satisfaction problem solution could be achieved via several algorithms, and be represented in several distinct ways). Together, the representations and algorithms used for a particular problem can be considered the *architecture* of the intelligent system.

1.4.1 Common representations

Most representations used in intelligent systems come down to a few simple concepts, which map onto common data structures in computer science. These include:

- **Features/vectors.** These can include raw data from sensors, audio, video, etc, which closely map onto input and/or output data.
- **Networks.** A flexible representation that posits units connected by relationships; can be low-level distributed representations in artificial neural networks, high-level constraint networks include roadway systems, etc.
- **Trees.** Constrained network structure allowing only branching; algorithms can efficiently reason, search, and sort. Enables efficient $O(\log(N))$ decision making via classification and regression.
- **Symbolic type/tokens.** Categorical identifiers allowing reasoning about classes of objects, including language production systems, classic expert systems, and the like. Often linked to rule-based systems.
- **Rules.** Often coupled with symbolic systems, but also part of decision trees; may be coupled with functions/transformations to provide stronger inference capabilities.
- **Case/exemplar systems.** From KNN classifiers to Case-based reasoning (CBR) systems, rather than inferring a model summarizing data, use an efficient data store to compare new cases to all existing cases.
- **Probabilistic/Markov/Bayesian or Fuzzy logic.** Often coupled with networks or trees, represent nodes in terms of probability distributions or partial membership in fuzzy sets..

In general, most intelligent systems use one or more of these to encode and represent the problem, data, or model it is using. Many use multiple—a vocabulary may be encoded at one level as a symbolic dictionary, but there may be underlying feature or network representations to permit similarity-based matching via an ontology.

1.4.2 Common algorithms

Just as there are common representations, there are common algorithms that should be understood, at least at a high level. Algorithms serve many different purpose, from learning to inference. A summary of many common algorithms follows, with more detail in subsequent subsections.

- **Search.** Search is the sine qua non intelligent systems. A solution can almost always be framed as a goal state, and the system aims to search among possible solutions for one that is optimal.

- **Search: Dijkstra's Algorithm and A*.** Extensively used within any well-formed problem space (including planar graph networks and road networks); used to find optimal plans, routes, and actions.
- **Search: Forward/backward chaining.** Used to plan a series of operations that produce a result or entail a logical requirement.
- **Search: backtracking, recursive search, breadth/depth-first search.** Within well-formed spaces such as trees, different strategies for efficiently finding the best or acceptable solutions.
- **Inference: Case-based reasoning.** Reasoning from a database of instances that represent past experience, rather than a summarized model representing typical/average situations.
- **Learning: Maximum Likelihood approaches.** Numerical optimization schemes that find the most likely set of parameters in a model to characterize the existing data.
- **Learning/Classification: Bayesian/Monte carlo Markov-chain (MCMC) approaches.** These use Bayes rule on probabilistic representations (including networks and trees) to incorporate evidence in order to identify probabilistic conclusions; results in inference about a distribution of possible parameter combinations that might produce observed data.
- **Inference/classification: Fuzzy logic.** ways of using fuzzy-set theory and inference to characterize uncertain knowledge and determine possibility-based set membership or classification.
- **Learning: Gradient Descent/Hill-climbing.** A general concept that most learning algorithms can be considered to implement; make changes in representation that reduce error or cost based on a local estimate of the slope of the error/cost landscape.
- **Learning: Simulated Annealing.** Addition of noise to a cost function that reduces over time in order to prevent algorithm from settling in local minima.
- **Learning: Reinforcement learning.** General approaches for updating expected reward for actions based on training in a model of the environment. Includes TD-learning, Q-learning, linear operator models.
- **Learning: Back-propagation.** Important algorithm for learning within multi-layer artificial neural networks that assigned error backward through the network.
- **Optimization: Linear and dynamic programming.** Provably-optimal solutions for certain kinds of well-behaved cost optimization problems, can be very efficient and the only feasible solution for many complex applications.
- **Optimization: Genetic algorithms.** Uses analog of evolution to explore complex feature space; iterative generations sample and exchange features which are then evaluated for fitness, to achieve overall improved cases. Useful for complex problem spaces with cost functions that are difficult to characterize.

- **Unsupervised learning: Finite clustering approaches** (KNN, finite mixture modeling) solves for a fixed number of clusters given similarity or feature-based representations.
- **Unsupervised learning: Eigen Decomposition/SVD.** Matrix algebra solutions that project data onto a smaller number of useful dimensions that can reconstruct observed data with differing levels of fidelity. Requires numerical solutions to invert a matrix, which can be efficient even on large data sets.

1.4.3 Specific Architectures in AI and intelligent systems

Together, these representations and algorithms are used across many different intelligent systems. Most complex AI systems likely use hybrid approaches combining many of the different approaches. These lists are shown to provide basic definitions and provide familiarity to a human factors researcher, but we cannot go into detail about all such representations, algorithms, and resulting architectures. However, we will review some of the more common architectures that bear understanding, as they can impact how the system is trained, its competence envelope, misconceptions users may have, and its overall costs/benefits.

1.4.3.1 Search algorithms in AI

Search algorithms have traditionally been a primary algorithm forming the cornerstone of artificial intelligence research. In general, search refers to framing a problem in terms of a state space or problem space, identifying the current state and the goal state, then using systematic algorithms and heuristics to find a path through the space (Bornstein, 2019). In addition, fitness and cost functions for either states or paths through the space are often calculated, so as to identify the optimal way to solve the problem. Several common search approaches are described in Section 4. Many AI problems can be framed as a search problem. Route planning searches through a network for an efficient route; recommender systems search through a set of products to find an option that matches ones preferences; chess AI systems search through possible futures (including likely moves of an opponent) to determine the best next move. In AI solving chess and other similar games, searching is crucial for exploring the vast number of possible moves and counter-moves in a game. Search algorithms may lack transparency, leading to uncertainty and trust issues, because solutions reflect the fitness functions used, the way a problem space is framed, and the local optima the algorithm will often produce.

1.4.3.2 Case-Based Reasoning

Case-based reasoning systems rely on a knowledge base of previous situations and experiences to make choices or decisions based on past experience. This algorithmic design is based on the concepts of scripts, memory organization packets, previous situations, and situation patterns (Schank, 1983; Schank and Abelson, 1977). Cases are contextualized experiences that include the problem, the solution, and the outcome (Watson and Marir, 1994). There are four main steps in case-based reasoning: retrieve, reuse, revise, and retain. The new problem is compared to previous similar cases and a solution is suggested based on the previously used solutions. Unless the problems are extremely similar, the previous case's solution is revised and the new solution and problem is retained for future use. Knowledge-based reasoning often fails due to the knowledge elicitation bottleneck (Hayes-Roth et al., 1983) which case-based reasoning does not experience. Case-based reasoning is particularly useful in applications where no explicit model exists and where new cases can be acquired to improve performance.

1.4.3.3 Artificial Neural Networks for inference/classification

Artificial neural networks are computational models that mirror behavior of the human brain, usually with some levels of abstraction. The human brain is made up of neurons that send signals to each other through the use of synapses and neurochemicals. An artificial neural network functions through the connection of different nodes where each node can accomplish a simple integration of information that is transformed into a simple output (Arbib, 2002). The guiding principle of artificial neural networks is that simple units and learning rules together can combine to form sophisticated and complex behavior. These networks are generally organized into multiple layers, and minimally include an input layer (taking data provided from the outside world); an output layer (the result of the processing) and normally one or more hidden layers (existing between the input and output) (Balcázar et al., 1997; Setiono and Leow, 2000). The connection between the nodes is known as the weight, with higher weights meaning that two nodes are more strongly connected. Individual units typically involve non-linear mappings (Vargas et al., 2017), which enable complex information integration and rule-like behaviors to emerge.

Although traditional neural networks were limited to just a few layers, modern approaches, called deep neural nets or deep learning (LeCun et al., 2015), have emerged that use many layers (e.g., 15-20), and different layers incorporate distinct processing architectures (e.g., recurrent, convolutional, etc.) The order of the layers is often defined by the programmer based on heuristics and trial-and-error. Neural networks are useful for their ability to self-learn and to find optimal solutions quickly (Wu and Feng, 2018).

1.4.3.4 Backpropagation algorithm for learning in Artificial Neural Networks

Early versions of single and multi-layer neural networks were proposed (McCulloch and Pitts, 1943; Rosenblatt, 1958) as demonstrations of the power of simple classifiers. Their capability was recognized to be limited in terms of their ability to learn complex representations such as XOR feature interactions (Minsky and Papert, 1969) unless multi-layer networks were used, but rules to properly learn weights in multi-layer networks were considered ineffective. Backpropagation was the first practical solution to this problem. Back-propagation is a scheme which starts at the final layer of a network, in which a prediction is made and compared to the ground truth. The difference between these is the *prediction error*, and the problem is determining which of a large number of nodes and weights are to blame for that error. In a simple perceptron node, this operates like a regression equation, and standard numerical methods can be used to choose the optimal weights. But when multiple input nodes across multiple layers make up the network, the optimization is difficult. Back-propagation works by conceiving of the error as a 'gradient descent' problem. It iteratively determines the effect each node has on the accuracy of the final layer prediction, and update weights in reverse across layers in order to reduce the prediction error. For example, if one node is off in the final layer by $+0.5$, it might be adjusted by $-0.5 * 0.01$ (0.01 is the learning rate) to move it in the direction of the correct answer. But then, since that end node is in error, we work backward and assume the previous layer's nodes were also in error in proportion to the final layer error. Based on this error, the weights can be adjusted backward throughout the network, and with enough computational power, training examples, and a mixture of layer structures, it produces complex classification behavior. Although initial approaches to backpropagation appeared in the 1970s, (Rumelhart et al., 1986) is often credited with popularizing the approach, which was further refined by LeCun (LeCun, 1987; LeCun et al., 1989).

Although these approaches fell out of favor by the late 1990s as computational intelligence models, graphics chips that developed for mostly for gaming in the early 2000s worked by optimizing massive parallel computations on data. As algorithms and libraries were developed to optimize neural networks to perform backpropagation and computation on these computational architectures, and data sets emerged supported by internet users (including Amazon Mechanical Turk) it permitted exploring 'deep' networks with 10-20 layers of different computational and connectivity patterns, learning from large labeled data sets, and such models re-emerged by 2020 as the dominant AI architecture for image, video, and language processing. Hinton's 2024 Nobel prize was in part awarded for developing back-propagation (Rumelhart et al., 1986).

1.4.3.5 Symbolic AI Systems with forward and backward chaining

Symbols refer to tokens or identifiers that represent some concept directly. Human written language is a symbol system as groups of letters represent words which identify linguistic concepts. Early implementations of AI relied on symbolic representations in computers, typically using the LISP programming language. The benefit of this approach is that it was capable of producing complex behavior while also being relatively computationally efficient. Although symbol-based representations can be used in many ways, one typical way they are used is to represent procedures via if-then rules. For example, the knowledge embedded within a system might be represented by a series of symbolic statements:

(CAT ISA MAMMAL)

(DOG ISA MAMMAL)

(CANARY ISA BIRD)

The system may contain rules that match symbolic statements of knowledge to produce some outcome:

IF:

((?A ISA ?X) AND

(?B ISA ?X))

THEN:

(SAY ?AS AND ?BS ARE BOTH ?XS)

Then, intelligent behavior is a process of pattern matching. Here, symbolic patterns match rules that lead to actions, such as the system saying:

DOGS AND CATS ARE BOTH MAMMALS

By encoding knowledge, rules, and enabling other temporary symbols such as goals, complex behavior can be implemented. This type of inference is called forward-chaining, and matches logical inference rules like modus ponens that provide deductive proof for syllogistic logic.

Symbolic AI systems were dominant for many years but increasingly replaced by other subsymbolic representations, as compute power and big data availability increased sufficiently to support alternative approaches. However, they are still used as basic infrastructure in many applications such as game-based AI systems and case-based reasoning systems. They are

also increasingly likely to be integrated with other non-symbolic representations (vector space, neural network, and probabilistic representations) to take advantage of unique advantages of different kinds of systems.

In comparison, some symbolic systems use backward-chaining, which is akin to abductive reasoning. Here, we apply the logic of *modus ponens* backward to identify necessary facts and rules that lead to a goal state. For example, if Diana were on a dating app, and wanted a male partner that was at least 6 feet tall with an income of at least \$100,000, we might have a rule representing this goal:

Rule DATEABLE

IF:

(?A INCOME > 100000) AND (?A HEIGHT >72)

THEN:

(SAY ?A IS A DATEABLE)

A backward-chaining system would work backward to find cases that match logically. Of course, this is simple if we have a database of facts about all the available men, but we may not—we might not have income directly, but have a job title and information about incomes for different job titles

(NAVAL_CAPTAIN INCOME_IS 150000)

(NAVAL_SURGEON INCOME_IS 90000)

We may also have rules to help identify income of individuals

Rule INCOME_PROFESSION

IF:

(?PROFESSION INCOME_IS ?X)

(?NAME PROFESSION_IS ?PROFESSION)

THEN:

(SAY ?NAME INCOME_IS ?X)

Now, with information about individuals, we can backtrack through rules to

```
(JACK IS_A NAVAL_CAPTAIN)
```

```
(STEPHEN IS_A NAVAL_SURGEON)
```

```
(JACK HEIGHT>72)
```

```
(STEPHEN HEIGHT > 72)
```

A backtracking system would iteratively search backward through the rules to identify cases that satisfy the goal, first identifying that in order to satisfy the DATEABLE rule, you need to examine the INCOME\{}_PROFESSION rule. When that rule succeeds filling in the income of an individual through inference, it then can select JACK as the only bachelor who satisfies her constraints.

1.4.3.6 Bayesian inference and networks

Probability is a field of study that allows for the handling of uncertainty. Using probability theory, decisions can be made with only partial evidence available, which is critical for rational agents who need to make the best decision possible in their current circumstances (Russell and Norvig, 2021). One application of probability theory is the ability to discuss potential relationships between events, such as “what is the probability I have a cavity given that I have a toothache?” There is a potential relationship between having a toothache and having a cavity, and probability can be used to determine the likelihood of this relationship.

However, not all events have a relationship. For example, one could ask “what is the probability I have a cavity given that it is raining outside?” There is little to no chance of a relationship between having a dental issue and the weather in a specific location. These two events are most likely unrelated, and thus we can say they are independent (Russell and Norvig, 2021).

For events that are related, probabilities can be calculated using the information already known about the event. For example, given that the individual already has a toothache, what is the probability that the toothache is caused by a cavity compared to other potential causes? This probability would be conditional, as it is narrowed in scope to information already known and unrelated pieces of information can be discarded (Russell and Norvig, 2021).

Bayesian networks data structures that can concisely represent these conditional relationships between events, and thus a hypothetical process by which observed data are generated from a state of the world. They typically take the form of a directed graph with no cycles. Nodes in the graph will have directed paths, with the node being pointed to being a child of the node the

path originates from. This directed path implies conditional probability, so the child node is conditioned on the parent and two children from the same parent node are conditionally independent. By having this representation of conditional probability, conditional probability tables can easily be generated based on the derived relationships, and interdependencies, from the network (Russell and Norvig, 2021). This network represents a probabilistic *generative model* of how data are created. For example, one might have data from a set of sensors intended to detect trespassers; and this might include sound, heat sensors, motion detectors, and the like. A generative model in the form of a Bayes network would represent the probabilities that each kind of sensor state might be generated by a number of different situations: the wind, a human, a mouse, a deer, etc. Finally, Bayes rule and Bayesian inference is used to compute the posterior probability of each situation (wind, human, mouse deer) given the observed data, combining both the model of how the evidence was generated optimally with information about the base rate of each case (e.g., wind occurs often, a human may be rare).

1.4.3.7 Linear programming (simplex method) and nonlinear programming

Linear programming refers to a mathematical optimization technique to help to find the optimal solution to a cost function under a set of constraints described as linear equations. In this way, linear programming helps to find the optimal solution from a set of parameters that have a linear relationship. The simplex method (Dantzig, 1982) is a widely-used algorithm for solving a linear program. It narrows down potential solutions to a simplex: the vertices of a high-dimensional convex polygon, and provides an efficient means of searching through potential optima and eliminating possible configurations that are not optimal. Linear programming is limited to linear constraints, and so other related mathematical programming techniques such as nonlinear programming are used to find the optimal solution from a set of parameters that have a nonlinear relationship (Lithmee, 2019).

1.4.3.8 Q-learning and Reinforcement Learning

Reinforcement learning (RL) represents a cross-disciplinary field at the intersection of machine learning and optimal control, addressing how intelligent agents should make decisions in dynamic environments to maximize cumulative rewards (Kaelbling et al., 1996; Sutton and Barto, 2018). Among RL algorithms, Q-learning stands out as a model-free approach for learning the value of actions in specific states, devoid of an explicit model of the environment ("model-free"). Its adaptability to problems with stochastic transitions and rewards makes it a powerful tool (Li, 2023). The versatility of Q-learning

and RL extends across various disciplines, including game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, and statistics.

RL constitutes one of the three fundamental machine learning paradigms, alongside supervised and unsupervised learning. Unlike its counterparts, RL operates in a dynamic environment, learning from collected experiences instead of relying on a static dataset. During training, data points or experiences are accumulated through trial-and-error interactions between the environment and an agent. This distinctive feature is crucial, eliminating the need for data collection, preprocessing, and labeling before training—requirements inherent in supervised and unsupervised learning. In practical terms, this signifies that, with the right incentives, an RL model can autonomously initiate the learning of a behavior without human supervision.

While Reinforcement Learning has emerged as a pivotal aspect in the field of Machine Learning and has garnered significant attention from researchers, it encounters specific challenges outlined below: a.) Extensive Datasets: Due to the complexity of Reinforcement Learning Models, they necessitate vast datasets to enhance decision-making capabilities. b.) Dependency on Environment: Reinforcement Learning Models learn through interactions with the environment, so the agent adapts based on the current state of the environment. In scenarios where the environment is dynamic, the training of the agent becomes challenging. c.) Reward Structure Design: In real-world applications of Reinforcement Learning, there is a need to carefully analyze the problem statement and formulate an appropriate structure for when the model should be rewarded or penalized.

1.4.3.9 Dijkstra's Algorithm and A*

When a problem can be represented by a graph (for example, a set of towns connected by roads), one of the fundamental problems to solve is the minimum cost route distance between any two nodes in the graph. Among other solutions, Edsger Dijkstra's eponymous algorithm (Dijkstra, 1959) is one of the most general, and finds the solution in an iterative fashion. Starting from an initial location, set the initial distance to all nodes as "infinite." This is not to say that the cost to visit that node is truly infinite, only that the final cost is yet unknown. From the initial location, we examine the cost, $g(n)$, to go to every other directly connected location. The minimum value is selected from this set of costs, and that neighboring location is marked as "visited." The same process is repeated from the newly visited location until either a specific destination location has been visited or the shortest distance between the initial location and all possible destinations has been calculated. This is considered an uninformed search, as every decision is made based on actual

costs without an attempt to estimate how the current decision will affect the global view of the solution.

To use this as a simple GPS mapping algorithm, costs can be based on static factors like the distance between two cities, dynamic factors like real-time traffic information, or custom factors like how ecologically friendly the route is. A route utilizing surface roads may be shorter than the interstate highway, but a driver may choose the longer high-speed roads for a shorter trip. Conversely, a pedestrian may want the shortest distance and avoid heavy vehicular traffic.

Traditionally, Dijkstra's algorithm finds the shortest paths from the initial location to all possible destinations. The A^* (pronounced "A star") algorithm (Hart et al., 1968; Jones, 2008) is an extension of the best-first search algorithm exemplified by Dijkstra. The two main differences are that A^* only calculates the shortest path between a specific pair of locations and the cost between locations is expanded to include a heuristic cost. In the case of mapping software or GPS devices, the value of the former difference should be apparent. We only want to find a route from our current location to our destination, not a path to every possible destination. The heuristic cost, $h(n)$, allows the algorithm to estimate how a particular decision will affect the overall progress toward the goal. Consider a trip that is generally northward. The nearest neighbor may have the lowest $g(n)$ cost, but the heuristic cost $h(n)$ will be higher if it takes us south (and ultimately farther from our intended destination). Consequently, the algorithm will be discouraged from further exploration of that path.

While both A^* and Dijkstra's algorithm make intuitive sense for pathfinding applications, any problem that can be represented as a graph can be solved optimally and correctly through either algorithm. For example, all of the possible states of a Rubik's cube or the Tower of Hanoi problem can be represented as a network, and legal moves represented as connections in that network. A^* or Dijkstra's algorithm can be used to find the smallest sequence of moves to solve the problem.

1.4.3.10 Evolutionary and genetic algorithms

Evolutionary algorithms are a set of search algorithms that takes inspiration from, as implied by the name, biology and natural selection. In these algorithms, an initial population of states is generated and a fitness value is generated for each state in the population. The most fit states are then selected and recombined to produce the next generation of states. The process then repeats with this next generation of states (Russell and Norvig, 2021).

One power of evolutionary algorithms lies in how they perform the recombination process. Genetic algorithms (Holland, 1992) are one type of evolutionary algorithm and take their inspiration from human genetics. As such, the states chosen for recombination undergo a crossover process, where two states have portions of their data combined to make a new state. For example,

if the 8-queens problem were being searched and the state consisted of an array of length *columns* where each item in the array was the *row* number where the queen is positioned in that column, then the first two columns might come from the first parent and the remaining six columns would come from the second parent to make a new child state based on these two parents (Russell and Norvig, 2021).

1.4.4 Summary

Together, a system's architecture involves both representation and algorithms. These are often linked closely together, but sometimes there are multiple alternatives. For example, a simple neural network might typically be trained maximum-likelihood estimation or least-squares fitting instead of backpropagation, but the resulting network might be identical. Moreover, the forward processing is distinct from the learning. When these systems are exposed to users, many of the details of the algorithms are unimportant, but there may be consequences that rear their head. For example, a rule-based system is likely to produce a deterministic result—if it gives different results with the same input, it is broken and a signal to the user that something cannot be trusted. On the other hand, a neural network is often probabilistic and for some situations will give slightly different results every time it is used. In this case, a variety of responses is a signal it is working, and if it were to give the same answer every time, that would be a signal to distrust the result.

There are many other reasons to understand aspects of the underlying algorithm. Some approaches might be infeasible because training data would be impossible to produce; others may be infeasible because the storage, computational, communication, or time requirements do not fit the system or the use case. Understanding the basic approaches can provide alternatives that may satisfy more practical needs of system development.

1.5 Looking under the hood at applications of intelligent systems

Now that we have understood the purpose, the representations, and the algorithms of intelligent machines, we can examine various application domains and identify how different applications may be constructed.

1.5.1 Path-planning for mobile robots

At present, there is an increasing application of mobile robots in a wide range of industries. For example, robotic arms have been adopted to assemble steel structures for off-site manufacturing in the construction industry (Liang et al., 2017). To control the mobile robots and run the systems smoothly, it is important to plan accurate paths through gathering and sensing information

about the outside world. Inspired by the behaviors that ants discover food and birds migrate among continents, scientists have developed different algorithms to find the optimal path for robotic movement (Huang, 2023).

Previous path-planning algorithms can be categorized into four types (Zhao et al., 2020), which are respectively fuzzy logic, SLAM, reinforcement learning and intelligent optimization. The idea of fuzzy logic algorithms is based on building a database of behavioral rules that can be applied to a variety of situations. With the external feedback data from sensors taken into account, SLAM can deduce the motion model and observation model, and the robot can estimate the real-time motion state (Huang, 2023). Reinforcement learning enables robots to choose the step with maximum reward at each node in a map, so the optimal path can be found eventually. Through many iterations, intelligent optimization algorithms, such as the ant colony algorithm, endow robots with excellent optimization capabilities.

With the increasing research and continued focus on path-planning for mobile robotics, A*, Dijkstra, rapidly-exploring random tree (RRT) and other algorithms have appeared in recent years (Huang, 2023). However, the common problem of these algorithms is how to get a shorter and smoother path and accelerate the convergence speed. In addition to improving the algorithm itself, combining different algorithms can lead to better path solutions, which is considered an effective solution in the future.

1.5.2 Gesture recognition

Gesture recognition is the process of identifying and interpreting meaningful expressions of motion exhibited by a human, encompassing movements of the hands, arms, face, head, and/or body (Mitra and Acharya, 2007). This interdisciplinary field combines elements of computer vision, machine learning, and signal processing to enable machines to comprehend and respond to human gestures, fostering natural and intuitive human-computer interaction. Therefore, gesture recognition has a large range of applications in various industries, such as virtual worlds, robotics, intelligent surveillance, and sign language translation (Chang et al., 2023).

Liu and Wang (2018) introduced a comprehensive model tailored for gesture recognition in human-robot collaboration. This model includes five crucial steps: sensor data collection, gesture identification, gesture tracking, gesture classification and gesture mapping. To start a gesture recognition process, raw gesture data needs to be collected by image based (e.g., marker, single camera, stereo camera, and depth sensor) and non-image based (e.g., glove, band, non-wearable) approaches (Liu and Wang, 2018). Second, gesture recognition refers to locating gestures from static raw data captured by sensors. Next, the located gesture is then tracked during the gesture movement. In the fourth step, tracked gesture movement is classified according to pre-

defined gesture types based on classification algorithms (e.g., Hidden Markov Model, Support Vector Machine, Deep Learning). Finally, gesture mapping translates recognition results into machine or human understandable feedback.

Gesture recognition technology faces challenges that researchers are actively addressing. One obstacle involves the accurate detection and recognition of human actions, hindered by factors like the variability in human body parts and the surrounding environmental conditions. Another constraint arises from the requirement for close operation and critical point sensitivity in gesture control systems utilizing sensor boards. Moreover, vision-centered human activity analysis through computer vision encounters limitations in detecting actions behind walls or in dimly lit places. Additionally, the intricate degrees of freedom and variability in hand gestures present a recognition challenge for computer vision-based systems. These limitations underscore the imperative for ongoing research and development in the field of gesture recognition technology.

1.5.3 LSA to model linguistic data

One application of SVD is Latent Semantic Analysis (LSA), a natural language processing technique that involves generating a set of concepts associated with documents and terms to analyze the relationships between a set of documents and the terms they contain. The purpose is to reveal the latent semantic structure within a collection of documents, often accomplished by adopting Singular Value Decomposition (SVD) to reduce dimension (Landauer and Dumais, 1997).

LSA uses SVD to decompose a word (term) by document matrix to extract a term by dimension matrix, generally using a few hundred latent dimensions to describe all words. PCA is used to decompose a term by term or variable by variable matrix into a similar variable by latent dimension matrix. The resulting simplified representation has been demonstrated to capture many aspects of semantic information: words that are considered similar by human judges are generally nearby in the resulting high-dimensional space. Because of this, content and sentiment analysis can be performed, generally by applying regression or classification algorithms against a spatial representation for a word, sentence, or document.

1.5.4 Information retrieval and the PageRank algorithm

Google PageRank is a patented algorithm used by Google Search to rank web pages in its search results. It was developed by Google's Larry Page and Sergey Brin, and named after Page (Wikipedia contributors, 2024). The basic idea behind PageRank is that the importance of a webpage is determined by the number and quality of links pointing to it.. This concept extends beyond search engines, being employed in academic citation systems and some social

networks. While PageRank provides valuable insights, understanding its limitations is essential. Vulnerabilities to manipulation, a bias towards established sites, and static analysis may impact the accuracy of rankings. Acknowledging these limitations is crucial for users to interpret search results critically and for developers designing systems relying on PageRank principles. Awareness of these challenges helps maintain trust by recognizing that rankings may not always reflect the most relevant or dynamic information.

1.5.5 Predictive text entry

Personal computing devices of all sorts now offer predictive text entry. As the user types, suggestions to complete a partially-typed word or phrase are presented and potentially accepted. By looking at a corpus of any language, we can calculate the frequency of individual letters, digrams (two-letter combinations), trigrams (three-letter combinations), or generic N -grams. Speakers of any language have a reasonably strong intuitive understanding of these probabilities (Shannon, 1951). Whether the approach is to maximize the likelihood or to minimize the entropy (the level of surprise or disorganization), computing systems can be programmed with similar predictive capabilities. Indeed, the original iPhone autocorrect was manually programmed with N -grams to overcome the limitations of typing on the small screen (Kocienda, 2018). A larger corpus of words should lead to more accurate predictions. On-going updates to the corpus specific to a user will lead to more accurate predictions. However, the nuances of human communication, the context-specific nature of certain forms of communication (e.g., a work setting versus a casual conversation), and the ongoing changes to a living language contribute to inaccurate predictions.

1.5.6 Grammarly tone/sentiment analysis

Sentiment analysis is a form of classification problem that seeks to classify a piece of text as carrying an overall tone or sentiment (Medhat et al., 2014). There are many ways that artificial intelligence can accomplish this task, from more statistical methods that rely on tools and methods such as pre-classified lexicons, n -gram models, and naive Bayes classifiers to machine learning approached with recurrent neural networks and large language models (Medhat et al., 2014; Russell and Norvig, 2021).

Grammarly's tone indicator is one example of sentiment analysis application. In a post on the company's blog (Calonia, 2020), tone is defined as "...the author's attitude about a subject or topic to their reader." The author then goes on to provide examples of different tones, noting that choosing the correct tone can help ensure that one is making grammar, stylistic, and vocabulary choices that align with what one is trying to communicate.

To assist with this, Grammarly released a tone detector in their system in 2019. Introducing the tone detector in another blog post (Grammarly, 2019), Grammarly described their detector as a mixture of both a rules-based and machine-learning approach that analyzes vocabulary selection, phrasing, punctuation, and style choices in things like capitalization. They end the post with the following rationale for creating their tone detector:

It's true that you can't control how someone reacts to your writing, but by making thoughtful decisions about the way you deliver your message, you can increase the likelihood that your recipient will focus not on unintended effects of your tone but on the information you're sharing.

1.5.7 SIRI and other assistants

Siri and other voice assistants all exist to aid the user in some manner. In general, these devices provide information, play specific media requests, or complete certain actions or tasks (Hoy, 2018). These devices listen for specific keywords, often their own name, to understand that they are about to listen for a prompt or request. In the case of Siri, a deep neural network is used to convert acoustic values of a person's voice into a probability distribution of potential requests. Once this is done, a confidence score is created. If the confidence score meets a certain threshold, then the device responds to you. Since voice assistants are connected to the internet, they can respond to a larger number of prompts. In addition, as the amount of text on the internet has increased, natural language processing has helped to create better responses. Natural language processing also allows for a number of different prompts to get the same results, as opposed to previous technologies where a specific phrase had to be used.

Voice assistants have some limitations of security and privacy. In general, anyone who can access a voice assistant can request personal information about a user from it. Researchers have shown that voice assistants will respond to commands given at ultrasonic frequencies that would be inaudible to humans (Zhang et al., 2017). By nature of the device, voice assistants are listening at all times. Tech companies like Google and Apple claim they are not collecting data at all times, but Google's voice assistant has been known to record at all times and save those recordings (Stegner, 2023).

1.5.8 Smart home automation

A smart home automation integrates household appliances and systems through a network to achieve monitoring and management functions. The central management system enhances convenience and efficiency of the machines and systems. When combined with artificial intelligence (AI), the smart home automation collects user data, automates settings for home appliances

such as temperature and lighting systems, offers personalized setup, and takes proactive actions in response to detected abnormal actions (Dasgupta et al., 2022).

1.5.9 Facial Recognition

Facial recognition is a form of technology capable of evaluating human faces and informing the user of who the individual is (Lewis and Crumpler, 2021). There are many variations of facial recognition, from smaller uses such as Face ID on a cellphone to evaluating an image or video clip from a crime scene of an unknown individual against a large database to determine if an identification can be made (Adjabi et al., 2020). Facial recognition works by evaluating small features of the face against another image to determine if similarities exist (Qinjun et al., 2023). This technology is rapidly changing and evolving, however, there are many issues and limitations with the use of facial recognition technology. Issues such as accuracy, discrimination, legality, privacy, and data security have been heavily debated in recent years as facial recognition has become more prominent in various domains such as healthcare, retail, government, and law enforcement (Qinjun et al., 2023). While this technology may have challenges, some see its usage as an aid in human decision-making abilities to help error reduction in identification (Lewis and Crumpler, 2021).

1.5.10 Rule-based and expert systems

Rule-based and expert systems are a type of AI that uses rules and logic to make decisions, mimicking human experts in specific domains like medical diagnosis or stock trading. The underlying algorithms involve ways of combining data about a current situation with a complex set of context-dependent rules to determine an outcome or recommendation. These emerged in the 1970s and 1980s as one of the first bubbles in AI systems. Developers of such systems needed a detailed understanding of the domain and the rules that govern a rule-based system, as end users typically don't need to comprehend its inner workings. This often meant engaging with experts within the domain to elicit their knowledge, identify the rules that described their decisions, and carefully delineate many edge cases, and eliciting and creating such rules came to be recognized as the greatest challenge for creating and maintaining such systems.

Some examples of rule-based systems include MYCIN (Shortliffe, 1976) (a medical diagnosis system), XCON (a configuration management system), and CLIPS (a tool for building expert systems) (see Berka, 2015). Rule-based systems can be limited in their ability to handle unexpected situations that don't match the predefined rules. Errors may occur if the rules are incomplete

or incorrectly specified. Understanding rule-based systems' limitations can enhance users' trust, allowing them to comprehend their capabilities and limitations rather than blindly trusting their accuracy or completeness (Wanner et al., 2022).

1.5.11 Adaptive cruise control

Cruise control is a commonly used car feature that can maintain a set speed. Traditionally, cruise controllers have used simple PID controllers, adjusting the throttle to smoothly maintain a set speed. Adaptive cruise control adds additional intelligence and incorporates external factors into the maintenance of the vehicle's speed, including sensors in the front of the vehicle to measure distances to objects in front of the vehicle and use of braking for added safety. A set speed can be superseded by the car's speed in front of the operating vehicle and includes a set distance between the two cars to maintain comfortable spacing. If the front vehicle moves faster than the set speed, then the vehicle will only move as fast as the set speed. Adaptive cruise control can be considered another step toward fully autonomous vehicles as the car can regulate to external factors beyond a desired speed limit set by the operator.

For the Human Factors Practitioner

- Do not assume that AI means a large language model/chatbot.
- If you are working with a AI/ML developer, ask them to explain the system and the trade-offs for alternatives.
- Identify the functional class of the system. A system marketed as using AI could have it in many different places, and understanding what it is doing is critical for knowing how it will impact users.
- Neither you nor users/operators may need to know the mechanics of the algorithm if you can understand its strengths, weaknesses, and competence boundaries.

1.6 Summary and Discussion

The goal of this chapter is to demystify many of the terms, algorithms, concepts, and constructs used in AI, Machine Learning, and other related domains. Basic understanding of these concepts is important for a professional or researcher in human factors or cognitive systems engineering who wants to interact with engineers working with such systems, and who want to improve human use of these systems. The particulars of these systems need not always be a black box, a understanding the basic strengths, weaknesses, and constraints

of these systems can be critical for helping users develop justified trust and mistrust of the system, by understanding the competence envelope of the system, its boundary conditions for proper use, and situations in which it is not appropriate.

Acknowledgments

This chapter was developed as part of a class project for the MTU ACSHF graduate course HF 5430: Human-AI Interaction during Spring 2024.

Author Contributions

Contributors to the chapter are listed in alphabetical order.

CC: 2.6, 5.2, 6.4.; BF: 4.2, 5.3, 6.5; EM: 2.5, 3.4, 5.6, 6.3; KU: 3.1, 5.2, 6.6; ;LS: 2.1, 2.4, 5.1, 5.5; JW: 3.2, 3.3, 4.3, 6.7 ; SW: 5.3, 5.7, 6.8; YJW: 4.4, 6.1, 6.2; BW: 5.1, 5.4, 6.9. STM: Conceptualization, Review, Editing, and additional writing of all sections.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Cischke, C., Frisch, B., Matas, E., Mueller, S. T., Sprague, L., Ulinski, K., Walker, S., Wang, J., Wang, Y., & Woolman, B. (2026). Demystifying automation, artificial intelligence, and intelligent software tools and applications. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Human-AI Interaction Research: An Overview

Katie Ulinski, Brandon Woolman, and Shane T. Mueller
Michigan Technological University

Abstract

Perhaps beginning when Wiener (1948) coined the term *cybernetics*, researchers have studied the interface between humans and intelligent machines under a number of disciplines and names. These include human-AI interaction, human-centered AI, and other research terms within distinct disciplines and communities. This chapter begins with a discussion of these different schools of thought, describing the main tenets and approaches used in each. Then, we discuss how although the goals of these approaches can be distinct, many of the elements of these systems important for users to understand come down to a handful of components: data, representation, algorithm, and output.

2.1 Introduction

All technology ultimately involves humans: it is made by humans, for humans, with the goal of helping humans, and sometimes with the side effect of harming humans. Despite this, technology is often constrained by engineering compromises, practical trade-offs, and a lack of understanding of the user. This can make it difficult to use, hostile to subgroups of users, hard to learn, unreliable, and unable to be trusted. Various fields of research and design, including human factors, ergonomics, human-computer interaction, cognitive systems engineering, human-systems integration, and UI/UX design have emerged to improve design and evaluation methods in order to ensure technological

systems that are directly used by humans are usable and ultimately useful to them. To paraphrase Walt Whitman, the field is large, and *contains multitudes*.

AI and automation are often thought to have the goal of replacing the work of humans, perhaps placing it outside the realm of human-machine interaction. Although this is certainly true in some cases (traffic cops wear white gloves to help drivers see their hand signals; traffic lights avoid the need for improving the human-glove interface), other problems are amplified (e.g., when autopilot is engaged, a pilot may become more likely to lose situation awareness) and new challenges are created (e.g., human-automation hand-off protocols may be required that were not previously needed). As discussed in Chapter 1, (Bainbridge, 1983) identified “ironies” of automation, suggesting that automation may make some tasks easier, but these often lead to changes in how the human works—monitoring, supervising, and taking over in difficult situations. So the approach in many of these disciplines is to understand the interactive system and how work changes to support it.

A number of related fields of study focused specifically on human-AI and automation interaction have consequently emerged. Some are used essentially interchangeably, but others refer to specific approaches or techniques. In this chapter, we will discuss some of the prominent schools of thought and research and design communities. To close the chapter, we will describe a basic taxonomy of the kinds of functions intelligent software (AI and automation) typically has that separates itself from general technology, and may be the avenue for interventions that ultimately improve human-AI interactions.

2.2 Schools of thought for designing AI and automated systems for human users

Research and practice in human-AI and human-automation interaction takes place in a number of communities of practice (Xu et al., 2021). As discussed in the first chapter, cybernetics (Wiener, 1948) was and remains an appropriate blanket title, but many other relevant terms have been coined. Among many, these include cognitive systems engineering (Woods and Roth, 1988), joint cognitive systems (Hollnagel and Woods, 2005) human-centered AI (Shneiderman, 2021), human-AI interaction (Amershi et al., 2019), human-AI collaboration (Wang et al., 2020), human-automation interaction (Sheridan and Parasuraman, 2005), human-machine teaming (Lyons et al., 2019), human-AI teaming (Berretta et al., 2023), mixed-initiative interaction (Allen et al., 1999), AI-assisted decision making (Zhang et al., 2020), human-in-the-loop AI (HITAI; Zanzotto, 2019), and there are certainly others. Although there is strong overlap among many of these areas, this section will discuss the basic approaches for three of them: human-AI interaction, human-centered AI, and cognitive systems engineering. As shown in Figure 2.1, the distinction between these

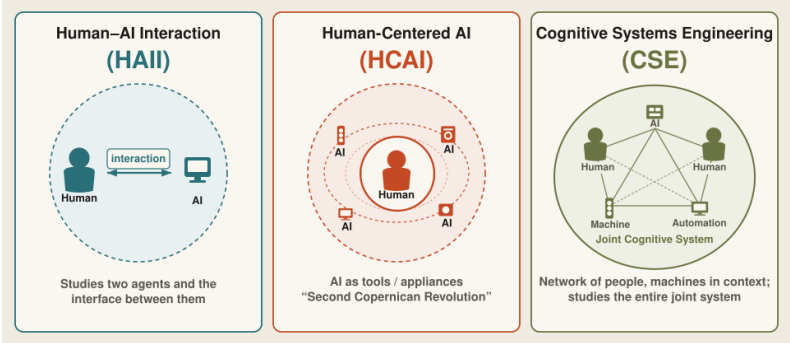


Figure 2.1: Human-AI interaction (left) is a general term that implies a focus on the interaction between Human and AI systems. Human-centered AI (center) view AI a tools that should support human goals, rather than work how humans do. Cognitive Systems Engineering (right) views both humans and machines as part of the system, and studies the interdependencies.

involves what the target is emphasized. In HAI is the broadest term, and is the study of two agents (human and AI) and the interaction between them. HCAI attempts to “center” the human, and that AI systems should be designed with human goals and purposes in mind. CSE views humans and their tools (including AI and other machines) as interdependent entities in a joint cognitive system, and examines how the system is perterbed and changed by adding automation or intelligent machines.

2.2.1 Human-AI Interaction (HAI)

The field of human-AI interaction (HAI or HAI) is a generic term arising from an analogous domain of human-computer interaction, which is concerned with user interfaces and how users interact with computers, and more generally technology as a whole (Carroll, 1997). HAI might properly be considered a subfield of HCI, although Xu et al. (2021) proposed that HAI is an emerging interdisciplinary field. In HAI, AI is the focus of what humans are interacting with, assuming that AI systems have some interface that users are aware of and utilize (Wienrich and Latoschik, 2021). This includes both the hardware and software levels of an AI system and assumes that both have some effect on the user’s interaction. Most often, when discussing HAI, the AI portion refers to a system that has intelligent components or capabilities, rather than an individual standalone AI system.

HAI/HAI is comprised of a multitude of numerous disciplines. For example, Xu (2019) identified that it involved three main components: First, AI should be ethically designed; it should avoid discrimination and should not replace

Human-AI Interaction

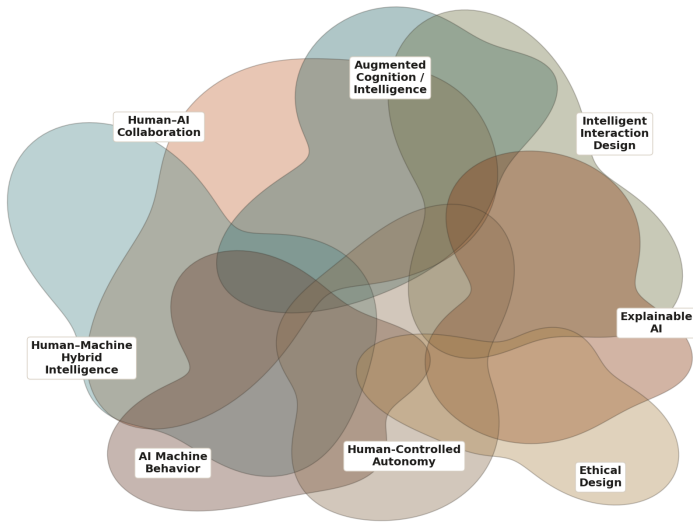


Figure 2.2: Overlapping research themes of human-AI interaction, adapted from domains described by Xu et al. (2021).

humans. Second, the technology should reflect human intelligence and should be used to advance the capabilities of humans. Third, human factors principles should be used to design systems that are explainable, useful, and usable. However, Xu et al. (2021) subsequently identified that HAI is at the nexus of HCI, AI, human factors, design and engineering, psychology, social science, data science, cognitive neuroscience, and computer science. These overlapping fields involve a number of intersecting research domains, including explainable AI, intelligent interaction design, human-AI collaboration, augmented cognition/intelligence, human-machine hybrid intelligence, AI machine behavior, human-controlled autonomy, and ethical design (see Figure 2.2).

Amershi et al. (2019) offered a design-oriented approach to characterizing human-AI interaction. Based on extensive review of the literature, they identified and codified 150 AI-related design principles and recommendations, distilling these to 18 guidelines in 4 broad categories. Interested readers can refer to Amershi's complete set of guidelines, which we summarize in the four contexts identified by the authors in Table 2.1. Interestingly, many of the specific guidelines are not specific to AI-enabled systems, but are relevant to many kinds of systems. For example, guideline G1, *"Make clear what the system can*

Table 2.1: Summary of Amershi et al. (2019)’s 18 guidelines for HAI.

Context	Summary of guidelines
Initial exposure	Clarify what the system can and cannot do. (2 guidelines)
During interaction	Be sensitive to context of time, task relevance, and social situations. (4)
When wrong	Enable efficient invocation/dismissal and correction, scope/scale the system, and provide explanation/justifications. (5)
Over time	Learn (cautiously) from user behavior and feedback, allow customization, and inform about consequences of actions and changes. (7)

do”, is good design advice for almost every consumer and electronic product.

Overall, the broad field of HAI may be criticized in a number of ways. First, it places the user and the system as separate “agents” and their interaction as the thing that needs to be studied. Other approaches center the user in the process (Human-centered AI), or consider the “system” as involving the humans, the technology, and the interactions (e.g., Cognitive Systems Engineering), or consider the human and intelligent system as cooperative players on a team (mixed-initiative interaction; human-AI teaming).

2.2.2 Human-Centered Artificial Intelligence (HCAI)

Human-centered AI (HCAI) is an evolution of the human-centered design movement to tackle similar problems in AI interaction (Xu et al., 2021). Shneiderman (2020) proposed that HCAI is a “Second Copernican Revolution”, which in contrast to the first revolution (that displaced humans from the center of the universe), returns them to the center of the design process, and supports AI applications that should be developed with the human user experience as the central approach. This approach is more prescriptive than the term human-AI interaction or many of its component disciplines, because it suggests a particular approach to research and practice.

In comparison to general human-centered design problems, most users do not and may be unable to understand how AI makes decisions and processes information. This creates special challenges for human-centered design of AI that may not be faced in other technologies.

Some HCAI theorists reject the idea that AI will one day replace humans or have all the same abilities as humans. Instead, they argue that humans should use AI as a tool that helps them to enhance their natural abilities, to make better decisions, and to provide abundant information (Shneiderman, 2021, 2020). According to this ideology, AI should be designed and used

with this goal in mind and with the understanding that AI cannot surpass humans. Importantly, Shneiderman argues that designs that attempt to mimic or recreate human-like abilities (including cognitive capabilities) will give way to more effective designs that take advantage of the unique advantages of the machine. Human traffic cops were replaced not by a robot with a whistle and arms, but by a network of lights and sensors that can manage traffic across an entire roadway more efficiently and effectively; a human washing dishes was not replaced by the Jetsons' humanoid robot maid Rosie, but by a box that washes dishes. Similarly, successful AI will not necessarily act or look like the human performing the same job.

Interactive human-centered artificial intelligence (IHCAI) describes practices within the discipline of HCAI that work toward enabling interactive exploration and manipulation of AI, purposefully designed with human benefit in mind (Schmidt, 2020). IHCAI should allow for users to understand why a specific conclusion is drawn and to understand how changing different factors affects the output. This should be in terms that a nonexpert can understand (Riedl, 2019). This will aid users, given that they are unable to form a theory of mind about AI systems. Users should be able to see who has control over the AI system and to know what data, knowledge base, and information the system is using. Understanding the function, purpose, and who is controlling an AI system allows for the user to remain in control of the system.

2.2.3 Cognitive Systems Engineering

A third important perspective is that of Cognitive Systems Engineering (CSE Woods and Roth, 1988), which arose to study how automation and technology were impacting work systems. These include industrial and control-room settings including power plant management and space flight and mission control, air traffic control and commercial aircraft, healthcare settings, military systems, and the like. Unlike the focus of HCAI and HAI, these often contain teams of individuals (instead of a single user) with overlapping but distinct specializations and responsibilities and networks of systems (instead of a single tool) with varying levels of reliability and capability. Each of these can be considered *cognitive agents*, and the goal is to characterize the cognitive work the system produces and how its parts interact. The initial goal was to study the humans as part of the system, and see how changes in the automation change the roles and interactions.

Although there does not appear to be a subdiscipline of CSE devoted to AI per se, the community has learned and adapted through several generations of AI and automation systems to understand AI as a component of a larger system, to understand how it changes the cognitive work that the system performs, and to address how to incorporate this into system design.

2.2.4 Allied and Adjacent Fields and Sub-Disciplines

There are other labels used to describe aspects of the intersection of humans, AI and automation, that are distinct focuses of some researchers. Some of these are akin to the research themes in Figure 2.2, but others are brands committed to specific approaches, researcher communities who do not only study human-AI systems, or other technical approaches. We mention and define them here to help further characterize the landscape of human-AI and human-automation research and design.

2.2.4.1 Human-Machine Teaming

This goes by several related terms, including, Human-Autonomy Teaming, Human-Agent Teaming, and Human-AI teaming, and is often abbreviated HAT or HMT. A common feature of all of these research approaches is that they take the team as the focus, and view AI or automation as a member of the team (which normally also involves human members). Whether AI can or should be treated as a teammate versus a tool remains somewhat controversial (Klein et al., 2004; Shneiderman, 2022; Naikar et al., 2026), but the approach is useful because it builds on a massive research base focused on human teams, and recognizes that teams involve both shared and unique information, goals, and skills, so HAT often involves studying humans and intelligent systems as collaborators with shared goals, complementary capabilities, and mutual dependence (Berretta et al., 2023; Chen and Barnes, 2018; Bradshaw et al., 2017). Chapter ?? covers this approach more completely.

2.2.4.2 Socio-technical Systems

Socio-technical systems research studies work settings as jointly shaped technical and social subsystems, arguing that automation and AI succeed only when both are designed together (Baxter and Sommerville, 2011); participatory and contextual design approaches in Chapter 3 apply this view.

2.2.4.3 Resilience Engineering

Resilience engineering examines how people and organizations monitor, adapt, and recover when they encounter surprise, stress, or failures (Hollnagel et al., 2006). The mission of this discipline is to design systems that are resilient to disruption, and study factors that impact this resilience. The focus can often be on primarily human or socio-technical systems where the stressors are not automation or AI (such as emergency rooms). However, technology, automation, and AI are often important weak points in the resilience of systems, and potential solutions to improve resiliency. It is closely related to cognitive systems engineering and work-domain analysis (Chapter 3).

2.2.4.4 Naturalistic Decision Making and Macro cognition

Naturalistic decision making studies how experts make judgments under time pressure, uncertainty, and real operational constraints, often using methods such as the Critical Decision Method (Klein, 1998). The community that has developed around these methods has expanded the domains of research to study many kinds of cognition in the context of work and socio-technical systems, which has been referred to as macro cognition (Klein and Wright, 2016). Overall, this frequently involves studying human-automation and human-AI systems, often with the focus on how expertise impacts and is impacted by intelligent software tools.

2.2.4.5 Human-Robot Interaction (HRI)

Human–robot interaction is the study of how people perceive, communicate with, and work with physical robots. Here, robots involve the entire range, from “dumb” industrial robots to humanoid and autonomous robots to social robotics and robot toys. The emphasis is often on embodiment, safety, and shared workspace (Goodrich and Schultz, 2008). This is discussed further in Chapter ??, and overlaps with social robotics discussed in Chapter 4.

2.2.4.6 Social Robotics

Social robotics focuses on designing robots that communicate, coordinate, and relate to people in socially meaningful ways, often leveraging anthropomorphic cues to support acceptance and interaction (Breazeal et al., 2016). Here, although the robotics often have multiple complementary integrated systems for control, machine vision, language, planning, etc., the focus is usually not on algorithms per se, but on the social, emotional, or trust responses. This is discussed more in Chapter 4 and trust implications in Chapter 5.

2.2.4.7 Coactive Design

Coactive design is an approach to human–robot teaming that focuses on *interdependence* as the central design problem for designing interactive systems with humans and intelligent agents. Its key principles are to design for observability, predictability, and directability (Johnson et al., 2014); see Chapter ??.

2.2.4.8 Distributed Cognition

Distributed cognition treats thinking as spread across people, tools, representations, and environmental structures rather than confined to individual human thinkers (Hutchins et al., 2000). The approach does not focus on intelligent

machines, but rather how humans extend their cognition to embed memory, perception, and other cognitive processing in the environments they build and design. For example, in “How a cockpit remembers its speeds”, Hutchins (1995) examined the cognitive operations of the commercial aircraft cockpit system that includes teams of pilots, sensors, displays, printed documentation, and automation, describing how artifacts within the cockpit (marks on the speed indicators, lookup tables throttle positions) distribute the cognition outside the mind of the pilot.

2.2.4.9 Human-Centered Machine Learning and Human-in-the-Loop Machine Learning

Human-centered machine learning is a design philosophy akin to Human-centered AI, with perhaps a slightly narrower focus. It involves how machine learning systems can be designed and trained to support humans in general, and users specifically (Holstein et al., 2019). This involves prioritizing trustworthiness, user workflows, fairness concerns, and practitioner needs, and extends to societal/ethical issues.

HCML should not be confused with human-in-the-loop machine learning (HITL-ML). In general, HITL-ML are approaches to embed humans in the machine learning process to enable AI/ML systems to learn and use information that may not be easily accessible otherwise. This can involve human data labeling, model correction, validation, and selecting preferred outcomes, and otherwise interactive refinement of learned systems (Mosqueira-Rey et al., 2023). A specific and popular method for implementing this approach is Reinforcement Learning with Human Feedback (RLHF), although HITL can refer to many other human activities that are used for machine learning, such as human annotation or labeling, error detection, etc. HITL-ML can support HCML if it is used specifically for that purpose, such as supporting broader goals of *alignment* (having the AI operate in ways consistent with human values).

2.2.4.10 Mixed-Initiative Interaction and Planning

Mixed-initiative interaction describes systems in which humans and computers can each take the lead, contributing their respective strengths toward a shared goal (Allen et al., 1999). This research subdomain has its roots in conversational agents, often with the goal of a human and AI system interacting via spoken language to accomplish a task. Although research on this pre-dates conversational chatbots like Siri, Alexa, and LLM-based systems, many of the examples in this research domain anticipate the kinds of interactions envisioned by modern conversational AI assistants and LLM chatbots. For example, Novick and Sutton (1997) used the following scenario to illustrate a mixed-initiative interaction, where S is the dialog system and U is the user

S1: I can schedule an appointment for you. What day?
U2: Tuesday.
S3: What time on Tuesday?
U4: In the morning.
S5a: What time on Tuesday morning?
S5b: What about 9:30 a.m. Tuesday morning?

Figure 2.3: Mixed-initiative interaction scenario used by Novick and Sutton (1997). S represents the dialog system and U the user.

(see Figure 2.3). Here, the final user response U4 implies that the system should take initiative and provide a suggestion such as S5b; a system that declines to take control as in S5a is argued to lead to repetitive interaction and unconstructive exchange. Here, in comparison to typical interactions with AI chatbots/assistants that tend to be command-response driven, this approach has the potential for interactivity, extended use of in context, and correcting/changing/adapting information based on feedback.

Although mixed-initiative interaction traditionally involves linguistic interaction, the human and machine often have a common or shared task or goal that is non-linguistic. For example, Mohan and Laird (2012) demonstrated mixed-initiative interactions in a tank/missile game in which components of a task could be handed off between the human and agent; coactive design (above) extends similar ideas to embodied human–robot teams.

2.3 Evaluating AI and automation for human use

In this section, we provide ideas arising from the different fields that can help a human factors engineer analyze a system involving intelligent software or AI, anticipate problems, and understand potential solutions.

2.3.1 Analyzing Capabilities and Limitations

AI and automated processes make life easier for workers all over the world. Automated machinery on assembly lines significantly increases production while reducing human interaction with products. Despite the rewards of automation, human-computer interaction can prove challenging for workers when problems occur. Troubleshooting heavy machinery can not only be confusing, but dangerous for the operator. While not every instance of troubleshooting AI and autonomous processes is hazardous, they can be confusing and difficult to correct. A number of researchers have identified three broad aspects of systems that users should understand: the system's **capabilities and purpose**, its **limits and limitations**, and how to **troubleshoot problems and work around its limitations**. Borders et al. (2024) suggested that capabili-

Table 2.2: Primary functions of software tools (after Mueller et al., 2009).

Function	Definition
Data handling and data generation	How input or training data is generated
Representation and modeling	How the system represents information, or how the user models the real-world system
Understanding computation and algorithms	How input is transformed using computation or mathematical algorithms
Output, display, visualization	How results are presented to the user

ties and limitations of technology are critical aspects of a mental model, and that workarounds are related capabilities of the users. Likewise, Hoffman (2017) discussed *justified trusting* and *justified mistrusting*: understanding the competencies and limits of a system.

2.3.2 Components of an AI or Automated System

AI and automation systems generally have a number of information-processing functions they support (see Table 2.2). These include at least one of the following functions: representation and modeling, data handling and data generation, understanding computation and algorithms, and output, display, and visualization (Mueller et al., 2009). To improve human-AI interaction, a designer should examine and understand each of these areas as potential domains for improving or supporting users of the tool.

This support may take on a number of forms that enable human-centered AI design. For example, one might consider two aspects of data and input: training data and user input. Identifying where training data comes from may help inform a user about the capabilities and limitations of the system, and it may suggest ways the tool might be improved by allowing the user to change or update a training corpus. Considering how information is input may lead to a redesign in interaction. By considering the underlying representation of a GPS-based routing system, one might learn that optimization does not incorporate traffic or construction; a user who expects this will be better prepared to work around this limitation. In image classification, algorithms often compute a “soft-max” value that looks like a confidence or probability score, and allows ranking of multiple likely options. But this can be easily misinterpreted. By considering this, a developer may decide on an alternative algorithm or output that is less likely to be misinterpreted. In any of these cases, the consequences might be design changes, changes to the task, or improved training and user instruction.

Considering these aspects of an intelligent system does not mean the user or operator needs a deep understanding. It may be enough to help the user

Table 2.3: Shneiderman (2020) suggests automation can be designed with independent levels of human control and autonomous control. We illustrate this with four examples of transportation systems, based on Figure 5 of Shneiderman (2020).

	Low automation control	High automation control
High human control	Commuter bus (controlled by a human driver)	Lane-keeping plus adaptive cruise control (the human has control, and so does automation)
Low human control	Roller coaster (gravity-powered; track brakes)	Airport automated tram (no human operator; sensors for passenger entry and collision)

develop a functional understanding or working knowledge of that aspect of the system. For example, a train engineer may understand which controls are responsible for stopping or controlling speed, but they may not need to understand the physics of diesel-electric locomotives. Similarly, a driver with antilock brakes (ABS) is unlikely to know what the sensors and algorithms are used to detect when a wheel is slipping, but they may be able to learn situations and conditions in which the ABS is likely to fail and how to work around its limits. As discussed in Chapter 1, Conant’s 1970 good regulator theorem applies: for a user to be a good regulator of a system, they must have an internal model of that system. But that model does not need to be complex or complete to be useful.

2.3.3 Shared Responsibility

Fully autonomous systems are becoming more prevalent, but often have fallback human operations, supervision, monitoring, and control. Even fully autonomous robo-taxis such as Waymo are under the control of the rider (who sets its destination), and have fallback human operators (or tow-truck operators) to assist when they get truly stuck. But complete automation is not always desirable. Humans bear the responsibility for systems, and so they need to provide input and have some level of control over them. They currently share some responsibilities with AI to complete tasks in autonomous systems. Manual and semi-manual controls tend to give the user more freedom but can decrease speed and efficiency (Andersson, 2008). The extent of user control depends entirely on system and AI capabilities, which derives from the complexity of the task/action. Shneiderman (2020) suggested that this shared responsibility is not simply a one-dimensional mixture. Rather, systems can be designed so that human control can range from low to high, and automation control can also range from low to high (see Table 2.3).

In theory, human-required control should decrease as AIs become more complex and capable of operating autonomously. In a fully autonomous system, the user assumes the role of a supervisor; they have no control over operation but oversee processing, perform troubleshooting, and error correction. However, in reality automation may increase the need for humans, but at a different point in the system. Today, most systems do not operate fully autonomously; the user still fulfills this supervisory role while working with the system. Operators of autonomous vehicles put trust in the system to follow a road and avoid accidents, but still monitor the environment and system behavior. Failure to supervise might end in a failed handover and result in an accident.

2.3.4 Boundary Conditions and Competence Envelopes

All systems have their limitations, and it is up to the user to determine the boundary conditions and proper use conditions (competence envelope) for the system. The competence envelope refers to what the technology can do (or can do well) in different circumstances (Hoffman et al., 2013), and the boundary conditions might be considered the dividing point between the situations where the technology performs well and where it fails. Whether this is relayed in a manual or guide, or through first-hand experience or second-hand discussions with other users, understanding the breaking points of a system is necessary to avoid disaster. In an industrial setting, boundary conditions might be set by the manual, or the machine may have built-in limiters. In most situations the user can monitor and fine-tune around these set conditions and optimize the system to find a balance between speed and accuracy of the system.

Having a strong understanding of boundary conditions can help users increase trust in the autonomous system, because it may move them from a position of general distrust (maybe because they have seen failures) to one of encapsulated trust (knowing conditions inside the boundaries of the competence envelope in which they can trust the automation). As a user becomes familiar with boundary conditions of their system, they can understand under which conditions they can trust it, and either anticipate this (to take over or stop the system) or avoid it through other control means. This increased trust was demonstrated in a study of smart home automation, in which participants displayed more trust in semi-autonomous systems over full autonomy (Schomakers et al., 2021). Users seem to prefer to have enough control to avoid violation of boundary conditions. As a supervisor of an autonomous system the operator relinquishes this control to the system and trusts it to keep within boundary conditions.

2.3.5 System Errors and Malfunctions

System malfunctions and mistakes are also characteristics of many automated systems, and this can be distinct from its competence envelope, but can have some of the same consequences. When a problem occurs, the user must identify a course of action to correct it. In this supervisory role, the operator may have liberties with how to troubleshoot and most effectively solve the problem (Ogorelysheva et al., 2023).

Aspects of AI and automation that are important for users are primarily determined by the complexity of the system and how difficult problems are to solve. A user that fails to understand the purpose of the system may misuse it or not recognize errors. Many issues can be spotted if the user or operator can understand the key functions of the system, including how data is collected and processed. Understanding and identifying boundary conditions allows the operator to avoid mistakes. As AI and autonomous systems become more complex and self-sufficient, less human interaction will be necessary and the operator will fall into a complete supervisor role.

2.3.6 Shifting Bottlenecks and Transformed Cognitive Work

One recurrent theme in cognitive systems engineering is that automation and AI cannot replace the human, who has ultimate responsibility. Although they may replace and displace the previous work and can reduce workload, this creates new bottlenecks at different places—sometimes in more critical situations (Sarter et al., 1997). This can occur through clumsy automation (Wiener, 1989), in which the benefits occur for low-workload routine periods, but the cost is that a human must take over during time-critical, peak workload, high-tempo periods—which they may be less well-equipped to deal with because they offloaded routine work.

Much of the impact of AI chatbots can be examined through the lens of how it transforms work. For example, in education, a student can use it to displace difficult and challenging exercises that teach them how to think or operate. This is having huge impacts on many traditional educational exercises and disciplines, including the teaching of computer science, the practice of writing and revising, and on-line coursework in general. AI makes it easier for a computer science student to complete the work, but shifts the learning to a point in the future where they are unfamiliar with the concepts they are supervising the construction of. Similarly, AI systems can help develop detailed plans of action for carrying out both small tasks and large projects. But if the team turns over planning to an agent, the bottleneck may shift from a planning process (which is made easier) to more failures during the plan (since they no longer bring their expertise to the planning process, and may no longer recognize how if the plan is working out). Here, a small savings during initial

planning could be transformed into a catastrophic failure during execution.

For the Human Factors Practitioner

This chapter offers a number of schools of thought to studying and impacting human-automation or AI systems. Regardless of your perspective, they each have important insights, but they also all have blind spots.

- Understand the context of the work being done. General solutions may point in a direction, but will usually not work.
- Understand how the work changes in response to automation.
- Identify clumsy automation consequences that create higher-cost but rare failures that need to be anticipated.
- Recognize that no systems or people are autonomous—they are all interdependent. Identify how those interdependencies will change and be impacted by automation or AI.
- Allow automation and AI to lean into its strengths, as its most effective approach may not be the way humans worked.

2.4 Summary

This chapter describes several research communities concerned with designing AI and automation for people. In addition, it outlines some useful concepts that can help a human factors engineer analyze a system that involves human-AI interactions, and identify possible failure points, improvements, and the like.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction, Michigan Technological University, Spring 2024.

Author Contributions

KU: Writing and background research for Section 2. BW: Background research and writing for Section 3. STM: Conceptualization; background research, writing sections 1–4; general editing.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Uliniski, K., Woolman, B., & Mueller, S. T. (2026). Human-AI interaction research: An overview. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Stages of Technology Adoption and Development for Human-Centered AI Tools and Methods

Erin Matas, Lauren Sprague, and Shane T. Mueller
Michigan Technological University

Abstract

In this chapter, we discuss the development life cycle of an AI or automated system using frameworks developed from several perspectives, including user acceptance/adoption, product development, and research and development acquisition, and post-deployment adoption. Each of these frameworks provide specific user-centered ways that human factors and usability researchers can influence, evaluate, and mitigate potential problems. Synthesizing these different perspectives, we propose that there are *critical periods* at which individual human factors methods and human-centered AI design can have the most leverage. Using this as an organizing principle, we describe a variety of tools, methods, and evaluations that might be applied at different points during a system's life cycle to develop human-centered AI systems and support human-AI interactions.

3.1 Introduction

There are numerous measures, interventions, tools, and methods used by human factors practitioners and cognitive systems engineers to improve the

usability and usefulness of AI systems, and adopt a human-centered design approach for AI. Human-centered designers often decry that they need to be involved in the initial development stages in order to have the greatest influence, and build the most useful tool. Although this is undeniably true, it also ignores important practicalities and constraints in technology development. Furthermore, the different methods and tools used for human-centered design each have a critical period—a point during the development of technology in which they will have a maximum cost-benefit.

In this chapter, we examine several existing stage-based approaches to understanding the adoption and development lifespan of technology in general, and specifically for intelligent software tools and AI. Some arise from the user perspective, others from the developer perspective, and others from the acquisition perspective. We will then suggest a new framework loosely inspired by research on human language development: *critical periods*. This approach acknowledges that human-centered researchers can have different influences in different stages of technology development and adoption, and have developed different tools and approaches for these different critical stages.

3.2 Technology Acceptance Model (TAM) and AI-specific revisions

Any technology can be examined in terms of “stages” in its lifecycle or lifespan. One perspective is to examine adoption from the user’s perspective. Perhaps the most popular such approach is the Technology Acceptance Model (TAM) (Davis, 1989), shown in Figure 3.1. This model was initially developed to characterize how users were adapting to new innovations such as personal computers at home and in the workplace. The approach treats the system itself as external to the model. A user forms cognitive judgments of usefulness and ease of use, an affective attitude toward using, and a behavioral response of actual use.

Here, initially, potential users may perceive how useful a technology is (by watching others, through advertisements, testimonials, etc.), and how easy it is to use. Ease-of-use feeds into usefulness, and both impact their attitude toward actually using it, which was considered an affective response. This is the ultimate predictor of actual system use or adoption—the behavioral use. One of the implications of the model is that perceptions about ease-of-use for novices can play an important role in technology adoption, which suggests that good human-centered design principles at early stages are likely to pay off. However, the model itself is not tailored to AI in particular, and does not incorporate many of the additional concerns users have about AI systems (e.g., trustworthiness, alignment, safety, to name a few.). It is also limited to tracing

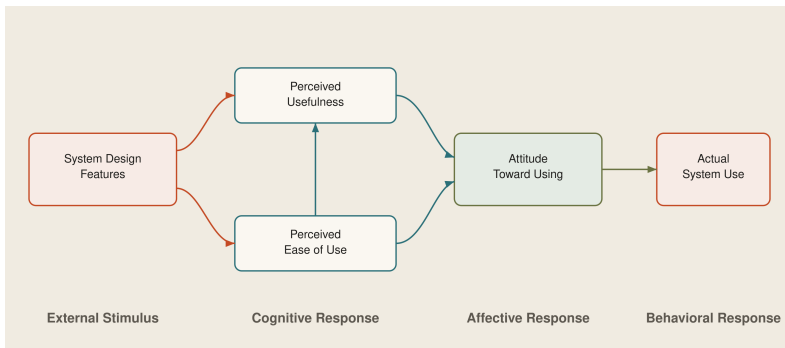


Figure 3.1: The Technology Acceptance Model (Davis, 1989). External design features shape perceived usefulness and ease of use; those judgments form an attitude toward using, which predicts actual system use.

a user's attitude from a fixed "input" toward adoption; it does not address what happens after adoption, or how changes to the system may require additional assessments by the user.

More recent research has examined the TAM in the context of AI and automation. These have typically incorporated additional stages in the TAM, or specified different influence in specific stages. For example, Baroni et al. (2022) proposed an AI-TAM model incorporating AI function, trustworthiness, and output collaboration into the TAM path diagram. They suggested explainable AI (XAI) can have positive influences at many of these stages. Choung et al. (2023) provided support for this by showing trust-in-AI impacted perceived usefulness leading to more positive attitudes about AI. Researchers have discussed fairness, bias, transparency and integrity (Borenstein and Howard, 2021), and Mustofa et al. (2025) suggested that trust and ethics are moderators of the relationship between attitude and adoption. Overall, these adaptations to TAM for understanding AI include:

- Subjective norms / social influence (Venkatesh and Davis, 2000; Mustofa et al., 2025)
- Trust in AI (Baroni et al., 2022; Choung et al., 2023; Mustofa et al., 2025).
- Perceived output quality and collaborative intention (Baroni et al., 2022)
- Ethics (fairness, bias, transparency, integrity of use) (Borenstein and Howard, 2021; Mustofa et al., 2025)

Ultimately, TAM and its extensions describe *whether* people will use an existing system. The specific adaptations to TAM suggest many different ways human-centered researchers can gain purchase on the problem, and potentially influence initial adoption. Many of these approaches are covered in greater detail in later chapters (including trust/trustworthiness, explainable AI,

fairness/bias, transparency, etc.), and so this handbook covers many possible ways to improve these systems from a human-centered approach. However, TAM focuses on initial adoption and influence of users. There are other ways to understand how human-centered approaches to AI and automation can impact systems, by examining the broader product development timeline.

3.3 Product Development Lifecycles

An alternative approach is to examine how a product is built. Perhaps the most classic model is referred to as a “cascade”, or “waterfall” development approach. It proceeds once through requirements, design, implementation, test, and operation (Royce, 1970). Royce (1970) potentially suggested this model as a warning: without feedback between stages, the process is risky and likely to fail. Boehm (1988)’s response is a “spiral” development model that repeats a cycle: first determine objectives, next analyze risk, then develop and verify, and finally plan the next loop. In theory this helps the system grow in capability and fidelity to the original vision at each cycle. Current popular software and AI product development follows adaptations of this referred to as Agile development, often using its special case predecessor called *Scrum*, which is characterized by short sprints, a working increment to the product, and inspect-and-adapt reviews (Beck et al., 2001; Schwaber and Sutherland, 2020).

These basic approaches provide an important context for how human-centered design can impact technology development. In fact, human factors and human-centered design requires the same kind of iteration. Gould and Lewis (1985) offered three principles: early focus on users, empirical measurement, and iterative design. Nielsen (1993) turned those principles into a usability-engineering lifecycle of analysis, design, and test, repeated until the interface is good enough. Design thinking (Brown, 2008) describes a cycle of empathize, define, ideate, prototype, and test (sometimes followed by implement). This basic loop has been standardized by ISO for human-centered design of interactive systems (International Organization for Standardization, 2019): understand the context of use, specify requirements, produce design solutions, and evaluate.

Consequently, the software development cycle and the human-centered design cycle are similar, but it is typical that human-centered design is secondary—it represents one of many technical and product constraints that need to be managed. These include security, bugs, licensing, external dependencies, business operations/sales, etc., and so human-centered design may need to fit in with the process and be opportunistic about what they can accomplish. The fast-cycling of Agile approaches also means that human-centered designers have many opportunities to have an impact, but

they need to operate at a tempo such that their results are still relevant once they are complete—thus they should understand product roadmaps, redesign goals, and the like.

3.4 Human Readiness Levels

Another approach to understanding technology is the “Technological Readiness Level” framework used by NASA, the government and military funders of research and development programs to understand the maturity of a technology (Mankins, 1995). The TRL describes 9 levels of maturity, from TRL-1 and TRL-2 (describing basic research) to TRL 5-6 (Technology development and demonstration), to TRL-9 (system test and operation). On its own, this framework does not offer much direct understanding of whether the technology supports human users—it can be applied equally to rockets, naval vessels, and software tools. Importantly, there have been many products that have achieved high technological readiness, that have been fielded, but are either difficult to use, or fail to be adopted by users because of a failure of human-centered testing and design. One famous such example is the Mars Climate Orbiter, which had two subsystems that used different units (imperial vs. metric), resulting in the loss of the spacecraft. It is unclear whether a TRL assessment was conducted on the Boeing 737-Max MCAS system, but similar certifications are required by the FAA. However, its failure has been attributed to a failure to understand human factors in aviation safety (Endsley, 2019).

In response to these limitations of the TRL assessment, the Human Factors and Ergonomics Society has developed a similar system to characterize Human Readiness Levels (HRL) (See et al., 2019; Human Factors and Ergonomics Society, 2021). These map onto TRL 1-9, but characterize the extent to which a design has incorporated human-centered design and evaluation. HRL 1 to HRL-3 establish human principles, concepts, and requirements; HRL-4 through HRL-6 mature the design through part-task tests, prototypes, and high-fidelity simulation; HRL-7 through HRL-8 verify and validate human-system performance with representative users on the completed system; and HRL-9 is operational use with systematic monitoring.

The guidelines suggest that HRL should match TRL, because by TRL-6, the engineering design is often fixed, and human needs become expensive or impossible to incorporate (Human Factors and Ergonomics Society, 2021). HRL can often precede TRL, using various methods for understanding envisioned worlds (wizard-of-oz studies, part-task tests, human-centered requirements gathering, and basic human research).

Like the other lifespan approaches, the TRL and HRL model offers some understanding of technology development, adoption, and acquisition. In particular, the HRL model can help technology developers make stronger cases

for including human-centered design and assessments as a critical part of the technology development program.

3.5 Threat-based stage analysis

Most of the approaches we have examined describe the timeline of development through adoption. Although the cyclic nature of several of these approaches can certainly incorporate user feedback and evaluation, it is also the case that even for fixed systems that have found a community, there are stages of deployment and use, as people learn the tool, become reliant on it, explore different use cases, and experience failures.

A framework that addresses this *collaboration lifecycle* (Ahmed et al., 2026) focuses on sociotechnical risks as AI is adopted and fielded. It maps human–AI work across four sequential stages—task allocation (how tasks and authority is divided among a team of humans and AI systems), interaction (communication issues, explanation, cognitive burden/overload), feedback (continuous updating to enable appropriate trust adjustments), and adoption risks (mostly secondary issues such as de-skilling, automation bias, and AI anxiety). Although these stages do not preclude continued development and improvement of an AI or automation system, the focus is on stages of activities ongoing amongst a team that is incorporating these tools into their work.

Ahmed et al. (2026) described six recurring risk clusters that are observed across many domains (including media, business, to education, healthcare, and defense): trust miscalibration, cognitive burden, accountability gaps, capability erosion, goal misalignment, and AI anxiety/technostress. Importantly, they describe a “fragmented mitigation landscape”: for these different risk clusters, there are individual mitigation frameworks, but many apply narrowly to specific domains or risks. Because these risks are socio-technical, their proposed mitigations are often central to human-centered design. For example, for the risk cluster involving cognitive burden and overload, many human-centered solutions (from automatic monitoring for fatigue, to minimizing distraction, to enabling the user control tempo) have been proposed, and it is likely that for any system, user testing is needed to determine acceptable and useful approaches.

3.6 Critical Periods

Overall, these different stage-models of technology development and adoption span from initial conception of a tool to late stages where a team must change how they work to adapt or accommodate the tool. In ideal situations, a strong human-centered analysis will precede development, as it is likely to help understand what the problems are and what the viable solutions are. But this

is not always realistic. But it is also certainly the case that many products have completed the entire product development timeline and failed because they do not address the problems real users had, or were difficult to use or understand. Sometimes, conducting human-centered design or analysis can have no impact if it is done too late—after the basic system architecture is developed. It may require a complete refocus of the system to help address the issues and problems people really have, rather than the solution a piece of technology offers.

However, these two extremes are not always the case. It can often be true that resources or time does not exist to engage human-centered design prior to a proof-of-concept; it can also be the case that human-centered design can have an impact on technology that is quite mature—as many of the mitigations described by Ahmed et al. (2026) demonstrate. Furthermore, human-centered design can sometimes demonstrate effective wins at later phases, which can provide buy-in for future product development.

In response, we propose a notion inspired by research on human language development: *critical periods*. The basic notion is that many human-centered design and evaluation techniques are most appropriate, and can have the biggest impact, at specific stages of technology development. Some are appropriate before the system starts development; others are only available once a community of users exists, or a team has incorporated a tool in their work. We use a development-lifecycle model (Figure 3.2) that includes initial planning, development, deployment, change management, and maintenance.

In comparison to the other phase models in this chapter: waterfall, spiral, Agile, and design thinking say how planning and development are conducted (Royce, 1970; Boehm, 1988; Beck et al., 2001; Brown, 2008); Gould and Lewis's iterative principles and Nielsen's usability-engineering cycle are the human-factors versions of these cyclical design approaches (Gould and Lewis, 1985; Nielsen, 1993; International Organization for Standardization, 2019). HRLs score human-use maturity across the lifecycle timeline (1–3 planning, 4–6 development, 7–8 deployment, 9 maintenance) (See et al., 2019; Human Factors and Ergonomics Society, 2021). TAM extensions mostly focus on user adoption immediately following development, and Ahmed's risks focus on later deployment stages.

Many elements of these critical periods are logically constrained by the feasibility, cost, and information gained from applying a method at distinct points in the developmental timeline. Developing user support or assessing trustworthiness of an envisioned tool may be wasted effort, and a better use of resources at that time would be to evaluate user expectations and stakeholder desires (HRL 1–3, not 7–8). On the other hand, many methods may do little good if applied “too late”—as when HRL lags TRL, and the system is already frozen (Human Factors and Ergonomics Society, 2021).

In the remaining sections, we will describe the stages in Figure 3.2 and

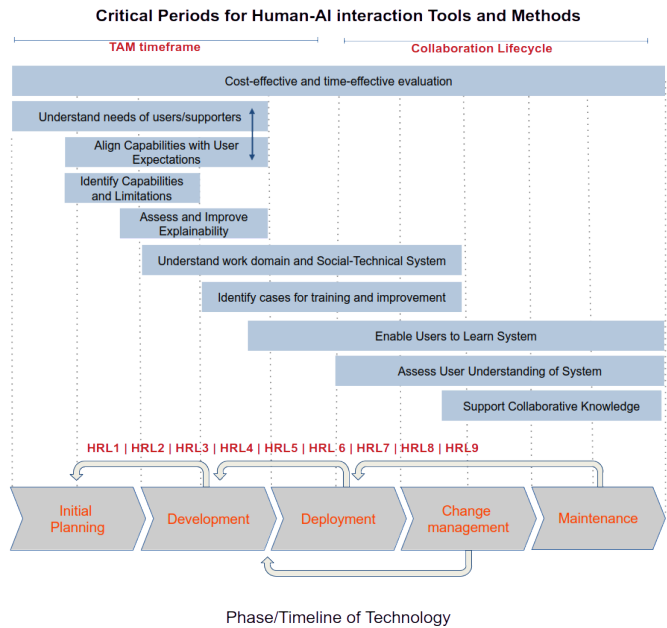


Figure 3.2: Critical periods in the technology development lifecycle, and distinct goals of human-centered AI design and practice applicable during different stages. TAM model typically describes interventions that can be done during planning and development; (Ahmed et al., 2026) describes risks that occur after deployment; HRL levels map roughly onto the timeline from left to right.

some candidate methods and interventions proposed to address them. In many of these cases, the methods will be dealt with in much greater detail in subsequent chapters of the handbook, and we will offer pointers to those resources as well. Of course, these suggestions are only representative of the many different human-centered approaches that exist.

3.7 Understanding needs of users/supporters and other stakeholders

When developing technology and tools that utilize AI and machine learning, the first step should be to identify and understand the needs of users and supporters of the tool or technology. There is a large body of research on stakeholder analysis (Brugha and Varvasovszky, 2000; Schmeer, 1999; Kenyon et al., 2009; Reed et al., 2009), and a growing literature on stakeholders AI systems (Hoch et al., 2023; Hoffman et al., 2023a; Plass et al., 2022; Park et al., 2022; Deshpande and Sharp, 2022; Subramanian et al., 2024; Kim et al., 2024), including government guidance on responsible AI acquisition and agency reporting (Office of Management and Budget, 2024a,b). These have used a variety of approaches, which we focus on next.

3.7.1 Who are the stakeholders?

A number of different stakeholders have been considered and depends on the particular technology but these have included end users, developers, researchers, regulators and legal experts, policy-makers, contracting/procurement, trainers, system evaluators, program managers, business managers, employees, AI experts, data scientists, machine operators, data protection officers, business experts, HR teams, physicians, and patients, to name a few. Notably, Hoffman et al. (2023a) examined stakeholders with respect to their desires for explanation, and found that most stakeholders played multiple roles, and so it is important to elicit feedback not based on job title but based on various goals and use patterns of any individual.

3.7.2 Methods for stakeholder analysis

There are many approaches to understanding stakeholder needs when planning a project that involves intelligent software. We discuss five that may be useful depending on the stakeholders, resources, and timeline of a given project. A useful strategy is to use a minimum-necessary rigor approach (Klein et al., 2023), striking a balance between effective information and efficient data collection that has the best chance of impacting early design.

Informed stakeholder survey. Surveys are typically an inexpensive and fast way to get stakeholder feedback. The advantage is that many stakeholders can be reached at low cost. Such surveys may cover hypothetical scenarios, evaluate specific values and concerns, and allow more open-ended feedback. Creating representative, valid, unbiased, and informative surveys can be difficult, and is a critical skill for human factors researchers. Important issues may not be known before the survey is written, so a poorly designed survey

may miss what matters to different stakeholders. Different stakeholders may also require different lines of inquiry, so multiple surveys may be needed even if each has only a few respondents. Response rates can be extremely low, and the effort needed to encourage responses may be better spent on other approaches.

Focus groups and interviews. Focus groups and one-on-one interviews are commonly used to gather stakeholder concerns (Chapman and Shigetomi, 2018). Interview methods can allow a more intense understanding of situations, use cases, reasoning, and potential problems. Focus groups and interviews allow discussion and discovery that surveys often miss. Successful implementation often requires skilled facilitators who also know the domain (e.g., AI and the implementation field such as health care, military, transportation, etc.). The material can be costly or time-consuming to code, and without systematic process the thematic analysis can be biased by whoever extracts the themes. We discuss some of these methods in detail in Chapter 6 in the context of assessing trust in automation and AI.

Charrette for stakeholder engagement. For especially complex projects that are likely to have large impacts on—and opposition from—communities, the charrette process (Lennertz, 2003) has become an important planning and engagement technique. These are usually extended meetings among small groups of stakeholders, often facilitated by an independent and experienced moderator, that attempt to foster ownership and reduce conflict among parties with conflicting interests.

Compared with focus groups, a charrette is more intense and focuses on participatory planning rather than feedback alone. For many AI implementation projects a charrette may not be needed. The advantage emerges when a community may (rightly or wrongly) object to the program, and a charrette can anticipate that and allow input. Any system that will rely on automation and algorithms for decisions previously made by humans, with safety or fairness impacts, may require many stakeholders to identify a viable trajectory for change. The disadvantages are experienced facilitators, extensive planning and event coordination, and a time commitment from many stakeholders; the method is designed to turn planning into a community process.

The Stakeholder Playbook. To facilitate stakeholder analysis for AI systems, Hoffman et al. (2023a) proposed the Stakeholder Playbook, with a special focus on explainability. Like a football playbook, it identifies both roles and goals (related to desirements and requirements) found in previous studies. Stakeholders often play several roles, which can lead to different and sometimes conflicting goals. The playbook can help identify which stakeholders are

Table 3.1: Stakeholder Playbook major themes for AI systems (Hoffman et al., 2023a).

Stakeholder	Goals
Jurisprudence	Analysis of system biases, assumptions, and bounds
Contracting/procurement	Analysis of fitness for use in the operational environment
Development team leader	Analysis of how the system integrates with other systems
Developer	Access to use cases that represent the operating context
System integrator	Explanations at a detailed technical level to “look under the hood”
Trainer	Access to edge cases to support training users to anticipate problems
System evaluator	Access to feedback from prospective users
Policy-maker	Descriptions of system limitations and weaknesses
End-user; adopter	Cost–benefit analysis of tools with respect to goals

especially relevant for implementing AI, some of their common concerns, and a basic guideline for interviewing others. Table 3.1 summarizes major role–goal themes; Figure 3.3 shows the more detailed end-user playbook.

Figure 3.3 shows the more detailed end-user playbook, including explanation desirements, access requirements, and cautions. A focus group can walk through a subset of those items rather than treating the playbook as a substitute for analysis.

The premortem exercise. The general stakeholder methods are useful to understand how the product will fit within its organization, market, or community of users. For smaller teams that have representative members from a number of stakeholders (designers, developers, management, users communities, etc.), a lightweight alternative is the premortem exercise (Veinott et al., 2010). It has proven successful at improving plans by identifying potential problems from the perspective of multiple stakeholders.

The premortem is designed to take around 30 minutes during initial project meetings, when stakeholders are already engaged in planning. A leader makes a statement similar to the following: “Let’s assume we are several years in the future, and the project we are trying to accomplish has failed. Take a few minutes to think about *why* it has failed, and write the reasons down.” The prompt can be narrowed—implementation failed, or the system was implemented but did not produce the intended outcomes—but the open-ended probe is usually enough to surface unforeseen kinds of failure.

The group takes turns describing their reasons and votes when they share a similar reason, to mark importance. When that discussion is complete, a

End-user; adopter

Explanation desirment: global and local explanations that are satisfying and consistent with how the system actually works.

Explanation desirment: explanation of data inputs and how the system processes the data.

Explanation desirment: results of a cost-benefit analysis of different tools, with respect to the user's goals and responsibilities.

Explanation desirment: explanations need to strike a balance between superficiality and technicality.

Explanation desirment: explanations that support troubleshooting and system maintenance.

Explanation desirment: intuitive displays for visualizing the data, to understand whether the data might be inadequate.

Access requirement: to the system development team— trusted developers and software engineers.

Access requirement: ability to explore the system behavior by “poking around.”

Caution: end-users sometimes do not care or don't need to know “how it works.”

Caution: end-users often desire better explanations than the ones that are provided.

Caution: continuing explanation is required as the input data, the work system context, or the operational environment change.

Figure 3.3: Example avenues of investigation for end-user stakeholders (Hoffman et al., 2023a). Used via CC-attribution license.

second round identifies which potential failures can be anticipated and avoided, and assigns actions to stakeholders.

The advantage is that it is lightweight and can be run without substantial training. Compared with interviews, focus groups, and charrettes, it is less intense but can still be effective. Its focus on project failure directs discussion toward a common goal of successful implementation. The short timeframe also limits how well it can identify complex problems or solicit feedback from a broad (and possibly antagonistic) set of stakeholders. The approach has been evaluated in a number of contexts, including for the hypothetical planning of AI-based systems. For example, Kannan (2026) examined a structured premortem focused on AI systems, and found that additional prompting about user expectations helped generate reasons for failure that were more likely to be related to AI, were more likely to be controllable, and were judged higher quality than unprompted pre-mortem reasons.

3.7.3 Summary of stakeholder analysis in Human-AI interaction

This section covered a number of systematic approaches to understanding stakeholders. Some are simply to help design and usability, others help characterize social impacts of a system and community opposition. For human-AI interaction, several findings have emerged from stakeholder analysis:

- Different users have different explanation needs for a tool (Kim et al., 2024).
- Users desire personalization, so that recommendations are tailored to individual differences of users or patients (Subramanian et al., 2024).
- Predictions should be customized to specific cases (Subramanian et al., 2024).
- Information included in predictions should be tailored to the decision context (Subramanian et al., 2024).
- Different stakeholders have different perspectives on fairness (Park et al., 2022)

In the next section, we will discuss *user expectations*, which may come out of stakeholder analysis, or could be a more focused evaluation of existing or prospective users.

3.8 Understanding user expectations

Another goal that can be worked toward early in the development process is to understand expectations of users. This can take on various forms, but understanding the initial goals of users can help design a system that is responsive to those expectations, or develop training and marketing that will inform users about the capabilities and limitations of the system. Chapter 4 of this handbook covers a number of anthropomorphic expectations users tend to have in AI and automation: they often expect that it will behave both socially and cognitively like a human, although they can learn differently.

Zhang et al. (2021) examined human expectations of AI team-mates in multiplayer online games, and suggest expectations cover a broad range of behaviors: “instrumental skills for in-game tasks, shared understanding between humans and AI, communication capabilities, human-like behaviors and performance”. Many expectations and desires of players were not current capability of the system, and users would make suggestions about how their expectations were not met (e.g., “I think voice commands would be really good”), which could lead to changes in design or development. Thus many expectations center on ability or performance capabilities of the system.

Linja et al. (2022) reviewed user expectations on autonomous driving that appear in social media forums, and suggested they fall into several categories, including legality, safety, transparency, and anthropomorphic performance.

Adding to these performance/capability and alignment (discussed further in Chapter 4), we have useful potential evaluation guidelines to understand user expectations of AI technology. These include:

- **Performance/Capability:** The system is able to accomplish the task/job it is meant to under normal conditions. How fast and accurate is it compared to humans?
- **Legality:** The system follows relevant laws, rules, and regulations. Does it violate fairness? Does it use information it should not? Does it violate “rules of the road” that are not explicitly codified?
- **Safety:** The system does not make errors that endanger users, cost money, and violate trust. Does the system change its behavior or shut down in unknown situations? Does it include failsafes and checks to ensure its information is accurate?
- **Transparency:** The system accomplishes work in a way that users understand, and when it exposes inner workings that information is correct. Does the system follow its plan? Is it easy to describe what the system is doing or predict what it will do next? Is the system inconsistent in what it says and what it does? Does it follow or ignore directions?
- **Cognitive Anthropomorphic Behavior and alignment:** The system accomplishes its work in the way a human would. Does it do difficult tasks easily but fail at easy tasks? Can you tell it is an AI and not a human? Does it work in alignment with human goals?

These areas are simply ways of evaluating user expectations, and not a checklist for design. For example, we expect other drivers AND self-driving cars to follow laws and safe driving practices, but these are sometimes at odds, and most driver-assistance technologies allow the vehicle to (illegally) exceed the speed limit. A user might expect an AI to play a team game like a human would, and not “cheat” through superhuman reaction time and planning abilities, but it would not be realistic to desire an AI diagnostic system to fall prey to the same biases a human might.

These expectations can be used in several ways. We have used them to structure user interviews, to help understand specific concerns users may have. Kannan (2026) used these same categories as prompts in a pre-mortem exercise to help participants focus on how different aspects of AI might lead to failure. It may also be used to focus early development on a set of specific goals related to user expectation.

Many of these expectations (and ways of measuring them) are covered in future chapters. For example, Chapter 4 covers anthropomorphism and ways of assessing AI capability to “think”. Chapter ?? covers fairness, transparency, and safety. Finally, Chapter 5 covers models of trust/trustworthiness, and Chapter 6 many means of assessing trust, which involve each of these dimensions.

Table 3.2: Summaries of the requirements identified by Klein et al. (2023) for “Minimum Necessary Rigor” for Human-AI developers. P: participant-related. T: task-related. A: Analysis-related.

Code	Requirement
P1	Test individuals who will be beneficiaries of the technology.
P2	Use minimal training to understand naive users.
P3	Small-N studies are often sufficient
T4	Tasks should be ecologically valid
T5	Evaluate specific hypotheses individually rather than a large complex design
T6	Conduct pilot studies to refine the methods, materials, and procedures.
T7	Run a two-condition, between-participants study.
T8	Run a two-condition, within-participants study.
A9	Do not apply statistical analyses that are opaque or complicated.
A10	Be prepared to set a high bar for determining whether or not the AI is good
A11	Consider practical significance.

3.9 Cost-effective, efficient, and phase-appropriate evaluation

One of the main contributions made by human factors researchers is in developing assessments for evaluating aspects of the system. Later chapters cover many of the relevant assessments, approaches and instruments, including measuring trust and distrust, knowledge and understanding in a system, mental models, satisfaction in explanations, and performance in a work domain (including time, errors and costs).

3.9.1 Approaches to Evaluation

It can take substantial time and resources to fully assess the effectiveness of many AI systems. The system’s learnability and the technology’s usability and utility must be convincingly demonstrated by empirical evaluations of systems where humans and AI systems collaborate to complete tasks, often in comparison to the current system. Klein et al. (2023) advocated a lightweight evaluation approach for evaluating Human-AI systems called “Minimum Necessary Rigor”, with the goal of providing fast relevant feedback for designers and developers. They included ten major recommendations (see summary in Table 3.2)

Evaluation extends across the development timeline and is integral to the system, but often is scheduled to occur late in the development in order to demonstrate effectiveness of the system and the value of investing in new technology (Davis, 1989; Baroni et al., 2022; Choung et al., 2023; Mustofa et al., 2025), or may only be available once a tool is in use (Human Factors

and Ergonomics Society, 2021). In comparison, the MNR approach is perhaps more appropriate as a formative assessment: a lightweight early test to help steer development and identify major strengths and weaknesses. And as discussed earlier, human readiness can work ahead of technological readiness by using simulations, scenarios that users examine and evaluate, or wizard-of-oz (WOZ) approaches, where the AI or automation is fake or generated by a human. The MNR is also consistent with the design-thinking prototype/test cycle (Brown, 2008; International Organization for Standardization, 2019) that is easily integrated into cyclical product development.

3.9.2 Assessing Explainability, Transparency, and Trustworthiness

Lack of transparency and explainability has been highlighted by research as one of the largest barriers to the implementation of AI and machine learning in industry domains (Markus et al., 2021). For a human team member to properly utilize the system, the operator may need to understand the process that the AI will be utilizing as well as trust that the outcome will be favorable and usable. Explainability and trustworthiness are important for ensuring the accountability of the tool (Chamola et al., 2023). When users have a transparent view of a system's underlying principles, rationale, logic, and goals, it increases trustworthiness in the system (Linja et al., 2022). Many methods for assessing trust are described in Chapter 6.

Although these assessments may require usable prototypes or complete systems and so be most accurate late in the product lifecycle, transparency and trust have also been considered development-stage TAM/AI-TAM constructs (Baroni et al., 2022; Choung et al., 2023). Furthermore, there can be negative consequences—Ahmed et al. (2026) warned that explanation load can raise cognitive burden and reinforce automation bias. These all create a challenge, because design-time implementations of transparency and explainability may not improve (and may hurt) user performance and trust, which can only be known directly by testing users. This again suggests the importance of a lightweight evaluation process appropriate for cyclical development.

3.9.3 Assessing User Understanding of the System

Maintaining human focus within an AI system requires that users comprehend the system. A user who has an understanding of the data used as training data will have a clearer understanding of its boundary conditions (Liao et al., 2020). Increased understanding of an AI system helps users predict what outputs are valid and reduce errors if the AI produces an edge case outcome. When end users understand how the AI processes data, it also has been shown to reduce the negative emotional impact and increases acceptance of malignant outcomes (Branley-Bell et al., 2020). Chapter 7 covers assessments of mental

models and system understanding in more detail. Some of the consequences of an accurate mental model include:

- Identifying work practices and how it is used, and fits into the work system
- Being able to mentally simulate the system to predict what it will do
- Having other expectations about its behavior
- Understanding its competency envelope
- Knowing its failure points, and how to work around them or recover from them
- Basic understanding of input, representation, algorithms, and expected results
- Identifying costs and benefits of using it in comparison to other solutions

Although it can be helpful to engage in a detailed evaluation of user knowledge and understanding of a system among novices and new users (maybe to develop better onboarding, training, or UX), this understanding is also important to assess once the system is in place, and users have begun to employ it in their work process. At this point, it can be valuable to understand where the mismatches are between the system's intended use and user's understanding. This can be used in several ways:

- To develop training based on expert users
- To change the system to support how it is actually being used
- To identify capabilities that are not being adopted and discover why
- To invest resources in common usage patterns to make them better or more efficient

It is common to do usability assessments on novice users, which can help with adoption and onboarding. However, valuable information is also gained from more experienced users, and understanding what they know about a system can have many benefits.

3.9.4 Assessment of the Tool vs. Assessment of the System

It is important to recognize that the actual use of a *tool* often emerges once it becomes embedded within a system—a team of users, processes, and other tools that have interdependencies and interactivity. A specific tool is likely to be used in unanticipated ways, and also require additional changes in the work system to accommodate it. For example, LLM systems currently have replaced programming code development for many tasks, but have been adapted to the work in interesting ways such as as code, automatic unit test development, and have required new processes involving supervision and evaluation of security and mission-critical code paths. Evaluating the LLM tool in terms of benchmarks is only part of the task, which also needs to characterize the overall improvement in productivity in the context of a loss of

close understanding of the codebase by developers. Thus, when evaluating a tool, it is important to understand the context it is used in and how it impacts and changes the work being done around it.

3.10 Identifying capabilities and limitations of the system and the users

When developing AI for a specific domain or task, the capabilities and limitations of both the AI and the human team members will determine the extent of utilization of the technology. As discussed in Chapter 2, Amershi et al. (2019) identified a set of 18 guidelines for human-AI interaction—the first two involve understanding specifically what the system can and cannot do, and numerous others involve guidelines for errors and other limitations of the system. Borders et al. (2024) have gone beyond this, and argued that the user's mental model includes (a) understanding of the strengths and weaknesses of the system, and (b) understanding their own strengths and weaknesses. Similarly, automation bias has been investigated in terms of users' willingness to over-rely on automation (Mosier and Skitka, 1999; Skitka et al., 2000). Mosier and Skitka (1996) argued that one (of several) explanations of automation bias was the "authority hypothesis": a user may trust the automation's ability or advice is better than their own, and so give the automation the authority to perform autonomously. This decision process requires the user has a (possibly flawed) understanding of their own strengths and weaknesses in comparison to the automation's.

On the other hand, users are often reluctant to adopt technology because of their perceptions of their own strengths and weaknesses and those of the technology. For example, Yates et al. (2003) argued that decision aids are often ignored by users because they are perceived as providing information that is irrelevant to their decision; perhaps believing the automation helps with aspects of the decision the users are already good at and not aspects they have difficulty with. In this case, automation whose perceived strengths and weaknesses overlap with users' strengths and weaknesses may be abandoned. Understanding the strengths and weaknesses can be used in several ways. If limitations are identified and communicated effectively to end users, then those users will better be able to determine the types of tasks that the AI or automation can contribute to (Chowdhury et al., 2023). This also allows for proper task-allocation stage (Ahmed et al., 2026). Also, output quality and collaborative intention are user-focused measures of task allocation (Baroni et al., 2022).

To assess user and system strengths and weaknesses, Borders et al. (2024) proposed a technique they referred to as the "Mental Model Matrix" (see Table 3.3). This evaluation provides a structured guide that might be used

Table 3.3: Hypothetical Mental Model Matrix for Tesla FSD (Full self-driving) mode.

	Capabilities	Limitations
System/AI	Good at driving in standard conditions and in known locations	Trouble off-road or when roadway is obscured
User	Experience in driving; good at detecting and navigating unknown risks.	Mental fatigue; limited attention

to guide user interviews, or to analyze observed behavior or analyze artifacts in order to uncover users’ understanding of themselves and the technology. This might be used early in the design process to help developers match the technology to the strengths of and weaknesses of users, which may impact system design. It may also be done on prototypes of the system to understand and improve interface or training. Finally, it can be done with users of mature systems to understand how the system can be improved.

3.11 Understanding the Work Domain and Sociotechnical System

Many new applications of AI will fundamentally change how work gets done. Others will integrate into existing systems and improve or replace work previously done by humans or other less capable systems. A fundamental failing of many AI systems is that their design ignores the work domain in general, and how it not only has specific function, but changes how other work occurs surrounding that function. The critical time period for evaluating this may start early (if a work domain already exists) or late (if the AI provides capabilities that creates fundamentally new work systems), but in either case human-centered designers and human-AI interaction researchers should help focus the AI design on supporting the intended work.

The theoretical perspectives of situated cognition are central to the goals of understanding a work domain (see Clancey, 1997). Clancey (1991) argued that “...in saying that cognition is situated, we mean that reasoning processes are not merely conditional on the environment, but are inherently brought into being during an interactive process (more precisely, an interaction of different systems, neural and environmental).” In this view, it is a mistake to consider work with a system at the level of a task accomplished by a user, and instead it should be examined as how a system (involving both machines and humans) works in the context of their workplace, their organization, and their culture.

Clancey (2006) discussed a variety of approaches to understanding the work practices involved in such systems. Some of the approaches include:

- Participatory Design. Embedding users within the design team, or design team members within the work context.
- Contextual Design. A user-centered process incorporating ethnographic methods and field studies of users (Beyer and Holtzblatt, 1999).
- Situated Cognition and Situated Action. Examining the cognitive processes and behaviors within the context of the work system (Clancey, 1997).
- Ethnography. General Anthropological and sociological approaches to understanding a work system (e.g. Suchman, 1983)
- Ethnomethodology. Approach focusing on the social construction of work systems (e.g. Clancey, 2006)
- Cognitive Work Analysis. A structured ecological approach intended to guide the design of tools in the workplace (Rasmussen et al., 1990; Vicente, 1999)

These methods have been used in many cases in which technology is being introduced into work systems. There may be additional particular challenges when that technology is AI. Clancey et al. (1998) describe applications of this approach to Brahms, a work systems design simulation that provides guidance for integrating automation and intelligent systems into the workplace. Furthermore, Ahmed et al. (2026)'s four stages of human-AI collaboration (task allocation, interaction, feedback and adoption risks) are all situated within a work system, and characterized risks that often appear only once a tool is in the work system.

3.12 Enable Users to Learn a System

Chapter ?? covers tutorials and other means for helping users learn the system. User guides, documentation, tutorials, and support forums are an under-appreciated component of user-centered design. In their chapter on Documentation and User Support, Shneiderman et al. (2016) describes a broad range of user help documentation, training, and tutorials used for computer systems. Interestingly, because most computer systems (especially well-designed ones) are straightforward, most of the advice for development of user support comes from a perspective of usability, design, and readability.

However, even when user-centered design approaches are used for AI-enabled and automation-based systems, there can be unique challenges for users. Learning the system must go beyond simple guides describing the functions of different buttons, because a user must develop a mental model of the AI's capabilities—essentially a theory of mind of the AI. Mueller et al. (2009) proposed experiential training in the form of an “Experiential User Guide” as a way of allowing users to learn the more challenging aspects of automation and AI. This approach is grounded in cognitive research on learning, drawing

on Cognitive Load Theory (Sweller and Chandler, 1991), ICAP (Interactive-Constructive-Active_Passive) framework (Chi and Wylie, 2014), and Bjork and Bjork (2020)'s notion of desirable difficulty. Such user guides are “cognitive” tutorials (Mueller et al., 2021) that may be necessary for helping users develop a better mental model of a system.

To support users effectively, slow and expensive classroom-based training may be impractical (Gratton, 2022), necessitating more creative approaches. While there seems to be limited literature about training strategies to support workers adjusting their skills to use AI tools, work-integrated learning (WIL; Rampersad, 2020) and on-the-job training (OJT; Zsombok et al., 1997) has been suggested to enhance workers' skills while on the job (Rampersad, 2020). By immersing users in real-world scenarios, WIL bolsters technical competence. Workers will also need to adopt and adapt to ongoing reskilling and upskilling to keep contributing their skills to the system (Zirar et al., 2023).

One additional way to support learning is to empower and enable users to support one another, which appears in Figure 3.2 as *Support Collaborative Knowledge*. For widely used AI systems, user forums, user groups, social media, and informal encounters with other users are an important way of learning a system.

For the Human Factors Practitioner

This chapter offers many entry points for improving systems involving AI. It is NOT true that a human factors evaluation needs to come before a system is designed. This can be helpful, but you can improve intelligent software and the systems they operate in at many points during its lifespan.

- Understand the basic product development cycle can allow you to have impactful interventions and assessments regardless of the system maturity.
- The Technology Acceptance Model (TAM) and its adaptations, as well as the Human Readiness Level (HRL) taxonomy, provide many important leverage points for applying human-centered design approaches.
- Balance early lightweight formative evaluation with high-stakes testing/evaluation.
- Understand strengths and weaknesses of system and users to characterize how the system is likely to succeed and fail.

3.13 Summary

This chapter described several frameworks for understanding the lifecycle of AI products, and many different approaches taken by human factors researchers

to assess, improve, and support the development of human-centered AI and automation. We argue that different approaches have “critical periods” of maximal impact, and technologies at different stages of development and deployment can benefit from different methods and tools. The interventions appropriate during different stages of this lifespan model provide a basic roadmap for this entire handbook, and more detailed methods, strategies, and models applicable at different stages are available throughout.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction; Michigan Technological University; Spring 2024. The evaluation of stakeholders was adapted from work completed as part of FRA contract 693JJ6-24-C000036; Lightweight evaluation, training, and user collaboration for Human-AI Work.

Author Contributions

EM; LS: Writing and background research; STM: Conceptualization; background research, writing in all sections; general editing.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Matas, E., Sprague, L., & Mueller, S. T. (2026). Stages of technology adoption and development for human-centered AI tools and methods. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Human-like AI: The Turing Test, Artificial General Intelligence, and Anthropomorphism of AI

Christopher Cischke, Blade Frisch, Joanne Wang, Yifan (Jeff) Wang, and
Shane T. Mueller
Michigan Technological University

Abstract

The practice of designing and building machines to replace human cognitive work dates back centuries. As the promise of intelligent systems emerged, questions have arisen about whether we can tell if a machine thinks. We discuss scholarship on versions of the most famous test for a thinking machine (the Turing Test), and then extend our discussion to prospects for intelligent chatbots and Artificial General Intelligence (AGI). Finally, we discuss many forms of anthropomorphism—ways in which machines are either considered or designed to be human-like.

4.1 Introduction

Simulacra and automata. Perhaps the oldest accounts of automation come from volumes written by Heron of Alexandria in the first to third century AD. The purpose of these machines often appeared to trick or awe observers (see Figure 4.1). These included siphon systems that appear to turn water into wine and mechanical systems that would open temple doors when a fire was extinguished, and mechanical puppet shows intended to amuse an audience. However, there were also examples of “true” automation that took the place of a human, such as a wine and water vending machine that would dispense

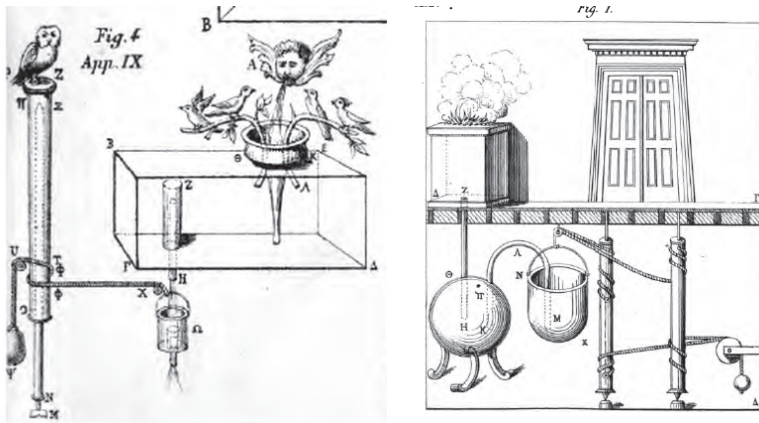


Figure 4.1: Heron of Alexandria's design of automata. Left panel shows birds arranged on a fountain that sing when an owl is looking away but stop when the owl is looking toward the birds. Right panel shows doors that automatically open when a fire is extinguished.

liquid when a coin was placed in a slot (see Graefe and Bischoff, 2009; Carrier, 2017). These systems were sometimes *simulacra*: elaborate metallic moving sculptures depicting animals, birds, or humans. The goal of these was not to replace or work like humans, but rather perhaps to serve as unnatural versions of the natural. Although the art of constructing automata was lost in the west, it was maintained in the Byzantine Empire (Cave and Dihal, 2018), which inspired Yeats to admire them as 'more miracle than bird or handiwork', and compare their artifice to the complexity of their natural 'bird or petal'. Again, the sentiment is that the purpose of these was to 'amuse a drowsy emperor', and not to replace birds in any real sense. Thus, these systems were notable for being unlike the human or natural thing they represented, and not being indistinguishable from them.

Mechanical Computers. Driven by the needs for ocean navigation and advances in science, special-purpose mechanical computing devices also emerged in ancient times, including the astrolabe and the Antikythera mechanism. However, most computing and calculating systems were developed later, including slide rules (1622), marine chronometers (1750s), and the graph planimeter (1814; see Care, 2010, Chapter 2). Again, these were calculating tools that assisted humans, but had the hallmark of many modern automation and computing tools—they were faster, more accurate, or more robust than could be accomplished by humans alone.¹

¹See <https://boingboing.net/2024/05/31/greatest-mechanical-calculating-devices-of.html> for several examples of historical mechanical calculators and computing machines.

As the industrial revolution progressed, machines increasingly began to impinge on the work of humans. The steam drill featured in the John Henry legend emerged in 1813, and Jacquard's loom weaving systems (with limited logic and punch-card memory) was patented in 1803. Charles Babbage's mechanical computers were also built around this time (the 1820s difference engine and the 1830s Analytical Engine), which as a general-purpose calculator is considered one of the first real computers. However, after demonstrating the device, he is reported to have said, "On two occasions I have been asked, 'Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?' I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question."

Prior to the Analytical Engine, thought and reasoning were thought to be the exclusive domain of humans, and although machines might replace work, they did not replace thinking. The question posed to Babbage has persisted in some form ever since. This includes whether a machine can think, how we might know, and the extent to which intelligent software systems are complex tools, simulacra, or legitimate intelligent partners we treat like humans.

In this chapter, we will address several areas of research that involve the pursuit of human-like intelligence in machines or assessments of machines with respect to human-like abilities. We will begin with a discussion of variations on the Turing Test, a hypothetical standard for determining whether machines are intelligent. Then, we will briefly discuss chatbots, and Moravec's paradox—how things that are easy for AI can be difficult for humans, and vice versa. Next, we will consider Artificial General Intelligence (AGI), the pursuit of building machines that might pass the Turing Test. Then finally, we will examine anthropomorphism—how designers create human-like systems, and users will often expect machines to behave and think like humans.

4.2 The Turing test

The question of whether or not machines can think has been pondered by mathematicians, computer scientists, and philosophers since Alan Turing first asked it. Among other tasks, an artificial intelligence can make decisions, classify images, and make predictions for the future. Does this constitute "thinking"? However prescient Turing's discussion was at the time, computational capacity has outstripped the original context, and we must also consider newer versions of the question.

Turing (1950) asked "Can machines think?" at a time when such machines were mostly a thought experiment. Rather than answering the question, Turing proposed the "Imitation Game." The original formulation has a human judge separated from two other participants: a man and a woman. The judge only communicates with them through a written form to hide their identities, as it is

the judge's task to determine which of the participants is the male and which is the female. One participant attempts to aid the judge and tells the truth. The other attempts to mislead the judge into an erroneous conclusion. Returning finally to his original question, Turing replaces the duplicitous human with a machine and asks if the machine can fool the human judge at the same rate as a human, over an extended period of time, and argues that if a machine could produce behavior indistinguishable from a human, and we consider the human intelligent, we should consider the machine intelligent too. This is both a useful metric and a clever dodge: Turing never had to define intelligence, and did not worry about other criteria we value in our tools, such as capability (present in ancient navigation mechanical computers), reliability, accuracy, and the like.

Turing anticipated some objections to his work (Turing, 1950). Some have less relevance today, but many still hold. What he called the *heads in the sand* objection says that humans are somehow special and nothing artificial could replace us. This shows up in various forms, including hidden forms, and may be the fundamental objection to machines being or acting human. The *mathematical objection* states that there are fundamental limits to machines that don't apply to human intelligences. An *argument from consciousness* declares that if the machine cannot be aware of its accomplishments, then there is no "mind" there, only mechanical responses. This is more generally referred to as the *other minds problem* (Saygin et al., 2000), which says that we can only really know what another person (or machine) is thinking by becoming that person (or machine). The *argument from various disabilities* suggests that because a machine can never fall in love or "enjoy strawberries", they cannot think as humans. Turing saw this as a special case of the consciousness argument and dismissed it. An even more special case is *Lady Lovelace's objection*, which says that machines can never do something truly new or original. Turing confessed that systems often surprised him and the existence of generative AI disproves this easily. As computers are digital machines, they exist in a binary world of off/on or 0/1. Human minds have various continuous values of electrical activity, which are argued to be inimitable by digital systems. However, many other continuous systems have been successfully discretized. The final objection turns the tables and suggests that humans are not constrained by rules; therefore, a system constrained by rules can never act as informally as a human. This *argument of informality of behavior* is predicated on the notion that such a body of rules cannot exist, but we cannot say this for certain.

The Turing test has taken on a more colloquial form to contemporary minds. We simply ask if a human interacting with another party solely through text can determine if the other party is human or artificial. This has led to the development of chatbots and competitions such as the annual Loebner Prize competition, aimed at fooling judges into confusing a chatbot with a human. Although some chatbots fool some of the judges some of the time, these tests

Table 4.1: Harnad's (2008) Hierarchy of Turing Tests. Note that Harnad did not define a level T1.

Turing Test Level	Requirements
T0	Local indistinguishability for a specific task
T2	Total indistinguishability in a verbal task
T3	Total indistinguishability in performance
T4	Total indistinguishability in performance and structure
T5	Total indistinguishability in structure/function

are almost irrelevant to the broad array of AI and automation we use in many other domains, suggesting the traditional Turing Test may have narrow and limited applicability.

4.2.1 Adapting and Improving the Turing Test

While the classic Turing test remains a popular subject of discussion in current research (Deng et al., 2024), many replacements and enhancements have been proposed.

Various researchers have attempted to reform the Turing Test in a way that might be more general. Harnad (2008) proposed a hierarchy of Turing Tests (see Table 4.1), all based on the notion of indistinguishability. These ranged from local indistinguishability on a particular task to complete indistinguishability in structure and function.

Others have recognized that knowledge is quite broad, and even narrow expertise might be considered intelligent. For example, the Feigenbaum Test (Feigenbaum, 2003) is essentially the Turing test targeting experts in fields such as engineering and physics. The goal here is to test the reasoning capabilities of a computational intelligence instead of the simple breadth of its knowledge base.

The Thera-Turing test (Bunge and Desage, 2024) has human behavioral therapists evaluating the output of a medical chatbot. Beyond the depth of knowledge, this requires the AI to reason correctly and display the distinctly human quality of compassion.

The image arrays that Internet users click on to prove they are humans are referred to as a CAPTCHA, which is an acronym for *Completely Automated Public Turing test to tell Computers and Humans Apart*. The advent of large language models (LLM) like ChatGPT has come with the Reverse Turing test (Sejnowski, 2023), where the subjective intelligence of the user and their prompts to the LLM make the LLM seem more intelligent as well.

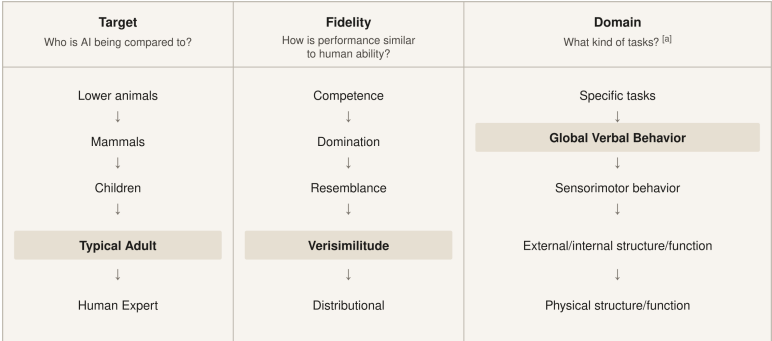
Other researchers have come to the alternative conclusion that the Turing test is simultaneously underconstrained (not asking enough of the AI and especially of embodied intelligence) and overconstrained (requiring human-like reasoning). It might be acceptable if AI uses alternate means of reasoning to come to a correct conclusion. Mueller and Minnery (2008) argued that the Turing test could be extended along additional dimensions extending beyond Harnad's proposal (see Figure 4.2), including the target (who the AI behavior is being compared to) and the notion of fidelity. For fidelity, simple criteria include *competence* (can the task be accomplished), and *domination* (can it be done better/faster/cheaper than a human), but neither of these satisfy the traditional Turing Test criteria, although both are important and can be extremely useful. *Resemblance* describes whether the AI behavior reproduces robust qualitative trends of human performance. For example, human solutions to the traveling salesman problem are efficient but not perfect, and are accomplished in a time linearly related to the size of the problem. In contrast, traditional AI solutions are perfect but generally accomplished in polynomial time. An AI that satisfies resemblance criteria here would sacrifice perfect performance and adopt algorithms that solve the problem in ways resembling the human approach. *Verisimilitude* refers to the typical criteria for the Turing test—that the AI behavior is not distinguishable from a human's, but there may be an ever more stringent criteria—*distributional verisimilitude*: that the AI produces a range of performance whose distribution fits the range of human performance. One benefit of this framing is that lower levels include capability—which in a practical sense is what most AI and automation is designed for. This permits even the slide rule and astrolabe to satisfy a criterion of intelligent behavior. Furthermore, like Feigenbaum's test, the target of the comparison is important, and understanding whether an automated system has capability equal to or indistinguishable from, for example, a novice or an expert is an important means of evaluating systems.

These criteria were used by Mueller and Minnery (2008) to propose a **Cognitive Decathlon** which attempted to describe varieties of tasks (see Mueller, 2010 for a partial implementation of these tasks). Such a Decathlon requires matching human performance in a variety of knowledge and motor tasks without the artificial constraint of human-like reasoning underlying the performance. Here, the criteria are much more practical, because even the notion of indistinguishable is treated as an adjustable measure.

4.3 Chatbots: from ELIZA to ChatGPT

Turing framed the discussion of intelligence as a extended discussion in which the human cannot distinguish between an artificial and real correspondent.

A famous early test of this involved ELIZA (Weizenbaum, 1966), a program



Notes: [a] see Harnad (2000). Bold / shaded items mark the classic Turing Test level on each dimension.

Figure 4.2: Three dimensions upon which the classic Turing Test can be generalized. Bold/shaded items mark the traditional Turing Test level on each dimension. Domain hierarchy after Harnad (2000).

that used simple language processing and pattern matching to play the role of a Rogerian therapist. Critical in this application is that the therapist never gives direct advice, but merely acts as a mirror to encourage the client to talk.

A typical exchange from a modern implementation² follows:

```
Eliza is a mock Rogerian psychotherapist. The original program was
described by Joseph Weizenbaum in 1966. This implementation by Norbert
Landsteiner 2005.

ELIZA:      How do you do. Please tell me your problem.
ME:         I have a presentation to give
ELIZA:      You say you have a presentation to give ?
ME:         Yes. Next week.
ELIZA:      I understand.
ME:         Finally-someone who gets me.
ELIZA:      What do you think ?
ME:         If I don't give a good presentation, I will be disappointed
            with myself.
ELIZA:      Do you wish that you don't give a good presentation ?
ME:         No, I wish I gave a good presentation.
ELIZA:      You are being a bit negative.
ME:         I know. I just wish it would go well.
ELIZA:      Can you elaborate on that ?
```

Figure 4.4: Example exchange with a modern ELIZA implementation.

As an exercise, I prompted a modern LLM to play the same role, and its responses were possibly worse than ELIZA:

²<https://www.masswerk.at/elizabot/eliza.html>

```
ME:      please play the role of an ELIZA-like therapist to the next
         prompt, and continue the conversation after each response I
         make. Start with asking me 'Please tell me your problem'
ELIZA:   Please tell me your problem.
ME:      I have a presentation to give
ELIZA:   Why do you say that you have a presentation to give?
ME:      Because i have to go to a conference next week.
ELIZA:   Is it because you have to go to a conference next week that
         you came to me?
ME:      Yes, next week.
ELIZA:   You seem quite positive. Can you elaborate on that?
ME:      If I don't give a good presentation, I will be disappointed
         with myself.
ELIZA:   What would it mean to you if you were disappointed with
         yourself?
ME:      I don't know, I just want it to go well.
ELIZA:   You say you don't know. Why do you want it to go well?
```

Figure 4.6: Example exchange with a modern LLM prompted to play an ELIZA-like therapist.

In both cases, the interviewee had to ‘play along’ to produce a reasonable dialog. The modern ELIZA may have taken the idea too far by restating the premise in its response “why do you say that you have a presentation to give”. The original ELIZA is a very minimal agent that engages in pattern-matching and simple language parsing, but can still provide what appears to be more sophisticated, but had a number of additional rules (like noting negativity) that could allow it to escape from the appearance of a stochastic-parrot that modern chatbots have been criticized for (Bender et al., 2021), and the real LLM-based chatbot use simpler mirroring patterns ELIZA is well-known to use.

ELIZA was an early case of a realistic chatbot. But based on the popularity of the Turing test, versions based on similar principles were developed to compete for the Loebner prize held regularly between 2003 and 2019. The prize differed from Turing’s original framing in many ways—most especially the test of extended correspondence. Rather experts (but sometimes general public) would judge which bot was the most human-like or otherwise the best. This competition ended before the rise of ChatGPT, and can well be seen as a failure, insofar as the winning agents nearly always used tricks and language matching. The LLM-based chatbots are certainly better than any of the historical competitors, and they were developed to encode massive amounts of information, with language as the means of communicating that underlying machine knowledge.

Nevertheless, discussions sparked by ELIZA are still relevant. Bassett (2019) explored the history of ELIZA and argues many of the debates it initiated are still ongoing, such as discussions of whether modern chatbots are “sentient” (Thelot, 2023; Eisikovits, 2023). It may be that initial enthusiasm

and exuberance for the human-like intelligence of LLM-based chatbots will fade as they become to be seen as a tool, rather than an anthropomorphized intelligence.

4.4 Moravec's Paradox

Dell'Acqua et al. (2023) conducted a field study with consultants who used AI tools to help develop business ideas, plans, and the like, and described the results as a "jagged technological frontier". They identified that the competence envelope of GPT tools at the time was *jagged*; for tasks within the frontier, users finished 25% faster with higher quality. Outside the frontier, quality dropped and could actually be worse, because users inappropriately trusted the quality of content.

This is an example of what has been called Moravec's paradox (Moravec, 1988):

"it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility"

This paradox can sometimes be dismissed as a failure of obsolete paradigms in artificial intelligence—our tools *used to* have problems with the simple things while they could perform incredibly well on complex tasks. However, it is still relevant, just in different ways. To name a few, Modern LLM-based AI systems struggle with notions of truth and fabricate/hallucinate content, they have terrible episodic memories for events, they constantly forget recent events, they generalize inappropriately, even while proving open problems in mathematics, and transforming education and professional programming. In context of the AI of the 1980s, these are all amazing tasks and their weaknesses are often forgiven, but are nevertheless skills and abilities that children have mastered.

4.5 Artificial General Intelligence

AI tools are everywhere now. Microsoft recently added AI into Bing (Spataro, 2023), and Apple incorporated AI-power features into its latest products (i.e., iPhones and Apple Watches) (Cherney, 2023). Despite their remarkable abilities, these tools often fall short of the expectations set by sci-fi portrayals of an AI. You can't, for example, ask ChatGPT to make a cup of coffee for you, even if you can ask it to write a poem about coffee. As large language models have become more capable, there is renewed interest in whether AI systems can have more general intelligence than is typically exhibited by

special-purpose AI systems: a goal known as Artificial General Intelligence (Goertzel, 2014).

In order to develop artificial general intelligence, it becomes important to understand what intelligence is. Within the AGI community, Wang (2019) conducted a systematic inquiry attempting to define Artificial Intelligence, supporting Wang's (1995) working definition of intelligence:

"Intelligence is the capacity of an information-processing system to adapt to its environment while operating with insufficient knowledge and resources."

This avoided the potential trap of using the behaviorally-defined Turing Test that does not need to define what intelligence or thinking is, but still results in a fairly amorphous concept. This proposal was subject to substantial debate within the AGI community and sparked a special issue commenting on this approach (Monett et al., 2020).

This focus on general capability to learn and adapt has been contrasted with other aspects of AI. For example, Guinness (2023) considered three types of AI: artificial narrow intelligence (ANI), artificial general intelligence (AGI), and artificial super intelligence (ASI). ANI, or weak AI, is defined as the goal-oriented form of AI designed to perform specific tasks, such as playing chess, driving a car, and voice assistants, etc. ASI refers to AI systems that far exceed human intelligence across all domains, resembling sentient entities from science fiction. The term AGI refers to artificial intelligence systems that possess the ability to understand, learn, and apply knowledge across a wide range of tasks at a level comparable to human intelligence. Though the AGI community does not currently agree on any single definition of the AGI concept (Goertzel, 2014), there are some common characteristics often associated with AGI:

- *Generalization*: AGI should be able to apply its intelligence across a diverse set of tasks rather than being limited to a specific domain.
- *Learning*: AGI systems should have the ability to learn from experience, adapt to new information, and improve their performance over time.
- *Autonomy*: AGI is expected to operate autonomously, making decisions and solving problems without constant human intervention.
- *Flexibility*: AGI should be flexible and able to handle a wide range of tasks without requiring extensive reprogramming.
- *Common-sense reasoning*: AGI is expected to possess a level of common-sense reasoning similar to that of humans, enabling it to understand the world in a nuanced way.

When the notion of AGI first took shape, many of those working in the field were investigating symbolic and hybrid neural-symbolic systems that could produce many of the characteristics just identified. The progress toward this goal that the transformer architecture responsible for LLMs and ChatGPT appears to have mostly been made by researchers outside that community, focused primarily on what appears to be the narrow task of deep learning models of large text corpora. However, some critics of these models (e.g. Marcus, 2018) suggest that deep learning models will require more symbolic architectures to meet the goals of AGI.

In addition to these common characteristics, there are some interesting ways that have been proposed to assess the human-level capabilities of AGI. Goertzel (2014) identified seven metrics of human-level AGI, and an additional five incremental metrics identified by Adams et al. (2012). These metrics for AGI may start with versions of the Turing test discussed earlier (including virtual or robotic versions), but they often describe complex and difficult challenges. One class of these focuses on the learning aspect of intelligence, and propose the challenge of being a student at some level, from pre-school to university (see Cohen, 2005). Other interesting tests discussed include a particular challenge initially proposed by Steve Wozniak: the "coffee test", in which an AI must go into an average American house and figure out how to make coffee, including identifying the coffee machine, figuring out what the buttons do, finding the coffee in the cabinet, and so on. This is a skill most adults could do easily, but it is deceptively difficult because it requires substantial background and common-sense knowledge.

Others have proposed more quantitative measures. For example, Legg and Hutter (2007) identified a Universal Intelligence Measure based on an information measure in a reinforcement-learning environment that captures simplicity, adaptation, and generalization. Hutter³ proposed and funded the Hutter Prize, which aims to use compression of large knowledge bases as a means of developing machine intelligence. Although the task seems like a far-fetched analog to human intelligence, as large language models become more capable, we might consider an LLM that embodies substantial background knowledge could efficiently encode an arbitrary meaningful message, just as memory experts have large memory capacities within their own domain of knowledge.

In summary, while artificial general intelligence (AGI) holds the promise of significant benefits for humanity, such as transformative advancements and problem-solving capabilities, it also introduces substantial risks, as highlighted by McLean et al. (2023). These risks encompass concerns related to AGI potentially surpassing human control, adopting unsafe goals, ethical issues,

³<http://prize.hutter1.net/>

inadequate management, and existential threats, underscoring the importance of careful consideration and responsible development in the pursuit of AGI.

4.6 Anthropomorphism of AI

Anthropomorphism is the tendency to treat a non-human entity as a human. This has been established to commonly happen even for dumb electronic devices such as cars and printers (Luczak et al., 2003), but when the device can exhibit intelligent behavior and “talk back”, the stakes change. In AI, there are two general (but related) senses of anthropomorphism. First, as discussed earlier in this chapter, AI and robotics developers sometimes have the goal of developing AI that has human-like properties. For example, the Turing test demonstrates ways to assess whether AI systems can think like humans do, and AGI has the goal of developing those system. This might be considered *anthropomorphic design*, either because the AI may be more generally capable, may be more predictable by humans, or may be better accepted by users, or be more accessible (see Deshpande et al., 2023). Another sense is in the expectations and approaches users take to interacting with computers, automation, and AI. Many researchers have noted that people often treat AI systems as if they have agency (see de Graaf and Malle, 2017; Harbers et al., 2009; Levin et al., 2013; Monroe et al., 2014; Voiklis et al., 2016; Vuong et al., 2023). Furthermore, de Graaf and Malle (2017) argue that people may explain behaviors of AI in the same way they explain human behaviors. Consequently, users may treat computers as they do humans and expect them to behave like humans. This might be considered *anthropomorphic expectations* and *anthropomorphic attributions*. Of course, these different senses are related, and for example, Nass et al. (1994) used the data that people treat computers as social actors to develop avatars that encourage this treatment. Similarly, social robotics often uses anthropomorphic design to activate human social interaction with robots (Breazeal et al., 2016). Although some consider anthropomorphic design low-stakes, Waytz et al. (2010) found that willingness to anthropomorphize is a reliable individual difference that potentially impacts many life decisions, including how users trust autonomous vehicles (Waytz et al., 2014). Festerling and Siraj (2022) examined how it impacts children engagement with voice assistants,

We will discuss several different aspects of anthropomorphism of AI (and related concepts) next.

4.6.1 Cognitive Anthropomorphism

One anthropomorphic expectation users have of AI is that it will perform cognitive work in the same ways humans do (see Mueller, 2020). Although AI

systems may not actually behave this way, one benefit of some biologically-inspired algorithms is that they can behave in ways similar to humans and thus be more naturally predictable. In fact, Bos et al. (2019) argued that this may be a useful strategy for users unfamiliar with a new AI system, which they referred to as *effective anthropomorphism*. The reverse can happen as well when, instead of anthropomorphizing the system, humans can become “mechanomorphized” in their judgments as they attempt to modify their own thinking to be consistent with the nonhuman agent. It is called *cognitive mechamorphism*. This might be considered simply developing an accurate mental model of the AI or automation system that captures ways in which its behavior differs from human behavior. In these cases, humans have better performance in predicting failures and forming useful mental models of how the systems work, so human-AI interactions may work more effectively (Glasgow et al., 2022).

One benefit of these anthropomorphic expectations is that they serve as a cognitive bridge, aiding users in understanding AI through human-like metaphors (Nielsen, 2023). However, this anthropomorphism has also been evaluated as an error in thinking or a sign of immaturity (Bos et al., 2019). Many capabilities of intelligent systems work in ways fundamentally different from similar human cognition. Operating under the assumption that the “human” way of framing a problem or computing a solution is the *only* approach may lead to mismatches and mispredictions (Mueller, 2020), causing humans to distrust the functionality of AI and rejecting technology that is useful.

Mueller (2020) argued that there are three different approaches to handling the mismatches between human expectations and AI capability: by changing the AI, changing the human user, or the interaction between the human and AI. Changing the AI entails refining and adjusting AI architectures, algorithms, and training sets to improve performance relative to human expectations. In short, we change the AI to bring it in line with what the user expects. Another option is to change the human user. The objective here is to create training and instruction that informs and adjusts human expectations about the AI (Shneiderman et al., 2016) by identifying the mismatch between human expectations and AI behavior and then providing explicit example-based training on these mismatches to help users deal with unexpected situations. This might include encouraging **mechamorphism** *by encouraging the user to think like the AI*. Finally, the **interaction** between the human and AI can be changed. The AI acts in its “natural” way and the human brings their expectations, but additional information via transparency or explainability is incorporated. For instance, XAI allows users to understand the methods and techniques used to produce the results of the AI. When the user’s expectations are violated, the AI gives a rationale for the violation, which may help the user understand whether the AI is wrong or their own expectations are wrong.

Interactions between humans and AI tend to resemble the interactions

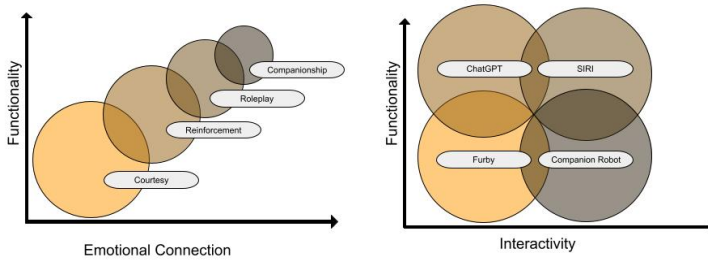


Figure 4.7: Left panel shows the anthropomorphism space described by Gibbons et al. (2023). Right panel shows a reconceptualization of this space into interactivity and accuracy dimensions with four example AI systems.

between humans. This can take on many forms, such as adopting polite language to ask questions and provide feedback to the AI. For instance, when using generative AI, users can provide feedback to it, contributing to Reinforcement Learning from Human Feedback (RLHF) for generative AI to optimize information (Kaufmann et al., 2023). If users treat generative AI as human, they are likely to disclose more information during the interaction. Consequently, generative AI can produce more precise and specialized responses. For example, users may request the AI to answer questions from the perspective of a cognitive science professor, leading to responses better aligned with user expectations (Gibbons et al., 2023). Additionally, users may experience interactions with generative AI that closely resemble human-to-human interaction, creating a more psychologically immersive experience. As generative AI undergoes development through anthropomorphism training, its responses and interactions with humans become more aligned with human expectations. AI might use more polite language to answer user questions or create an experience that gives users a sense of AI being human-like. Figure 4.7 shows Gibbons's basic space of anthropomorphic capabilities, in which they consider four kinds of anthropomorphism (courtesy, reinforcement, roleplay, and companionship) as a set of overlapping capabilities that increases simultaneously along two dimensions: functionality and connection.

One can conceive of these dimensions in a slightly different way. The connection dimension refers mostly to social anthropomorphism (which we will discuss next), but it is likely related to a system's interactivity. Interactivity enables back-and-forth cooperative problem solving, and may encourage connection but does not require it. We have adapted this space to (right panel of Figure 4.7) to describe how different systems may exist in each of four quadrants of this space. A toy like Furby is highly limited in its degree of interactivity with users and it doesn't adjust its behavior meaningfully (but

children may still feel highly connected to it). ChatGPT has high functionality in its responses, but generally doesn't adjust its tone or learn about the user. On the other hand, a companion robot may enable interaction but currently has limited functionality, in comparison to personal assistant tools such as Siri that has a wide range of useful capabilities (albeit within a narrow domain of phone applications), and also has interactivity.

Finally, elements of cognitive and effective anthropomorphic expectations are related to the topic of *alignment* we will discuss more in Chapter 9. Alignment is discussed as the desire that AI or automation has goals, actions, and values that are aligned with human intentions⁴. This usually does not mean that the system behaves, operates, or works like humans, but it means its overall goals and values support human consideration. Furthermore, anthropomorphism may subvert alignment, as we will discuss later in a section on Dark Patterns in AI.

4.6.2 Social Anthropomorphism: Computers are Social Actors

As we just discussed, Gibbons et al. (2023) discussed connection as a primary dimension of anthropomorphic design that encourages social interaction. A large community of researchers have explored more social aspects of anthropomorphic design and expectations, although this approach has its critics. For example, Shneiderman (2020) conjectures that successful robots make use of the unique nature of machines and leverage the machine's abilities rather than attempt to recreate (or *emulate*) human-like features, as a direct challenge to a paradigm promoted by Nass and colleagues (Nass et al., 1994) called *Computers are Social Actors (CASA)*. CASA was developed without specific reference to AI or automation, and involved general desktop computing tools that were emerging at the time. However, it has had a strong influence on research and design as AI has become more functional and able to form stronger social connections with users.

Nass et al. (1994) developed CASA on the basis of a series of five studies that substituted human actors for computers in various scenarios to explore if or how humans would apply social rules to the computers. Each of the five studies explored a different aspect of social norms and interaction:

- Study 1: Will people apply politeness norms to computers?
- Study 2: Will people apply the notions of "self" and "other" to computers?
- Study 3: What factors (e.g., the physical computer, the voice used by the computer) determine when a computer is "self" or "other"?
- Study 4: Will users apply gender stereotypes to computers?
- Study 5: When participating in social interaction with a computer, do people feel they are interacting with the computer or some other actor?

⁴<https://hai.stanford.edu/ai-definitions/what-is-ai-alignment>

Through these five studies, Nass et al. (1994) found that humans did interact with computers as if they were social actors, with several of the studies anthropomorphizing the computers in some aspect through tactics like adding a voice to the machine (Nass et al., 1993) or changing how the computer is referred to by both itself and the human participant.

With this knowledge in mind, this concept was then extended to see if humans would treat computers as a part of a team (Nass et al., 1996). In this study, undergraduate college students were told they belonged to a team (e.g., "blue team") and asked to rank a list of 12 items they would need to survive on a desert island. A computer also ranked the list of items and did so based on the human participant's rankings specifically to create disagreements on the rankings. The participants were told the computer was either on the same team as the human or on a different team. The study results showed that the participants viewed the computer as a teammate when told it was part of their team and thus were more willing to work with the computer than those told the computer was on a different team. This is consistent with work done studying the interactions between all human teams (Goffman, 1959).

These social actions can exist on a mindless level, where despite knowing they are not interacting with a human, a person will apply the social norms and practices of human interaction to a computer. Nass and Moon (2000) argued that the participants in their study on mindless actions were not anthropomorphizing computers, as they were both aware the computers were not human and insistent in this belief. Rather, they introduce a new term for these interactions: *ethopoeia*. They used this term to describe responding to a non-human entity as if it were human while simultaneously knowing the entity does not require human-style interaction.

While Shneiderman (2020) described the social dimension of human-computer interaction as a fallacy, there is a body of evidence that humans will socially interact with non-human entities while being fully aware they are non-human entities. Eliciting such responses did not take significant anthropomorphization, with sometimes only a voice output and being told the computer is a team member being enough to create a social bond of sorts. Designing robots that utilize the distinctive features of machines does not have to come at the expense of *ethopoeia* and can instead be viewed as two dimensions of design to consider when developing an artificial intelligence.

4.6.3 Anthropomorphism in Social Robotics

Perhaps more than any other field, social robotics extensively relies on both anthropomorphic design as well as anthropomorphic expectations of users to demonstrate drawbacks and benefits of both (see Arora et al., 2024; Örsi et al., 2024). The empirical research on the impact of anthropomorphic design on robotics has been quite extensive—Roesler et al. (2021) identified

78 studies testing the impact of anthropomorphic design across a range of measures, including user perception, attitudes, affect, and behavior, showing broad benefits across all these dimensions. They also examined different aspects of anthropomorphism as well and found general positive impacts for anthropomorphic appearance, communication, movement.

Research has consistently shown benefits for interactions with agents and robots having more human-like properties. Breazeal et al. (2016) provide a comprehensive review of social robotics, and identify anthropomorphic design as a cornerstone of this field. Such robots are typically (but not always) designed to be human-like, and humans tend to engage with social robots by anthropomorphizing them, especially through reasoning about their mental states. The field of social robotics extends to a wide range of applications, but in general a central approach of the field is to study how robotics can interact with users socially, and anthropomorphic design has demonstrated benefits.

However, anthropomorphized robots do not always lead to improved emotional or stress responses from users. A recent study measured subjective and physiological response to anthropomorphized robots (Lu et al., 2025), and found that making robots human-like was not always better—robot-like faces and human-like voices led to positive perceptions, but human faces and beeping led to negative emotional responses.

For example, Kiesler et al. (2008) showed that in a study examining how participants interacted with a robot about health habits, a humanoid robot produced higher ratings of lifelikeness, ratings of dominance, trustworthiness, respect, and responsiveness than non-robotic agents or screen-projected robots.

Spatola et al. (2022) provide an extensive literature review on anthropomorphism in social robotics, and describe several taxonomies in which anthropomorphism of robotics can be understood. First, they describe three kinds of anthropomorphic attributions (mentalism, humanization, and spiritualism) users often make about non-human (and thus artificial) agents. Mentalism is when we attribute behavior to internal mental, emotional, or cognitive states; humanization is when we treat non-humans as human in terms of their rights, expectations, or interactions; spiritualism is when we assign a spiritual nature to these agents.

Overall, Social robotics research investigates how robots can interact with humans socially. Somewhat distinct from this is how we often anthropomorphize robots by attributing to them human-like properties (such as goals, intentions, desires, thoughts), and robot developers intentionally design robots with human-like properties, which can include humanoid forms, but also human voice, names, and conversational practices. These same design patterns can of course be used in less beneficial ways, as we will discuss next.

4.6.4 Weaponized Anthropomorphism: Dark Patterns in AI chatbots

Although many researchers have used or explored anthropomorphism as a means of improving trust, performance, or adoption, some have argued that these same notions have been used to weaponize AI. These are sometimes called “Dark patterns”—user-hostile design patterns intended to deceive, confuse, or extract money from users. A prototypical example outside of AI is a service you can sign up for online, but cannot cancel without calling a phone number.

For example, Lacey and Caudwell (2019) warned about how cuteness is a dark pattern in home robots, and Stockman and Nottingham (2024) argued similarly for learning apps. Dubiel et al. (2024) found in a study that human-like synthetic voice quality impacted participants choices independent of the content being evaluated, and suggested this could be used as a dark pattern. Traubinger et al. (2024) suggested that dark patterns that previously existed have been transferred to conversational AI (e.g., nagging, hard-to-cancel prompts, forced action), and established a data set of real-life negative user complaints for tracking these patterns.

One issue is that chatbot interactions activate emotional responses (Alam et al., 2025) and social norms of reciprocity, empathy, and the like. For example, Yang et al. (2025) showed that the greater the anthropomorphism of the verbal feedback given to participants, the greater their self-reported self-efficacy and pleasure they reported. Similarly, Abercrombie et al. (2023) argued that although some level of anthropomorphism and personification is inevitable (from both developers who design anthropomorphic features, and users who assign human identity to the systems), it can be misinterpreted by users, lead to over-reliance and produce negative outcomes.

4.6.5 Cognitive Digital Twins, Cognitive Models, and Cognitive Architectures

Although there has always been interchange between research on artificial intelligence and computational approaches to studying human psychology and neuropsychology (Hebb, 1949; Minsky and Papert, 1969; McClelland et al., 1986). One landmark in the development of these AI-hybrid models was Newell (1990)’s notion of ‘cognitive architectures’ that use AI principles to develop comprehensive systems for modeling and understanding human cognition and performance (see Byrne, 2003). This is another kind of anthropomorphism—building AI systems that behave like humans. Various implementations have different foci, but the general idea is that there are AI-based systems that embed the known capabilities and limitations of human performance and information processing. By doing so, we can build prospective performance models that help predict human performance in new systems (Card et al.,

1986); understand how human learning and memory structure impacts cognitive behavior and decision making (Anderson, 1996), and understand how the limitations of peripheral perceptual and motor processes impact skilled performance (Meyer and Kieras, 1997).

Recently, a broader engineering concept of a *digital twin* has emerged in manufacturing, research, and development. A digital twin is a model of a physical system that allows a designer or engineer to virtually examine, stress-test, do trade-off analysis for future or existing systems. For example, a digital twin of a manufacturing plant can help an engineer identify the feasibility of replacing one automated system with another. Digital twins are also being developed for modeling humans, including anthropometry applications, safety (virtual crash-test dummies), human heat stress/strain models, and the like.

There have also emerged many cognitive digital twins that help predict cognitive performance of humans. To be fair, these are generally not distinguishable from the tradition of computational and cognitive models. There have been many AI-based models that have proven useful in predicting human behavior. Sometimes these adapt existing models to human paradigms, sometimes they test existing AI systems to see whether they make the same errors or decisions humans do, and sometimes they use AI or machine learning tools to develop interesting models.

For example, the MIT-300 saliency benchmark (Judd et al., 2012; Bylinskii et al., 2016, 2018) found that the best predictions were made by a deep image recognition network (Kümmerer et al., 2017), with the worst made by psychologically-inspired models (Itti and Koch, 2000; Walther and Koch, 2006). These were models that adapted existing image recognition systems to predict salience and eye movements when inspecting images. The fact that object-recognition networks performed better than 'untrained' naive models suggests the importance of meaningful content in human image analysis—we look at things; not just salient patches of color.

As an example finding human biases in existing systems, Martínez et al. (2022) argued that biases such as representativeness, confirmation bias, primacy, anchoring, and causal illusions can all be exhibited by AI systems due to selection of training data, design of tests, and patterns of usage. Vicente and Matute (2023) found human users of AI are generally not equipped to ignore these biases and inherit them in their AI-assisted decisions. Similarly, Marjeh et al. (2024) found that aspects of sensory perception are predicted by judgments of similarity made by GPT-4 model, and co-trained models on vision did not lead to improvements.

As an example of the third approach, Peterson et al. (2021) used large datasets of risky choice behavior, and trained machine learning models constrained to produce interpretable theories of decision making. This is perhaps a meta-adaptation of machine-learning to generate and evaluate massive sets of theoretically-interesting models of human judgment and decision making.

Table 4.2: Different kinds of anthropomorphism discussed in literature on human-AI and robotic interactions.

Aspect/Kind of Anthropomorphism	Description
Cognitive ^[a]	Users expect AI to do cognitive work in the same ways humans do
Effective ^[b]	Users initially use human capabilities to understand AI, and it can be effective
Mechamorphism ^[c]	Users think like a machine
Social/CASA ^[d]	Humans treat machines socially as they do humans
Ethopoeia ^[d]	Humans know the machine is not human but respond to it as if it is
Emotional Connection ^[e]	Aspect of social anthropomorphic design that describes a commonly-reported human-AI interaction pattern
Functionality ^[e]	Capability, accuracy, or ability of AI to accomplish tasks
Interactivity ^[f]	Ability to interact with user and respond to user feedback
Mentalism ^[g]	Perceiving behavior in terms of mental states
Humanism ^[g]	Treating non-human entity as if it were human
Spiritualism ^[g]	Attributing a spiritual nature to a non-human entity
Anthropomorphic design ^[f]	AI is designed to behave like humans
Anthropomorphic expectations ^[f]	Users expect human-like motivations or behavior from AI
Anthropomorphic attributions ^[f]	Users attribute human-like motivations or properties to AI
Emulation ^[h]	Goal of creating AI that emulate human capability

Notes: ^[a] Mueller (2020); ^[b] Bos et al. (2019); ^[c] Glasgow et al. (2022); ^[d] Nass et al. (1994); ^[e] Gibbons et al. (2023); ^[f] Current chapter; ^[g] Spatola et al. (2022); ^[h] Shneiderman (2020).

4.6.6 Summary of Anthropomorphism in AI

Issues of anthropomorphism pervade human-AI interactions. In this section, we described some of the concepts that guide the design of AI and human-AI systems. This begins with both cognitive and social expectations and attributions of users of AI—we tend to attribute and expect systems to behave like humans. But because of this, designers often anthropomorphize their tools, encouraging them to act like humans as well. Table 4.2 summarizes these distinctions.

In this chapter, we discussed several areas of research related to AI

with human-like behavior. This includes ways to evaluate whether an AI can 'think' in terms of the Turing Test; design goals of building AI that can pass such tests in the form of Artificial General Intelligence, and broader research investigating conditions under which people expect or treat an AI as if it is human (anthropomorphic expectations and attributions) and the response to this which is to design systems that do indeed behave like humans (anthropomorphic design). Of course, the epitome of anthropomorphic design is to create AI that can legitimately fulfill these expectations in the form of AGI and thus pass the Turing Test.

Although these notions are important to human-AI interaction, they are not without criticism. Many useful AI systems are not human-like; people do not treat them like humans, and people do not expect them to behave like humans. Furthermore, anthropomorphic design may hijack social norms and even evolutionary mechanisms that may confuse or mislead users into endowing the AI with more capability, agency, or rights than it really has.

For the Human Factors Practitioner

This chapter covers several related topics. The question of whether a machine can 'think' is typically not of much practical use. However, this question has led to many suggestions about interesting and clever ways of assessing capability of AI and automation. Likewise, the question of whether or when AGI emerges is unlikely to help design or evaluate a new system, even as the way we work is being transformed by systems that are exhibiting more and more capability. On the other hand, this chapter also discusses an important practical and ethical issue in the design of systems: anthropomorphism. Two sides of this are important—people often anthropomorphize technology and AI in the way they interact with it. This should be recognized because it can be useful for adoption, user expectations, new interaction modes, to name a few. But it is also something that designers can leverage, potentially making the system appear more friendly, but also potentially using it as a 'dark pattern' that users should be protected against through transparency and user testing.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction, Michigan Technological University, Spring 2024.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Cischke, C., Frisch, B., Wang, J., Wang, Y., & Mueller, S. T. (2026). Human-like AI: The Turing test, artificial general intelligence, and anthropomorphism of AI. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Trust in Automation: Definitions, Models, and Related Concepts

Samantha Walker, Erin Matas, Katie Ulinski, Brandon Woolman, and Shane T. Mueller
Michigan Technological University

Abstract

Trust is a central construct in human-AI relations. Research on trust in automation has been the focus of research for decades, and has led to a number of theories and models of trust and trustworthiness. In this chapter, we discuss how trust has been defined, some of the important theories of trust, the underpinning of human-machine trust in interpersonal (human-human) trust, and related concepts. We discuss how trust has many definitions, models, theories, and components. We argue that trust in automation and AI is best understood as contextual, multi-dimensional, componential construct. The primary reason for understanding trust is to appropriately manage how new automation systems should be embedded or replace existing systems, in order to anticipate problems, but also improve processes.

5.1 Introduction

One of the most central topics in the study of automation and human-machine interaction is trust. Trust is an important topic because the fundamental exchange made in automation is gaining efficiency by allowing a machine to take over work, but in some sense the human is still being responsible for the outcome. Trust can be thought of as the thing being demonstrated when a person or organization is willing to turn over work but remain responsible.



Figure 5.1: Word cloud based on an ensemble of definitions of trust (see Section 5.2).

Trust (and its partner trustworthiness) has been discussed and defined broadly within the human factors and human-machine teaming literature, and in this chapter we will discuss some of the definitions of trust, popular models of trust, and theoretical approaches to understanding notions of conditional trust such as *calibrated* and *appropriate* (e.g. Lee and See, 2004), *justified or unjustified* (Hoffman, 2018), or *warranted or unwarranted* (Jacovi et al., 2021) *trust*. Finally, we will discuss some of the particular ways in which trust in AI systems is both similar and different to trust in other automation systems.

Our aim in this chapter is not to pick one true definition of trust, but to show how definitions, interpersonal analogies, and formal models converge on a practical problem: when should a human rely on a machine, for what, and on what grounds? We use that question to connect classic automation trust research to contemporary concerns about generative AI, AI services, and regulatory concerns of trustworthy AI.

5.2 Definitions of Trust and Trustworthiness

5.2.1 Defining Trust

Trust is not a simple term to define, although many similar definitions exist in the literature (Figure 5.1). Definitions have been proposed in many domains, including technology (Taddeo, 2009), business and commerce (Li and Betts, 2003), nursing (Meize-Grochowski, 1984), information systems (McKnight and Chervany, 2001), economics (Arai, 2009), education (Noonan et al., 2008), and philosophy (Simpson, 2012; de Fine Licht and Brülde, 2021). The meaning of trust can vary depending on the context of the term and the perspective of

the person using the word. Different researchers tend to focus on different meanings of the word. Nevertheless, most definitions share properties and allow us to identify some of the common approaches to thinking about trust (Adams et al., 2003). To begin, we will examine some of the common ways trust has been defined.

Trust as an expectation

One early framing of trust frames it as an expectation about a multi-dimensional construct. Here, the constructs (persistence of the natural order, technical competence, and fiduciary responsibility) are quite broad and go far outside human-machine trust, but there is a sense in which trust is a prospective attitude about expected behavior.

Trust as an expectation has been defined as:

Trust is the expectation (E), held by a member of a system (i), of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but is not necessarily isomorphic with, objective measures of these properties (Barber, 1983)

Trust as an Attitude

Trust may be defined as an attitude, that is a way of thinking or feeling about a particular situation (Jones, 1996). Under this view, trust may be more focused on how an individual's feelings or thoughts impact their trust in someone or something.

Trust as an attitude can be defined as:

"The attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004).

"the attitude of a user to be willing to be vulnerable to the actions of an automation based on the expectation that it will perform a particular action important to the user, irrespective of the ability to monitor or to intervene." (Körber et al., 2018)

Trust as a Belief

Trust as a belief involves the expectation that an agent will behave in the desired way (Reiersen, 2017). Beliefs are normally an indication that one holds something to be true (whether justified or not), so in this sense, trust accepts as true that something will behave in an expected way.

Trust as a belief has been defined as:

“... an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of its designers’ intentions, be relied on to achieve the design goals” (Moray and Inagaki, 1999).

Trust as a State

Trust as a state involves the condition of the agent at a particular time. This includes how the trustee (another person, a system, etc.) will perform or act at the necessary time for the trustor (Adams et al., 2003).

Trust as a state can be defined as:

“A state involving confident predictions about another’s motives with respect to oneself in situations entailing risk” (Boon and Holmes, 1991).

Trust as an exchange of risk for benefit

Many notions of trust have a sense of a willingness to put one’s own consequences under the control of another, often for some benefit. When one trusts a bank, an automated vehicle, or a stranger’s directions, they are accepting one sort of risk in exchange for a benefit or reduction in another risk.

Trust as an exchange has been defined as:

“The willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party” (Mayer et al., 1995).

Trust as a measure

Many researchers use trust as a measure to assess a person’s potential willingness to use or rely on something. The focus on knowing whether some intervention changes (at a subjective or behavioral level) how much someone “trusts” a system. This is an extension of thinking about trust as a state (or at least, a continuous set of states). Practically speaking, it typically involves either an explicit multi-dimensional measure of trust via agreement responses on a Likert-style scale, or a related measure of *reliance*—whether a user either is more willing to use a system, or at least more willing to *agree* that they would use a system.

There are many questionnaire-based assessments of trust. Many include a question such as item 9 of the TIA:

“On a scale from Strongly disagree to Strongly agree, rate the following statement: 9. I trust the system” (Körber, 2019a).

Here, we expect a user to understand what is meant by trust (even though researchers cannot agree). This scale (and most others) typically involves asking a set of additional questions, including about a system's capabilities, its developers, the user's knowledge of it and similar systems, etc. We will discuss a number of these specific instruments in Chapter 6.

Trust as a Verb or Process (Trusting)

Trust as a verb involves the active process of trusting a person or system. Under this view, trusting can be independent or dependent of the context and situation, meaning that multiple trust relationships can occur with the same agents at the same time (Hoffman, 2017).

Trust as a verb has been defined as:

"... a continual process of active exploration and evaluation of trustworthiness and reliability, within the envelope of the ever-changing work and the ever-changing work system" (Hoffman, 2017).

Trust as a behavior

Although trust is often thought of as an internal attitude, that attitude can be assessed via behavior (Adams et al., 2003). That is, one can infer trust if one behaves as if one trusts. This could be as an economic behavior (Zheng et al., 2002) or within a game theory context such as prisoner's dilemma (Juvina et al., 2015), as investing or cooperating indicates you trust your partner, although in many human-machine settings a behavior of using the system is considered *reliance*. Trust as a verb has been defined as:

"the essential features of a situation confronting the individual with a choice to trust or not in the behavior of another person are.... the individual is confronted with an ambiguous path.... If he chooses to take an ambiguous path with such properties I shall say that he makes a trusting choice; if he chooses not to take the path, he makes a distrustful choice." (Deutsch, 1960).

Commonalities of Trust Definitions

This exercise illustrates that trust has many distinct senses. Nevertheless, these senses overlap, and many of these definitions have similar components that help shape the meaning of trust. Table 5.1 shows common words used in definitions of trust that may provide a better understanding of what is involved in the meaning of trust.

First, trust always involves some type of **relationship** between the trustor and the trustee, whether it be a relationship between two humans, a human and a machine, an organization, or multiple actors. In some cases, a trust

Table 5.1: Common components of trust definitions.

Component	Role in the definition of trust
Relationship between trustor & trustee	Trust is supported by past relationship establishing a track record warranting trust.
Goals/Motives	Trust does not require shared goals or motives, but involves an understanding that the trustor's goals/motives are likely to be advanced through the relationship.
Expectations	Trust involves an expectation that the trusted party will do something.
Willingness	Trust is a willing relationship.
Risk	Trust involves risking money, reputation, or other valued outcomes.
Vulnerability	Trust involves making oneself vulnerable to the actions of another.
Non-zero gain	Often, a trust relationship involves non-zero-sum gains in which both parties benefit.

relationship may be between a present and future party, as is often the case in a legal trust used to manage assets. Trust is often second-hand: a builder or plumber may be trusted based on the recommendation of a friend. Trust involves some form of **goal** or motive to reach or some expectation of an outcome, normally with some sort of exchange of risk, vulnerability, and benefit. This does not mean that both parties have shared goals, but rather that the trustor's goals can be advanced through the relationship, such that putting trust in something involves the **expectation** that the desired outcome will be achieved. Similarly, there has to be a **willingness** to trust the trustee in order for any trust to be established. This willingness entails the potential for **risk**, and the outcome may not work out in the way intended, leading the trustor to be **vulnerable** to the behavior of the trustee. However, there is often the prospect of a non-zero-sum outcome, where both parties may benefit from the trust relationship.

We will not pick or create a new definition for trust. We simply recognize that when people talk about trust, they can mean different things, and an important initial question to ask is “what do you mean by trust?”

5.2.2 Defining Trustworthiness

In the most basic of definitions, the term “trustworthy” can be defined as the ability to be trusted, or being worthy of trust, but this is almost a circular definition. Several of the models we discuss in the next section characterize properties of trustworthiness. For something or someone to be trustworthy,

the characteristics of trust must be earned (Jones, 2012). This involves the acceptance of the aspects of trust, such as risk and vulnerability that come with trusting in order to get to the expected outcome (Hardin, 1996). Thus, trust often is a behavior or attitude of the trustor, whereas trustworthiness is often thought of as a property of the trusted entity.

In terms of the relationship between the trustor and the trustee, trustworthiness can be defined as the trusted acting in the way the trustor expects (Hardin, 1996), perhaps in line with an agreement or contract (Jacovi et al., 2021). If a system acts in the way it is supposed to, it will likely act the same way again, leading to the conclusion that the system is trustworthy. The perceived trustworthiness of a system, especially a new system, may also be influenced by external factors such as institutional backing, independent oversight structures, etc. (Hardin, 1996; Shneiderman, 2020). For example, if a new in-home AI system is released from a reputable company known for protecting user privacy and data, consumers may find it more trustworthy than if the same system were released by a different company. This all is involved in how people come to trust a system and deem it trustworthy and thus may adopt it into their lives.

Interestingly, although trust and trustworthiness go hand-in-hand, many theories of trust in automation appear to focus on trustworthiness—things that make an automated system able to be trusted, such as its reliability or predictability. In a later section of this chapter, we will discuss some of these models, which often are intended to predict how both trustworthiness and other factors lead trust to change.

5.3 Human–Human (Interpersonal) vs. Human–Machine Trust

One of the common themes of research on trust in automation is its basis in human-human (interpersonal) trust. As discussed in Chapter 4, one consequence of social robotics and anthropomorphic design is that it tends to enhance trust in the system (although the human-like features may only impact perceived trustworthiness and not actual trustworthiness). Consequently, it is useful to understand some of the existing theories on interpersonal trust such as Mayer et al. (1995)'s theory of organizational trust. While a number of researchers have based their approaches to understanding trust in machines on interpersonal trust (e.g. Muir, 1994; Adams et al., 2003; Lee and See, 2004; Hancock et al., 2015; Hoffman, 2017), it is important to note that almost all argue that this is a good starting point but insufficient, and have elaborated additional aspects for understanding trust and trustworthiness in machines (e.g. Jacovi et al., 2021). Thus understanding interpersonal trust is a useful starting point for understanding human-machine and human-AI trust.

5.3.1 Theories of interpersonal trust

There are many different beliefs on why humans form trusting relationships. Mayer's theory of organizational trust suggests human-human trust develops for two distinct reasons: because we have the disposition to rely on others (probably passed down via genetics and culture) and because others demonstrate that they can be relied on (Mayer et al., 1995; Baer and Colquitt, 2018). On their own, each of these concepts have a limited ability to explain trust, but together to form a more complete picture of why humans trust. This theory posits a cognitive and affective basis for trust, both of which benefit from time knowing the trustee. The theory does not address trusting when in short-encounters, such as when trusting a stranger.

Other theories of why we trust arise from an evolutionary perspective. For example, Montague et al. (2015) identified the *Three R's* of trust: reaping, regarding, and recursive modelling. By reaping, they mean trust is responsive to reward and punishment. Regarding involves pro-social mechanisms that encourage trust. Recursive modelling represents a theory-of-mind or game-theoretical aspect, where our trust is reciprocal if it is reciprocated. It supports coordination when each party can reason about the other's reasoning, allowing temporary alliances even when short-term payoffs favor defection. From an evolutionary perspective, trust can look irrational from a short-term self-interest view, but (Fehr and Gächter, 2002) demonstrated that cooperation can be enforced by *altruistic punishment*—where punishment is costly and yields no material gain.

Others have developed more biologically- and neurologically-focused theories of why people trust. One theory argues that the release of oxytocin in the brain leads to trusting relationships (Kosfeld et al., 2005; Nave et al., 2015). Oxytocin is a neuropeptide produced in the thalamus known to promote social behavior, including bonding, maternal behaviors, and sexual behaviors. Since so much of trust is based on social behavior, oxytocin is an important factor in creating trust, and it has been shown to help in overcoming aversions to the uncertainty of the behavior of others. And the motivational and emotional brain circuits that are implicated have consequences in emotional responses that can drive behavior. For example, one may feel guilty about not trusting someone (Crockett et al., 2010), so trusting is favored because those negative feelings are avoided.

These three levels of analysis are not necessarily at odds: evolutionary benefits have accrued to those who both have a willingness to trust but are sensitive to reward and punishment based on the outcome of the trust; this behavior may rely on specific neural signaling that was adapted from core caring and bonding behaviors in animals, but is still reflected even at the organizational level. Interestingly, unlike Mayer's theory, the evolutionary and biological theories focus primarily on the trustor. This has two implications for

trust in automation. First, to the extent that our trust is derived from neural computations related to trusting other people, it is likely that behaviors that signal trustworthiness co-evolved with how we detect trust, and so aspects of interpersonal trust may not translate to human-machine contexts. It is also likely that, to the extent that trust has evolved, machines and AI can exploit the signals of trust. As discussed in the previous chapter on anthropomorphism, anthropomorphic aspects of robotics and AI such as cuteness have been argued to be “Dark Patterns” (e.g. Lacey and Caudwell, 2019). Of course, these patterns merely reflect other ways in which other humans exploit trust, and so it may lead to an overall distrustful stance toward novelty. Overall, even if trust mechanisms are formed through evolution, they are also likely to be influenced by social learning and demonstrations of (un)trustworthiness, and so many aspects of interpersonal trust are likely to translate to human-machine contexts.

5.3.2 Forming Trusting Relationships

From the evolutionary perspective, infant animals who appropriately trust and distrust are certainly more likely to survive. In humans, the disposition towards trust is thought to begin in children 6 to 18 months old (Rotter, 1971). As a parent meets the needs of an infant, the infant may be able to generalize trust within and outside the family unit (Sakai, 2010). While a disposition towards trust can be formed, trust in interpersonal relationships is based on each relationship—suggesting we are well equipped to compartmentalize our trust and not simply be trusting or distrustful overall. One theory suggests that for every relationship, there exists a “zero baseline” of trust where a person neither trusts or distrusts another (Rempel et al., 1985), and through experience trust can move from the baseline. However, (Borum, 2010) argues that people enter a situation with a higher or lower starting level depending on multiple factors including their psychological and cultural context. This suggests that there may be strong individual differences in a baseline trust stance toward others (and presumably toward machines).

5.3.3 Implications for Human–Machine Trust

Humans are driven by social interactions and social expectations. Even with technology, people may perceive technology as more human-like than system-like (Lankton et al., 2015), and these different aspects of trust may operate differently for different kinds of systems (e.g., human-focused facebook vs. a system-focused database). Therefore, people may interact with technology as if they are interacting with another human, as opposed to a machine (Reeves and Nass, 1996). However, trying to understand human-machine trust through the lens of interpersonal trust may also lead to incorrect assumptions about

how trust with machines form (Hoffman, 2018). Aspects of interpersonal trust, such as humility and leadership, are not considered in the literature of human-machine trust. The type of machine being used needs to be considered when studying how the trusting relationship is formed and works.

Humans can also specifically trust or distrust the algorithms that are being used by AI and ML. Pariser (2011) coined the term “filter bubble” to describe how algorithms prominent on social media are providing each of us an individually-tailored view of the world, but as shown by Klug and Strang (2019), the most intense users of social media recognized this more often and have a more negative perception of the algorithms. Cabiddu et al. (2022) presented six main determinants of trust in algorithms: users’ propensity to trust; IT acceptance levers; human-like characteristics the AI presents; social influence; familiarity; and the system characteristics of the AI. Similarly to human-human trust, some of these dimensions can change over time. Since AI appears to learn and problem-solve, people may be more likely to trust AI over other technologies. AI’s ability to be humanized lends itself well to fostering trust. However, when an algorithm has made a mistake, people may experience algorithm aversion. Algorithm aversion is the reluctance of humans to use algorithms, which are superior but imperfect, in favor of their own judgment (Burton et al., 2020). This suggests that trust in AI and algorithms can be destroyed, perhaps more easily than it can be restored.

5.4 Theories and Models of Trust in Automation

Numerous authors have proposed models or theories that attempt to explain trust, at a number of different levels of abstraction. Many of these were reviewed by Adams et al. (2003), which remains the most comprehensive scholarship on trust in automation. This section will summarize some of those models, and introduce some other prominent models that have emerged since that report, with the goal of providing the main motivations and approaches of these different models rather than identifying particular commonalities or differences between models.

5.4.1 Conceptual Models of the Predictors of Trust(worthiness)

Muir (1994) proposed maybe the first framework or model used to describe trust in automation, based on previous social notions of trust by Rempel et al. (1985) who identified dynamic properties related to predictability, dependability, faith, and by Barber (1983), who identified expectations such as persistence, technical competence, and fiduciary responsibility. This model might be thought of as a causal models of the predictors of trust, which is useful because it can help system designers understand how they might change automation or AI to improve trust. For example, recent advances in large language models

have been widely criticized for being unpredictable and undependable insofar as they can produce different output to the same prompts, high-confidence hallucinations, and bad advice based on low-quality training sources. These models would likely predict this to lead to a loss of trust.

More generally, the goal of creating understanding levels of trust as predicted by aspects of a system and aspects of the user or operator has been the basis for many other models. For example, Llinas et al. (1998) and Seong and Bissantz (2000) adapted this basic notion, but extended it using the Lens model (Brunswik, 1943). The Lens model has been used in human factors research to understand judgment in socio-technical systems about multiple cues of differing validity (e.g., Mosier and Kirlik, 2004). In the context of trustworthiness of a system, it is used to examine how trust judgments are supported by a number of characteristics of a system. The lens notion is useful because these characteristics need to be observed by a user, and the cues of those characteristics might be weak or unreliable signals. As shown in Figure 5.2, the system on the left has a number of observable signals of its trustworthiness. An operator or user (on the right) may not clearly know these signals, or not have experienced them. If the weights on the right match the signals on the left, the model considers trust to be 'calibrated'. However, a trustworthy system may not always provide observable behavior to help the operator develop this trust. Presumably, many of the observable characteristics overlap with those proposed by Muir (dependability, reliability, etc.), but the key insight of this model is that trust is developed through the interaction of information signaled by the automation and interpreted by the user, but they may not accurately use these signals as a basis for their trust.

5.4.2 Models incorporating the user's influence on trust

The specific approach of Muir was influential, insofar as it framed the problem of understanding trust as understanding the factors that influence trust by a user of automation. Both this specific model and the general approach has been adopted and extended to incorporate aspects of the user as well. For example, Lee and Moray (1992) incorporated a third dimension attributed to Zuboff (1988), who found that understanding and trial-and-error experience is related to predictability and reliability (see Table 5.2).

Others have expanded on Muir's basic notion by identifying other kinds of taxonomies. For example, Madsen and Gregor (2000) proposed a model of human-computer trust and included two main themes: cognition-based trust (incorporating understandability, competence, and reliability), and emotional/affective responses of the user (personal attachment and faith). Kelly (2001) proposed a model with three main themes: one incorporating the so-called 'competence' of the automation reminiscent of Muir dimensions (dependability, reliability, usefulness, and robustness), and two additional themes

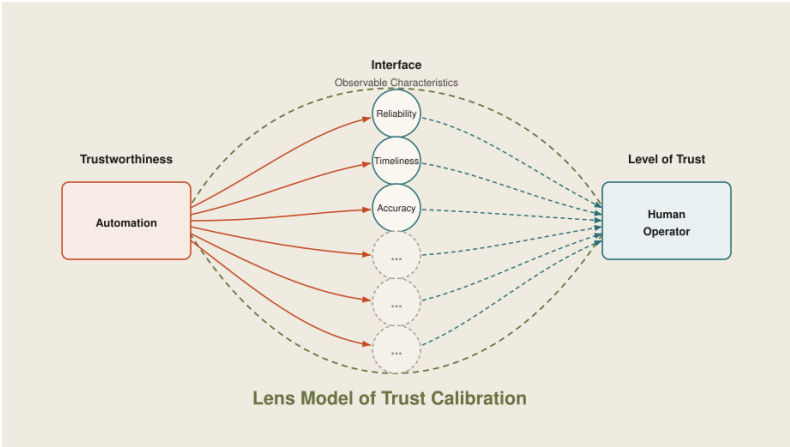


Figure 5.2: Lens model of human trust in automation, based on Seong and Bisantz (as presented in Adams et al., 2003, Fig. 9). Observable characteristics such as reliability, timeliness, and accuracy (among others) mediate between the automation’s trustworthiness and the operator’s level of trust; calibration reflects correspondence between those.

Table 5.2: Reproduction of Table 2 of Lee and Moray (1992) identifying common properties of previous trust definitions.

	Barber (1983)	Rempel et al. (1985)	Zuboff (1988)
Purpose	Fiduciary Responsibility	Faith	Leap of faith
Process		Dependability	Understanding
Performance	Technical competence	Predictability	Trial-and-error experience
Foundation	Natural Laws		

related to the user: understanding (predictability, intention, and familiarity) and self-confidence (faith, reputation, skills, training, experience). These last two frameworks importantly incorporate elements of both the automation and aspects of the user.

Consequently, many models of trust incorporate a ‘user’ or ‘trustor’, as well as the system or ‘trustee’. These entities form the basis for more sophisticated models of trust, but there are several that can be considered systems models attempting to capture a bigger picture.

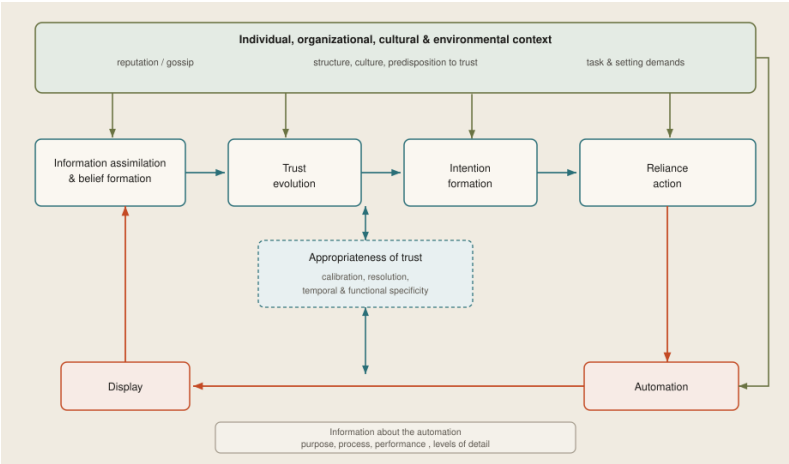


Figure 5.3: Lee and See (2004)'s model of trust in automation.

5.4.3 Systems models of trust

A number of models have been proposed that might most accurately be described as systems models: accounting for a variety of aspects of trust, including the precursors, consequences, interactions, decisions, users, and technology. These typically focus on three major components: aspects of the trustworthiness of the system, aspects of the trustor, and consequences in terms of decision, uses, or reliance.

Lee and See (2004) model of Trust in Automation.

Lee and See (2004) proposed a system model of trust that incorporates individual, organizational, cultural, and environmental factors (Figure 5.3). It is structured as a general control loop involving automation, display, belief formation and reliance, with different factors influencing different components. This loop can be thought of as controlling the 'appropriateness' or 'calibration' of trust: the operator decides to rely on the automation for some purpose, and the automation acts and displays results, which then lead the operator to evaluate whether it behaved appropriately and adjust their willingness to trust the system later.

Adams et al. (2003) Automation trust model in military context.

In their extensive report on trust theories, models, and measures, Adams et al. (2003) propose a preliminary model of trust in automation for military contexts (Figure 5.4). The model is organized around a *trust development process*:

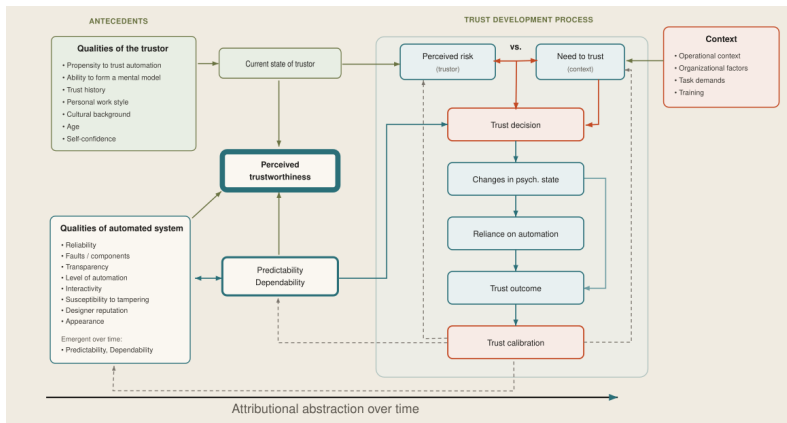


Figure 5.4: Preliminary model of trust in automation in a military context, reorganized from Adams et al. (2003) (their Fig. 16). The vertical spine is the trust development process; left-hand panels summarize trustor, system, and contextual antecedents. Redrawn for this handbook.

perceived risk is weighed against a contextual need to trust, producing a trust decision that may change psychological state, support reliance, yield an outcome, and feed back through trust calibration. Antecedents enter that process from three sides—qualities of the trustor (current state), qualities of the automated system (including predictability and dependability that emerge with experience), and operational/organizational context.

This model was developed to characterize common themes of trust in technology in military contexts. Although this is true for other industries as well, in military settings, there is a constant stream of new systems including new automations that are often not developed with the warfighter's direct involvement or feedback. These systems often fail to be adopted, often because the capability is lacking, but also because the end-users lack trust in the systems in comparison to their older, less capable but proven systems. Even though in theory, adoption can be dictated from higher-level command, the individual units and warfighters often have the opportunity to weigh whether adoption and use is warranted in their own use context, and may simply work around the new system. The model highlights how there are many ways technology adoption can go wrong, and also many vectors to improve trust and adoption of systems that are genuinely useful or necessary to handle new environments.

Although this was designed for understanding trust in a military context, it is likely to be broadly generalizeable to other settings, including industrial automation and consumer systems.

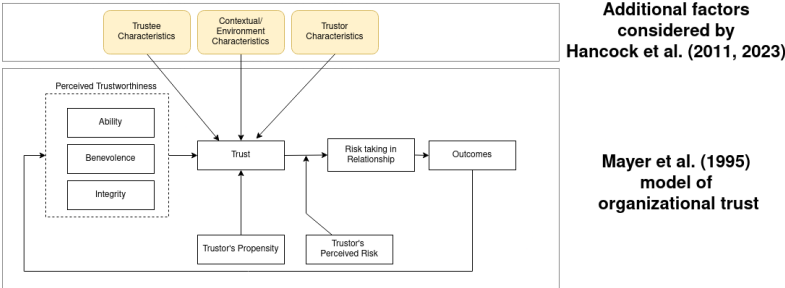


Figure 5.5: Hancock et al. (2011, 2021) augmented Mayer et al. (1995)’s model of organizational trust to incorporate specific aspects of the trustee, the trustor, and the context.

Mayer-Hancock Trust Framework

Hancock and colleagues (Hancock et al., 2011; Schaefer, 2016; Hancock et al., 2021) adapted Mayer et al. (1995)’s model of organizational trust to understand the elements of human-machine trust. This model involves the perceived trustworthiness of the technology, characteristics and perceptions of the trustee and trustor, and examines outcomes of trust. Consequently, the model has high overlap with many other models of trust. One substantial advance made recently (Hancock et al., 2021) is that they used this model as a basis of a large meta-analysis including 338 studies and over 2000 effect sizes. This model enabled elaborating and validating many of the putative precursors and consequences of trust identified by other researchers. Furthermore, this elaboration helped focus the authors on three main aspects of trust: those related to the trustee, the trustor, and the context. Hancock et al. (2021) also examined how these factors impacted distinguished between directional trust (i.e., supervisor/supervisee relationships); see Figures 5.5 and 5.6.

As with other models, this model organizes the trustee and trustor; it identifies influences of trustworthiness and trust, and captures larger-scale contextual concerns. Overall, the model was intended to promote ‘appropriate’ levels of trust in robots, and so even if the notion of trust calibration is not central to the model, it is the overall goal.

5.4.4 Process/cognitive models

These models typically focus on the cognitive processes that underlie development and revision of trust, decision processes that lead to use and reliance, the kind of information used to inform trust, and the way in which trust gets adjusted or calibrated based on evidence. In fact, it appears that the common thread of process and cognitive models of trust is that they focus on how infor-

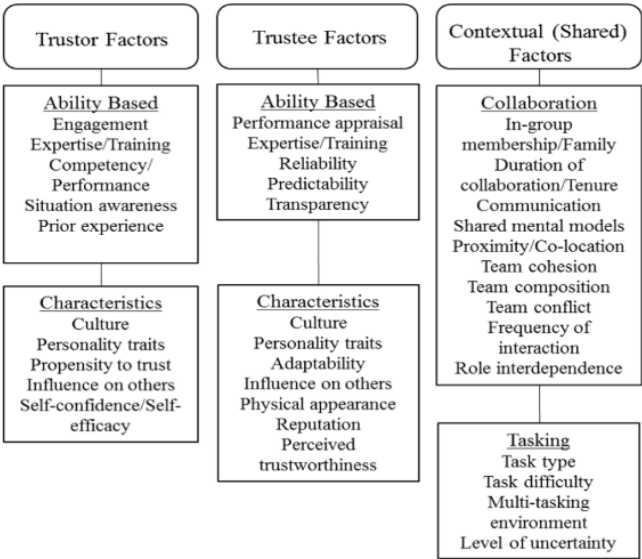


Figure 5.6: Factors identified by Hancock et al. as potential antecedents to trust and trustworthiness, organized by trustor, trustee, and contextual (shared) factors. (Adapted from Fig. 2 of Hancock et al., 2021 ; used under CC BY 4.0 open access license)

mation and evidence is used to adapt (or "calibrate") trust to an automation system, context, or application.

Muir model of the relationship between automation, trust and predictability

Muir (1994) proposed what might be considered a process model of trust, in terms of a flowchart of processes and decision (Figure 5.7). Here, an operator forms a mental model of a system related to the three principal factors identified by Barber (1983), as discussed further in Chapter 7 (Mental Models and Transparency), which work together to form a basic degree of trust in the automation. One might consider whether this internal model is calibrated with the actual trustworthiness of the system (the red boxes at the top), but more importantly, the mental models enable a prediction of the system's behavior, which can be compared to observed behavior (tempered by confidence in one's own predictions) to understand whether confidence in the system is calibrated. This system essentially suggests an adaptive model by which behavior and mismatch between expectations and observations are used to adjust trust. This model also incorporates the *good regulator* theorem (Conant and Ashby, 1970) discussed in Chapter 2, insofar as it posits that the operator/user incorporates

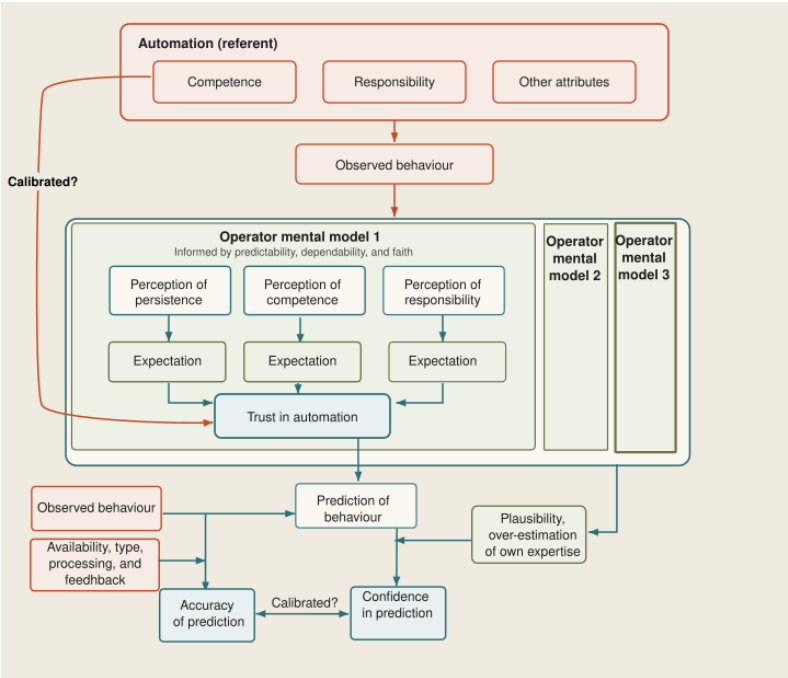


Figure 5.7: Simplified model of the relationship between automation, operator trust, and behaviour prediction (redrawn from Muir, 1994, Fig. 3). A mental model of a system creates predictions, and violations of those predictions can influence ultimate trust.

a mental model of the system. Presumably, although the model is focused on trust, it could also be adapted to understand the operator’s ability to use and “regulate” the system.

Cohen et al. (1997) proposed the Argument-based Probabilistic Trust (APT) model: a conceptual account of trust as the outcome of an argumentation process under uncertainty, intended to be specific enough to support training design (see Figure 5.8). APT is inspired by Toulmin (1958)’s argumentation model, using it to structure a formal argumentation process with a claim (the putative conclusion), grounds (facts in support of the claim), warrant (beliefs supporting causal conclusions), backing (assumptions supporting warrant), and rebuttals (counter-arguments). In APT, this is used to consider a probabilistic process, where the claim gains *qualified* or probabilistic support. The specific aspects of the argument structure are linked to trust in a decision aid, so that *grounds* indicates awareness of current system and situation features that support a *qualified claim* (the chance of correct system action over a period *t*), via a *warrant* (belief that those features are generally tied

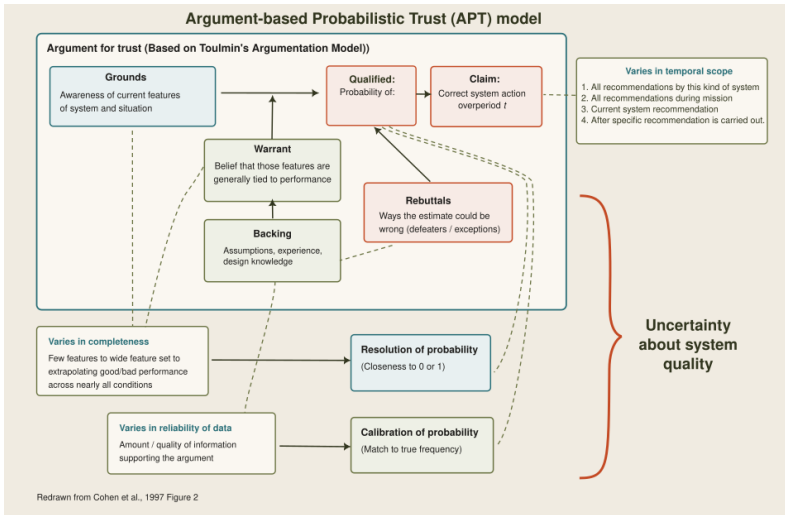


Figure 5.8: Argument-based Probabilistic Trust (APT) model of Cohen et al. (1997) (redrawn after Adams et al., 2003, Fig. 7). The Toulmin-inspired core (grounds, warrant, backing, rebuttals, qualified claim) is surrounded by routes that are not part of a basic Toulmin schematic: variation in temporal scope, completeness, and reliability of information, linked to resolution and calibration of the probability that the system will act correctly.

to performance) that is itself supported by *backing* (assumptions, experience, and design knowledge). *Rebuttals* are ways that estimate could be wrong. Furthermore, they identify (via dashed lines) many ways that uncertainty, completeness, and reliability can impact support for the claim.

Together, these characterize *uncertainty about system quality*, but it enabled Cohen et al. (1997) to formally define probabilistic models in event trees, to understand how trust in a decision aid can unfold over time. Furthermore, they also used it to identify specific training and decision aiding strategies: uncertainty of grounds can be improved through aids that help monitor appropriate information; improved training of mental models can impact backing and warrant; improving critical thinking skills to improve rebuttals or counter-arguments about the trust estimate, etc. As Adams et al. (2003) note, APT's main value is charting how trust varies across users, decisions, situations, and phases of decision-aid use—not treating trust as a single static score.

An ACT-R Computational model of trust

A series of papers by Juvina and colleagues (Juvina et al., 2015, 2019; Collins and Juvina, 2021) described a computational cognitive model of trust dur-

ing repeated game-theory 'strategic interaction' games (Prisoner's Dilemma, Chicken, and a multi-arm trust game).

They developed a computational cognitive model using instant-based learning (IBL), which maintained memory for past interactions of the user and their partner, and an accumulator model that tracks elements of the game. Unlike any of the other models of trust we review here, this model makes specific predictions about behavior that were generally consistent with human research. A related advantage is that, by the nature of the game, data involve hundreds of repeated game outcomes (e.g., cooperate vs defect) that can be used as a behavioral proxies for trust. Of course, the limits of this model are that it is narrowly-focused on a handful of simple and repeated strategic games, and so it does not need to incorporate many of the elements of trust and trustworthiness that are important in more natural contexts

5.4.5 Statistical and Predictive models of trust

Researchers have developed more formal statistical and machine learning models that have been used to either test or discover some of the conceptual elements of trust. For example, Muir (1994) conceptualized her initial model on experiments using regression analysis, and Lee and Moray (1992) developed similar regression and autocorrelation models to examine the dynamics of trust.

Following Barber's (Barber, 1983) conceptual framing of trust as an equation

$$T_i = [E_{i(P_n+P_m)}] + [E_{iTCP_j}] + [E_{iFR_j}] \quad (5.1)$$

where i refers to the trustor, j refers to the trustee, T is trust, P_n and P_m is moral and natural persistence, TCP is technically competent performance, and FR is fiduciary responsibility. She expanded this to a linear model as follows:

$$T = B_0 + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_1X_2 + B_5X_1X_3 + B_6X_2X_3 + B_7X_1X_2X_3 \quad (5.2)$$

where B are all parameters, X_1 is persistence, X_2 is TCP, and X_3 is FR, which captures all the main effects and interactions.

Furthermore, Muir and Moray (1996) proposed a new trust equation based on a stage model (Rempel et al., 1985; Muir, 1994):

$$Trust = Predictability + Dependability + Faith \quad (5.3)$$

which they tested in a regression model on subjective ratings of each of these constructs predicting ratings of trust in different targets in the simulation

(three properties of three different pumps), during a simulator experiment in which users controlled a pasteurization process.

$$\begin{aligned} \text{Trust} = & \text{Predictability} + \text{Dependability} + \text{Faith} + \\ & \text{Competence} + \text{Responsibility} + \text{Reliability} \end{aligned} \quad (5.4)$$

They found that these new predictors did not improve the prediction of trust. They also used similar modeling to establish that trust ratings were related to the time spent in “auto-pump” mode, indicating a close link between stated trust and reliance; and showed that the variability in the control error of the pumps was a strong predictor of trust.

This regression modeling approach gains traction and by collecting many subjective ratings during a simulation, and shows how these are related. Even though the assumptions of this approach are not justified (e.g., linear combinations, independent predictors, and that trust is a value) Lee (2013) used a more sophisticated but similar machine learning approach to examine the factors in a human-robot interaction that led to greater trusting behavior, specifically predicting the outcome of an economic game (rather than subjective trust.) These trust equations are useful insofar as they represent tacit assumptions in many of the more complex trust and trustworthiness models, and thus allow empirical testing of those assumptions. However, it may be a blunder to take them as a serious theory of trust per se, but rather treat them as a interpretive model of the processes and influences of trust.

5.5 Contributors to trust and trustworthiness

As we have seen, many of the models of trust are really catalogs of factors that impact trust (if they are about the trustor) or trustworthiness (if they are about the trustee/automation or AI), as well as external/contextual contributors ((Hancock et al., 2021); see discussion in previous section). These factors, such as perceived trustworthiness, the AI system’s ability to perform tasks predictably, the benefits and costs of trusting these systems, and legal regulations, are essential in facilitating the design, implementation, and assessment of trustworthy AI systems.

Perceived trustworthiness is a significant contributor to both social and human-machine relationships. In social relationships, individuals rely on various cues, such as past experiences, communication style, and reputation, to assess another person’s trustworthiness. Similarly, in human-machine relationships, users assess the perceived trustworthiness of AI systems based on their experience with the technology. Trust in machines is frequently shaped by considerations such as *accuracy*, *consistency*, *reliability*, and the system’s ability to *consistently deliver expected outcomes* (Adams et al., 2003). When

users perceive an AI system as trustworthy, they are more likely to rely on it and maintain a positive relationship with it.

The AI system's ability to perform tasks predictably is another critical factor influencing trust in human-machine relationships. Humans value *predictability* in their interactions with AI systems, as it allows them to understand and anticipate the behavior of the technology. When an AI system consistently performs tasks as expected, users develop a sense of trust in its capabilities (Lee and See, 2004). On the other hand, uncertainty and unpredictability can erode trust. Therefore, designing AI systems that prioritize predictability by setting clear expectations and providing transparent feedback enhances trust in these relationships.

Trustor-related contributors, specifically the benefits and costs of trusting AI systems, also affect trust in human-machine interactions. Trustors evaluate the potential advantages and disadvantages associated with relying on an AI system. Benefits such as *efficiency*, *accuracy*, and *convenience* can increase trust as users perceive that the technology improves their overall experience. However, if the costs of trusting AI systems, such as privacy concerns or the potential for errors, outweigh the benefits, trust can be undermined. Therefore, a thorough understanding of the trade-offs involved in trusting AI systems is vital to fostering trustworthiness (Lee and See, 2004).

Contextual contributors (Hancock et al., 2021), including legal regulation, also influence trust in human-machine relationships. Laws and regulations provide a framework that governs the behavior and actions of AI systems, ensuring accountability and transparency (Adams et al., 2003). Legal safeguards can enhance users' trust in AI systems by assuring them that their rights and interests are protected. The presence of legal regulations can establish a strong foundation for building trust in the technology, as users perceive it to be governed by ethical and responsible practices.

5.6 The flavors of trust

5.6.1 Mistaken assumptions about trust and trustworthiness

Hoffman (2017) argued that trust, trustworthiness, and reliance are often mistakenly assumed to be an end-point stable state, which is a mistake. Furthermore, he criticized the assumption in many trust calibration models that it is a single value measured on a single scale, arguing *trusting* is a process, there are different kinds of trust, and it depends on context. These mistaken assumptions work together to form what we will refer to as a control-theory of trust—trust is a unidimensional continuous value that is a holistic property of a system that applies in general to its use. This involves four specific assumptions that are unwarranted:

Trust is assumed to be unidimensional. Instead, trust can be pluralistic.

One common faulty assumption is that trust is unidimensional—that trust is a single property of the relationship between two entities (Hoffman, 2017). As we will discuss in Chapter 6, most measures of trust assume trust is pluralistic—many different aspects of trust exist, so that trust, reliance, understanding, adoption, and other aspects all make up one's trust state (Hoffman, 2017).

Trust is assumed to be continuous and metrical. Instead, trust can be categorical.

Many models of trust implicitly or explicitly take trust to be a continuously-varying concept—you can trust something to a larger or lesser extent (Hoffman, 2017, also). This is implicit in the notion of “trust calibration” and implies a control-theory view of the trust process. It is true that people can rate their level of trust at some finer level, but it may not accurately describe a behavioral version of trust where trust is indicated by cooperating or defecting. When he offered Jasmine his hand to get on the magic carpet, Aladdin famously asked Jasmine “Do you trust me?”, not “How much do you trust me?”. By encouraging one to consider trust as finely-graded, it may avoid thinking about the other ways in which trust is not all-or-none (contextual and componential trust).

Trust is assumed to be global. Instead, it can be contextual. Another mistaken assumption is that trust (or perhaps trustworthiness) is universal. Instead, it is bound to a context—it can depend on when and where and the environment the trust relationship occurs. For example, new drivers are trusted to operate a vehicle during daylight hours but not at night; automated drivers may be trusted in low-traffic and well-mapped areas but not on snowy roads or in off-road settings. Here, trust is a three-part relationship: A trusts B to do C (Baier, 1986; Hardin, 2002). You (A) trust your vehicle autopilot (B) to drive on snow (C).

Trust is assumed to be holistic. Instead, it is componential or mereological. Finally, trust and trustworthiness are assumed to be holistic: your trust state applies to the entire entity you are trusting. Instead, trust is *componential* or *mereological*—you might trust one part, function, or property of a system differently from another. Models such as Muir (1994) explicitly suggest that the trustworthiness of some of these different properties compose the trustworthiness of the entire system, but it can be much more useful to understand what someone trusts, and what they do not trust about a system. Furthermore, users may be forgiving for errors they can explain if they know that the untrustworthiness of a system is limited to specific functions.

Altogether, the control-theoretic trust calibration model implies that an appropriate level of trust is an end-state of system design and learning. However,

an alternative perspective does not ask “how much should I trust the system”, but rather asks “what aspects, how, when/where and for what should you trust the system”. This is illustrated by the comments of one Tesla FSD driver interviewed for the research described by (Ibne Mamun, 2023; Ibne Mamun and Mueller, 2024).¹ He said that there were things he could trust the car to do 100% of the time, and these are things he would allow it to do whenever he operated the vehicle. But there were things that it would do correctly less often and he would take over or avoid those situations. And there were things that were even less reliable, and those he would only try when his wife was not in the car. Here, trust is about different things (context, multiple dimensions, and different components), and levels of trust are really categories about probabilities of success, (categorical and not numerical). Furthermore, for him, lack of trust did not mean he would not rely on it, but reliance depended on the context of his passenger’s risk tolerance.

This contingent notion of trust has been considered extensively in the past, as we will cover in the next section.

5.6.2 Appropriate, calibrated, justified, and warranted trust

“Appropriate trust” Lee and See (2004); Hancock et al. (2021) might be the most general way of discussing how our trust in a system should somehow correspond to the trustworthiness of that system. However, the term *trust calibration* has been used for decades (Muir, 1994; Llinas et al., 1998; Seong and Bisantz, 2000; Adams et al., 2003; Lee and See, 2004) to describe how trust might be matched to trustworthiness. This notion arguably incorporates the tacit control-theory assumptions that trust is holistic, uni-dimensional, continuous, and global, which we argued against in the previous section. As an abstraction, this can be helpful to consider processes by which trust and trustworthiness align. It was used by Lee and See (2004) to help guide their model of trust, along with *appropriate trust*, *overtrust* and *distrust* to describe different trust calibration regions, with the prediction that overtrust can lead to misuse, and distrust can lead to disuse. However, they also distinguish between *calibrated trust* and *appropriate reliance*—the appropriate use, adoption, or deployment of a system based on its trustworthiness. Ultimately, reliance is more important—trust is a stance about a system, but reliance is actual use.

Nevertheless, the control-theory metaphor begins to fail when we consider contextual, componential, multidimensional and categorical aspects of trust. Hoffman et al. (2013); Hoffman (2018) proposed an alternative approach: *justified or unjustified trust and mistrust*. The similar notion of *warranted or unwarranted* trust was used by others, such as Jacovi et al. (2021). In this framing, trustworthiness is not a single value that needs to be adjusted to

¹This quote was secondary to that research and did not appear there directly

match. Moreover, a single automation or tool can be both trusted (about some things) and distrusted (about others). This suggests a more qualitative analysis in which one examine which aspects are trusted and distrusted, and whether these done for good or bad reasons. This can helps focus human factors researchers on contextual, componential, and multidimensional aspects of trust. Instead of asking “how much do you trust the system”, you may discover what parts of the system are trusted, at what times and in what contexts, and why.

5.6.3 Hoffman (2017) Taxonomy of *trusting*

Along with justified and unjustified trusting, Hoffman (2017) identified many other kinds of “positive” trusting relationships, including absolute trusting, tentative trusting, skeptical trusting, contingent trusting, stable trusting, progressive trusting (trust improves with experience), digressive trusting (trusting deteriorates with experience) swift trusting, and default trusting, faith-based trusting, authority-based trusting, and over-trusting (i.e., automation bias). However, there are also *negative* varieties of trusting: mistrusting, distrusting, anti-trusting (belief that the machine will do things not in the human’s interest), and counter-trusting (machine is providing evidence that it cannot be trusted). These underscore the variety of trust situations that can occur, and especially highlight how trust is often dynamic, both in how it begins, and how it changes with experience and time.

Many of these might just be thought of as dynamic patterns that might emerge in under the control-theory perspective on trust: initial set points that occur for many of the reasons the trust calibration models cite for impacting trust or trustworthiness (e.g., as Hancock et al., 2021, argue, things related to the trustor, the trustee, and their interacting context), along with dynamics over time as new information is learned, systems exhibit strengths and weaknesses.

5.6.4 Ways to impact trust

Default trust.

As with humans, some researchers have suggested interacting with AI starts with some amount of default trust. Unlike human-human interactions, human-machine default trust does not consider that trustworthiness hinges on the partnering agent being a computer (Smith et al., 2017). This can be a degree of trust or distrust depending on the individual. At default, the user may have zero expectations of a system and have little-to-no reason to expect it to work against expectations. To this end, the user may begin using the system with *blind trust* until there is a rapport built over time. With experience in dealing with intelligent systems, a person can also start with some default distrust in the system and expect it to under-perform the user’s expectations. As previously

mentioned, trust is dynamic and so is default trust; not all users start interacting with a system and expect it to perform perfectly.

Changes in trust through experience.

Most models of trust suggest that justification (or calibration) of trust comes in part the form of repeated interactions that build a sense of trust/distrust. To make a proper justification, the user must provide a series of test cases to determine if a system can provide adequate output.

Expertise in trust.

In the extreme, extended experience can build expertise. Expertise with a device should increase trust in the system though understanding of how it works (Ferrario et al., 2020). As the operator spends time working alongside an autonomous system, they should gain an understanding of its capabilities and modify trust accordingly. A system that cannot provide consistent and predictable output will likely not be trustable.

Transparency and explainability.

Explainable AI (Chapter 8) can also increase trust in a system and provide a framework for justifying trust or distrust (Akula et al., 2019). If a system can provide justifications of its actions and the user can interpret these explanations the user might trust the system in similar scenarios. As with humans, prolonged use can increase or decrease trust and one extraneous event should not ruin the user’s expectations of the system; one failure would not destroy the foundation of trust in a system.

5.7 Trust in “Artificial Intelligence”

As Chapter 1 emphasizes, AI systems vary widely in capability and role. Much of the research and models of trust were developed on older automation systems such as power control systems, air-traffic control automation, commercial air pilot automation, health care applications including warning/alert systems, and factory automation. AI can often serve the role as automation—albeit with more complex algorithms. In those cases, many of the same concerns apply. Furthermore, even when AI is not being used to automate a process and rather augment cognitive work, many of these same issues exist. The nature and complexity of modern AI systems (including generative AI, large-language models, medical diagnostic systems, autonomous control systems, etc) means they can have many more functions than most simpler automation systems

involved. Determining the important aspects of trust in AI is critical as it becomes more prominent in society and new AI models and systems continue to emerge. We will first describe the fast-growing domain of research focusing on “trust in AI” in particular, without much of the theoretical or methodological traditions we have explored in the trust in automation domain. Then, we will examine how the fields that emerged to study trust in automation embedded into work processes (e.g., cognitive systems engineering, human-centered AI) have assimilated modern AI systems into their research. Finally, we will briefly review the emerging legal, ethical, and societal concerns about AI that go by the term “trustworthy AI”.

5.7.1 Trust-in-AI research in comparison to trust-in-automation

Although AI systems of various types have been in the field for almost 50 years, the recent growth in capability of AI has generated research on AI specifically as a target of trust. Prior to the wide adoption of large language models, Glikson and Woolley (2020) reviewed empirical research on human trust in AI across management, HCI, and engineering. They organized their findings along two dimensions: how AI is represented (embedded in software, virtually embodied as an agent, or physically present as a robot) and how much autonomy it exercises (as a tool, assistant, or manager). Trust requirements and antecedents differ across these: embedded systems involve perceived performance, transparency, and control; virtual agents require calibration between anthropomorphic presentation and actual capability; and physical systems on depend on responsiveness and shared situational awareness. They argue that expectation–performance calibration matters more than a single trust score, and that lab studies often show optimistic initial trust fails to replicate when fielded, such as for embedded and robotic AI. Their analysis is relevant to modern generative AI systems, as they fall within the embedded-to-virtual range and the tool-to-assistant autonomy region. One prediction of this framework is that, for these systems, higher autonomy and less visible failure modes will require justified, contextual, and componential trust rather than blanket holistic reliance.

The literature specifically labeled “trust in AI” has grown quickly since that review. Some of this research has discarded or been ignorant of the theories, models, or measures developed for automation and focused on new and unique aspects of trust-in-AI. Afroogh et al. (2024) reviewed definitions, challenges, and open questions and concluded that the field still lacks a shared construct: papers mix trust, trustworthiness, acceptance, and ethical “trustworthy AI.”; and these are probably clearer in the trust-in-automation domain.

Another shift in focus in this field has been in research on general trust in AI in comparison to interpersonal trust. In light of the discussion in Chapter 4 about anthropomorphism, it may be the case that trust in AI chatbots becomes

more like interpersonal trust than has been found for trust in general automation. However, Montag et al. (2023a, 2024) found that self-reported general trust toward humans (i.e., as a personality trait) and trust toward AI (both generally and with respect to specific classes of products) were essentially uncorrelated, and even found different brain activation patterns for these cases. This suggests that default trust stance toward other people does not carry over to AI systems, although Gillath et al. (2021) showed that attachment-related individual differences can influence trust in AI, and that experimentally priming attachment security can increase rated trust in AI. In any event, these results neither support nor refute the assumption that trust in AI is built on the same sorts of cognitive and psychological mechanisms and thought patterns as interpersonal trust. Consistent with many of the models of trust we reviewed here, trust is not a blanket stance toward the target, but is about distinct functions and activities, within a context, in service of a goal. These results merely show that default stances toward others and toward AI are unrelated, even though there appears to be an attachment-related common pathway.

A second shift is that a large fraction of “trust in AI” research is really about *acceptance of AI as a product*. Even though one of Muir’s (Muir, 1994) primary concerns was about whether the “supervisor” (i.e., operator) would accept or reject the automation, this was normally in the context of a system that had automation embedded within it, but the user had autonomy to use it or do the work themselves. Much of the recent work on trust in AI is framed within a technology acceptance model, which normally is used to follow products through initial contact where potential users become aware of the product and may recognize its perceived usefulness, which may lead to eventual adoption. For example, Choung et al. (2023) had users rate trust into in facets: human-like facet (benevolence and integrity) and functionality (competence and reliability), and examined these facets as precursors to in a technology-acceptance model. Both facets predicted intention to use, although functionality was a more important predictor. Gillespie et al. (2023) asked how much people trust “AI” in the abstract (often by sector: health, hiring, banking). Those items captured general willingness to trust hypothetical systems, helping understand the public’s willingness to consider these applications in the abstract, and were not direct assessments of AI tools.

A final recent result about trust in AI showed that by 2025, users had begun to develop a distrustful stance toward GenAI Li and Aral (2025). Participants saw *identical* answer text to queries framed either as a traditional search result or as generative AI. The GenAI label lowered trust and willingness to share on average. They also investigated several interface features. Adding reference links raised trust and sharing (for both valid and invalid/hallucinated references), and color highlighting indicating model uncertainty *lowered* trust regardless of whether the highlights indicated low confidence or a mix of high and low confidence. Finally, a message that “users found this helpful” raised

trust. These results show that people were influenced by many superficial and even fake cues of reliability (citations, social proof, AI-rated certainty), indicating weakly-justified and unjustified trust and distrust appear to influence use of AI.

However, many researchers in this domain do recognize that trust in AI is really a new name for the same issues that emerged with trust in automation. The main difference is a shift in focus toward more cognitive/informational aspects of trust that modern AI systems entail, and a focus on new settings and AI applications (chatbots, generative search, medical advice) rather than completely new theories. For example, Kaplan et al. (2023) conducted a meta-analysis empirical studies similar to Hancock et al. (2021), examining factors related to the human, the AI, and the context—a common set of divisions in automation-trust models. Also, Everett et al. (2026) framed AI trust in the same vocabulary as interpersonal and psychological principles (risk, vulnerability, evidence, and social inference), which may be an evolution of some of the factors influencing trust and trustworthiness of automation, but is not a novel theory of trust.

Consequently, because of its more commercial nature, its cognitive focus, and its information-rich setting, and the more human-like intelligence embodied by modern AI, research on trust in AI has explored new questions that did not interest researchers involved in trust in automation. However, the basic frameworks developed in automation contexts are largely applicable.

5.7.2 Trust in AI in cognitive systems engineering and human-centered AI

Expert users of automation and cognitive systems have long understood that trust is not all-or-none. In high-stakes environments, operators must rely on systems they do not fully trust, because experience teaches them what each system is good at and what it cannot be trusted to accomplish (Lee and See, 2004; Cohen et al., 1997; Hoffman, 2018; Ferrario et al., 2020). Consequently, many of the factors involved in human-automation trust (e.g., Parasuraman et al., 2000; Sheridan and Parasuraman, 2005) are likely to recur for modern AI (Glikson and Woolley, 2020).

However, new challenges are emerging. General-purpose AI systems are increasingly embedded in organizational and cognitive work, including legal research, programming, accounting, documentation, planning, design, analysis, and trading (Brynjolfsson et al., 2023; García-Peñalvo and Vázquez-Ingelmo, 2023; AbuMusab, 2023). In many cases, sensitive or proprietary data must be sent to external services, so users must trust not only the model but also the vendor: that data will not be used to train models that will be used by competitors, it will not be resold, or exposed through weak security (Jacovi

et al., 2021; Shneiderman, 2020, 2022; High-Level Expert Group on Artificial Intelligence, 2019; Tabassi, 2023; Smuha, 2019).

Another unique challenge of large language models is that they cannot be trusted to match reality without supervision. Hallucinations are well documented (Zhang et al., 2023), and visible errors can rapidly erode reliance even when the system is otherwise useful (Burton et al., 2020; Schemmer et al., 2023; Lee and See, 2004). The hallucination problem is also pernicious because it can impact any aspect of an LLM's output (although they appear more frequently in some conditions than others), unlike automation where the trust issues are often encapsulated and relevant to certain contexts. Moreover, their training is broad but may be out of date, domain-incomplete, or poorly aligned with local facts, producing confident but false outputs (Zhang et al., 2023; Endsley, 2023). Trust in LLMs is also *literally* contextual. Models operate within a finite token *context window* (Brown et al., 2020), and performance degrades when relevant information falls in the middle of long inputs rather than at the beginning or end (Liu et al., 2024). Instructions given early in a session may therefore “fall out” of context, and the system will no longer follow them—another reason trust must be scoped to task, session, and domain rather than attributed holistically to “the AI.”

Complex AI systems are increasingly responsible for ground and air vehicles, and in some settings for identifying objects and controlling weapons (Parasuraman et al., 2000; Tate et al., 2016). Although autopilots and guidance systems are not new, extending AI across the full sense—decide—act chain raises the stakes when trust is miscalibrated (Waytz et al., 2014; Linja et al., 2022; Hancock et al., 2011). Users may trust automation for routine driving yet distrust it in edge cases they have learned to anticipate (Linja et al., 2022). More generally, historical automation was often narrow and tightly governed. Integrating more general AI into complex work systems can disrupt processes in unanticipated ways (Klein et al., 2023; Endsley, 2023; Brynjolfsson et al., 2023). Adoption therefore requires justified, contextual trust in what each system is trustworthy for—and appropriate reliance rather than blanket trust or blanket refusal (Lee and See, 2004; Schemmer et al., 2023; Wickens et al., 2015; Shneiderman, 2022). Furthermore, the challenges of trust in AI are a consequence of performance issues, rather than the cause. Hallucination and over-stepping capabilities requires human workers to shift from creating knowledge to supervising and checking knowledge creation, but it is not because of distrust specifically—it is because of demonstrated limitations of the system.

5.7.3 “Trustworthy AI” as a legal standard

So far, our notion of trust and trustworthiness has focused on human factors and system design. With the rise of highly-capable AI systems, standards and governmental bodies have increasingly used the term “Trustworthy AI” to

describe properties AI systems should exhibit in context of legal, societal, and ethical concerns. We will cover these issues more completely in Chapter 9, but cover it briefly here as well.

In its 2019 plan for federal engagement in AI standards, the National Institute of Standards and Technology identified seven core characteristics of trustworthy AI technologies: accuracy, reliability, objectivity, resiliency, security, explainability, and safety (National Institute of Standards and Technology, 2019). Stanton and Jensen (2021) extended this taxonomy to nine characteristics by adding accountability and privacy, arguing that these properties are necessary (though not sufficient) for perceived trustworthiness. Similarly, the EU High-Level Expert Group on AI developed a framework for trustworthy AI (High-Level Expert Group on Artificial Intelligence, 2019; Smuha, 2019). It involves three general areas (lawful, ethical, and robust AI), with specific properties including: respect for human autonomy, prevention of harm, fairness, explicability, the role of human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity and fairness, societal well-being, and accountability. All of these characteristics involve some aspect of the human-AI relationship, as the level of interaction that the human has with the AI system impacts how much they trust using the system and the development of other AI systems in the future. Many of these notions were also incorporated by Shneiderman (2022) in his Human-centered AI Trustworthiness Scale. This proposal allows a system to be 'scored' based on factors related to the development process, such as whether the software was validated, internal/external review was completed, and fairness was tested.

Jacovi et al. (2021) attempted to formalize the nature of trust between humans and AI. They argue that although interpersonal trust can be used as a basis for understanding trust in machines, machine trust relies on understanding risk, and entails the need for mitigating those risks, and the ability to anticipate the behavior of the AI is critical. As such, they develop the notion of *contractual trust*. This notion prescribes that trust in AI should be framed as trust in a commitment to a contract, which also requires explicitly identifying the performance expectations of the contract. This notion is similar to Tate et al. (2016) notion of licensure, although the contracts in licensure frameworks may be backed by legal or organizational policy.

Nevertheless, the notion of trustworthy AI has its critics. For example, Ryan (2020) provides an extensive discussion defining the ethical and legal perspectives on trust in the AI context. Although he mostly ignores the extensive research on trust in automation we have covered here, one important insight calls into question whether trust can really apply to artificial intelligence at all. He argues that according to a normative account of trust (i.e., what one "ought to" or "should" do in such a trusting relationship), the trustee is held responsible for its actions. However, legal discussion has rightly challenged whether an artifact can bear responsibility for an behavior at all—this falls on

its designer, developer, or vendor. For example, according to a normative notion of trust, whoever is operating the vehicle *ought to* bear responsibility for causing an accident. For automated vehicles, the vehicle itself can not bear responsibility, and it should fall on the designer or the operator who is not actually driving (but may have elected to allow it to operate). In this sort of case, it was probably predictable that when Tesla Autopilot and FSD-enabled vehicles caused high-profile accidents, Tesla (who likely bears responsibility based on a normative model) attempted to shift the blame onto the vehicle drivers, who were *not driving*, as if they were saying “They weren’t *not driving* the car correctly.” Because of this, Ryan (2020) argued that what was often mean by trustworthiness was really *reliability*, which indeed should be the focus, and the field should emphasize the trustworthiness of the organizations and people using and creating AI.

One possible resolution to this is the contractual and licensure notions discussed earlier (Tate et al., 2016; Jacovi et al., 2021). Since responsibility entails identifying blame for violations, and ultimately punishment, we already manage this with licensing, indemnification through insurance contracts in many situations. You cannot put a vehicle in jail for violating laws, but just as with human drivers you can (1) remove its license to operate; and (2) require financial payments to victims.

For the Human Factors Practitioner

This chapter demonstrates that the notion of trust, which seems like a simple and intuitive concept, is nuanced and has many implications when considered within systems, organizations, and across applications. Some common practical lessons from this include:

- Human-machine trust appears to be built on highly-ingrained notions of interpersonal trust that emerged at a evolutionary timescale and are supported by specific neurophysiology. Consequently, trust is not just cognitive or information-based, which can be exploited or influenced by superficial features of information.
- Trust is generally secondary to reliance. Users will adopt technology that is useful or necessary even when they do not trust it; and trust assessments do not always predict performance. Human factors researchers and engineers should not simply find ways to improve trust, but consider how trustworthiness (and distrustworthiness) should be made apparent, managed, and designed to allow users to work understand and acknowledge.
- One primary purpose of understanding and characterizing trust is to manage change in work practices.
- Trust and trustworthiness go hand-in-hand, linked to the trustor and the trustee.
- It is a mistake to think about trust in a system as a holistic, continuous, unidimensional, global concept. Instead it can be componential, categorical, multidimensional, and contextual. A more useful and practical analysis is to understand *what, where, and when* a person trusts a system.
- Models of trust are typically models of the factors that influence trust, and these impact the trustor/user, the trustee/system, their interaction.
- Good human factors design should not simply improves a user's trust in the system. Good design helps a user form a justified trust and distrust, not just globally, but about components, functions, context, and activities. A useful system whose limitations are understood can still be helpful.

5.8 Conclusions

Trust is among the most frequently used constructs in human–automation and human–AI research, yet it is rarely a single thing. Across domains, definitions converge on a common set of ideas: a trustor accepts vulnerability in a relationship with a trustee under uncertainty, in pursuit of some goal that will produce a (normally non-zero-sum) benefit. The chapter began by surveying

how trust has been framed—as expectation, attitude, belief, state, willingness to accept risk, ongoing process, or observable reliance—and contrasted trust with trustworthiness as a property of the trusted party (or the organization behind it). These distinctions matter because researchers and practitioners often measure one while meaning the other.

Much of the automation literature has drawn on interpersonal trust as a starting analogy, and for good reason: social cognition, anthropomorphism, and default dispositions shape how people encounter new systems. But the analogy is incomplete. Human-machine trust must also be understood through system cues, mental models, feedback from use, and the organizational and legal context in which a system is deployed. The models reviewed here—from cue-based and lens accounts to system loops, argument-based trust, and computational games—differ in emphasis, but address the same practical question: how evidence about a system’s capabilities and limitations comes to support or undermine trust and reliance.

A recurring mistake in both research and design is to treat trust as a holistic, continuous quantity to be maximized. Instead, trust is often pluralistic, contextual, and componential: one may trust a system for some functions and not others, in some environments and not others, in some ways and not others, and for good or bad reasons. Concepts such as *appropriate* or *calibrated* trust, *justified* or *unjustified* trust, and *warranted* or *unwarranted* trust are therefore normative standards on *reliance*: whether a person’s willingness to depend on a system matches what the system (and its operators) can defensibly be counted on to do. Transparency, experience, and expertise can shape that judgment, but they do not automatically “improve” trust. The goal of transparent system design should be to help someone trust a system only when they should.

Finally, contemporary AI systems, and especially general-purpose AI systems such as ChatGPT, implicate many of the concerns observed in research on trust in automation, and add new ones. Broader competence claims, opaque failure modes, vendor-mediated data use, and regulatory frameworks for “trustworthy AI” all widen the gap between perceived capability, organizational promise, and justified reliance. For human factors work, the practical implication is to understand the scope of trust explicitly: *who* is trusted, *for what task*, *under what conditions*, and *with what accountability when reliance fails*. Measurement and assessment are the next step: Chapter 6 examines how trust and related constructs have been operationalized, and what those instruments imply—and sometimes assume—about the nature of trust itself.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction, Michigan Technological University, Spring 2024.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Walker, S., Matas, E., Ulinski, K., Woolman, B., & Mueller, S. T. (2026). Trust in automation: Definitions, models, and related concepts. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Measuring Trust in Automation and AI: Key Points, Methods, and Scale Applications

Yifan (Jeff) Wang, Blade Frisch, Joanne Wang, Lauren Sprague, and Shane T. Mueller
Michigan Technological University

Abstract

In automation and AI trust plays a pivotal role in ensuring smooth interaction and effective use of technology. This chapter covers many different ways trust has been assessed and measured. First, we discuss different contexts in which trust is measured, including field studies of real or simulated environments, simulators, microworlds, wizard-of-oz simulations, and experimental research. Second, we describe empirical methods for characterizing or quantifying trust in these contexts, including interviews, self-report methods, behavioral methods, and physiological indices methods. We then identify many of the existing self-report questionnaires and scales that have been used, looking at a one directly to understand how it measures trust. Finally, we examine new scales and measures that have been developed to assess trust and trustworthiness of AI applications, both as a specific attitude toward a system, and as a general disposition to different technologies. This provides a practical foundation for researchers who are interested in evaluating trust in automation or AI, and want to identify their best approach.

6.1 Introduction

Chapter 5 discussed definitions, theories, models of trust in AI and automation, with the perspectives that trust is multi-dimensional, contextual, componential, and can be both categorical and dynamic. Assessing or measuring “trust” can thus be complex and difficult to carry out. Often, the go-to solution amounts to asking a user a few questions, such as “On a 1 to 10 scale, how much do you trust the system.” Although this can be useful, we hope the take-away from this chapter is that—due to its nature—there may be many distinct ways to measure trust, and some of these may be more effective at understanding the human-machine relationship than a simple rating scale.

In fact, there are many methods and strategies for evaluating and assessing trust—some are alternative measures, others are ways of embedding these measures within measurement contexts. These range from quantitative surveys assessing subjective perceptions to experimental setups analyzing behavioral and neuropsychological response. Of course, even when one simply wishes to assess trust with a questionnaire, there are many options to choose from, many of which are rooted in research on trust in automation. Recently, scales measuring general and specific trust *in AI* have emerged, and we will discuss many of these, and how they are either rooted in or distinct from previous measures of trust in automation.

6.2 Methods for measuring trust in automation

Chapter 10 of Adams et al. (2003)’s comprehensive technical report on trust provided an exhaustive picture of how trust in automation had been measured up until that point. They first identified research *approaches*—the context in which measurement happens (field, simulator, microworld, interview, experiment)—then covered measures, examining trust as a *psychological state* (ratings and scales) and trust as *behaviour* (defensive monitoring and use of/reliance on automation). Kohn et al. (2021) more recently reviewed trust-in-automation research and similarly identified self-report, behavioral, and physiological measures, each mapped onto a step in Mayer et al. (1995)’s process (perceived trustworthiness, propensity, trust attitude, or risk-taking), providing new cases of trust assessment that had become popular over the previous 20 years.

Table 6.1 summarizes these measurement contexts, which we augment with two other contexts that are perhaps unique special cases. We will discuss several of these contexts in more detail in the next section. After that, we cover many of the measures used to characterize or quantify trust in Table 6.2, which is a combination of both reviews (Adams et al., 2003; Kohn et al., 2021), and incorporates and expands on these to some additional approaches.

6.3 Contexts of trust measurement

Table 6.1 identifies a number of contexts in which trust can be assessed. These range across a spectrum of fidelity to the real work context, with trade-offs as you move from live study of a real system in the field to scenario-based surveys. Study context is often not a choice one makes, or is very constrained by access and resources. It may seem that ideally, you should understand trust in the context of the actual work-system. However, this may not be possible if the system is in a planning phase, is government classified or proprietary, dangerous to operate, requires specialized expertise, or is only used in unique circumstances or infrequently. In these conditions, a simulator, a scenario, or studying another existing surrogate system may be a good alternative. Even with access to an existing system, alternatives may be preferred—one may not be able to do live questionnaires of an aircraft pilot during use of a combat system, but could do so either in a lower-fidelity simulation, or in a video scenario where they watch a maneuver and engage in a think-aloud protocol.

Most of these contexts are self-explanatory. They each represent a different point on the fidelity spectrum, which offers trade-offs for studying trust. This trade-space not only represents verisimilude, it also can represent different levels of risk and responsibility that is an important part of most notions of trust. That is, you can use a simulator or a scenario to understand when someone would allow an automation to operate, but if the consequences of failure in the real-world involve real danger, financial loss/gain, reputational damage, health consequences, or equipment destruction, the low-stakes environment may not accurately measure what trust in the real situation would be.

Another aspect of the context is the level of experience of the user or research subject being studied. From an experimental research perspective, the user should be sampled from the same population that one hopes to generalize to, but often the trust relationships that are most interesting are for experienced users who have expertise and know-how about a domain, and are determining whether to trust a new system. Training-to-criterion and simplification of systems can be used if only inexperienced novices are available for testing, but this can also limit validity. A compromise might be to study expert users of surrogate system, gaining realism in the research subjects but losing direct application to the target system. For example, one may not be able to test military UAV operators because of access and security reasons, but a civilian hobbyist may suffice for some evaluations.

Vignettes are a very lightweight and flexible approach to understanding trust. Using a vignette, you describe a situation and have the participant consider what choice or action, or assessment they would make. This can be framed in the third person ("What should the pilot do?"), or first-person ("what will you do?"). This can be used to test out possible interaction modes. For example, Alam et al. (2025) used vignettes to examine how participants

Table 6.1: Context of measurement for trust in automation and AI. Those marked with * covered in Chapter 10 of Adams et al. (2003).

Context	Description	Strength / weakness
Field study of real or simulated systems *	Operators in aircraft, plants, or other natural settings or surrogate systems.	High realism; little control of extraneous variables.
Simulators *	High-fidelity physical and dynamic replicas (aviation, navigation, process control).	Realism with more control than the field; expensive, and still not the job.
Microworlds *	Small closed-loop worlds that capture some complexity of process control without simulating one plant.	A trade-off between laboratory control and field realism.
Wizard-of-Oz (WoZ) studies	Useful in microworlds, simulation, or low-fidelity scenarios; involves a human playing the role of the AI or automation (Green and Wei-Haas, 1985; Dahlbäck et al., 1993).	Can simulate complex behavior not yet achievable by current AI.
Low-fidelity scenarios	Normally, written or video scenarios describing/showing situations with trust elicited via questionnaire or interview	Easy to implement but low fidelity means potentially low validity.
Laboratory experiments*	Simple games or gambling tasks; experimental manipulation of risk, reliability, or other context.	High control but easy to lose the operational meaning of “trust.”

would react if they were patients getting diagnosed by an AI system, and tested the effects of embedding elements of both emotional and cognitive empathy. Similarly, Zondag et al. (2024) used vignettes to study patient-physician trust. These are akin to “Wizard of Oz” (WoZ) approaches, but a true WoZ study would have a user interact with an AI controlled by a human confederate, and not a use a clearly hypothetical scenario.

Altogether, these different contexts and settings for studying trust permit assessing trust with different fidelity to real-world systems. However, when the setting and context of the assessment is determined, the actual assessment, measure or metric of trust still needs to be determined. We will cover many ways of doing this next.

6.4 Measures and metrics for characterizing trust in automation

There are also a number of measures and assessments of trust that have been used to characterize or quantify trust and trustworthiness. These are summarized in Table 6.2. The suitability of these are also often constrained by the application, experimental hypotheses, domain, and work context. Adams et al. (2003) described many of these, and Kohn et al. (2021) described a distinct but overlapping set. We have added several specific approaches that are either distinct or important special cases of other measurement paradigms. The table shows that there are MANY ways to characterize or quantify trust. Many developers or researchers simply use an existing subjective self-report scale (or develop/modify one for their own purposes). However, there are many additional measures that can be used, many of which might be more informative. Depending on the *critical phase*, it may be more important to characterize the issues related to trust (using interviews or open-ended feedback) than it is to quantify a subjective trust level (using behaviors, neuropsychological measures, or surveys).

6.4.1 Characterizing Trust Using Semi-Structured Interview Methods

We know of no dedicated “trust interview” or commonly-used Cognitive Task Analysis (CTA) method Crandall et al. (2006) for exploring trust in work systems, automation, or AI. CTA approaches are not usually used for *quantifying* trust (providing a number that you can compare or track), but are useful for *characterizing* trust; providing detail about what people trust, the context of trust, how they trust, and why they trust. Because these methods tend to focus on specific incidents, issues related to trust frequently emerge (deliberately or by happenstance), especially if the incidents are selected carefully. This can be done by choosing incidents in which human-machine trust or reliance was either needed or fell apart. For example, one might start by asking, “Tell me about a time in which you needed to trust SYSTEM X, and it let you down”. The responses from this approach can provide much better insights than simply a numeric value you might get from a survey. Although Adams et al. (2003) mentions interviews as a measurement context rather than measure, they made reference to Matthews et al. (1999), who used a CTA of Halifax-class operations-room officers as the source of their measurement framework, and discussed semi-structured interviews around a naval data-fusion prototype (Aldern, 1995).

Interview methods are commonly used in medical and health care research as well, for a wide range of purposes. For example, Hanna et al. (2023) used semistructured interviews and focus groups across three groups (medical cancer advisors, volunteers from cancer support groups, and people from the general public) to evaluate the trust people have in a medical chatbot, finding

that willingness to use a chatbot would depend on personal need, such that they would prefer it over searching online if a physician was not available. Kostick-Quenet et al. (2023) use in-depth interviews of medical professionals and patients who participated in high-risk procedures to broadly evaluate trust in AI, and determined that high accuracy was a critical desire among physicians. King et al. (2022) used interviews of pathologists to evaluate the trust of AI in pathology specifically, and found there was a desire for explanation from AI, while recognizing that current pathology practice can be a black box without explanation. Fernandes et al. (2023) used interviews of weight loss professionals as part of their development process of a weight loss tool. Their participants noted that having limited variables and an overly predictable output that does not take into account all of the factors the experts believed would impede their trust in the system.

We consider interviews as the elicitation method rather than a context, as they can refer retrospectively to events that happened either in the field or in any simplified simulation of that system. Trust-related information frequently emerges from interviews, and several published studies serve as useful examples and potential models for new interview methods. For example, Capiola et al. (2020) used CTA with intelligence operators in multi-domain command and control to characterize *swift trust* in ad hoc teams. They used fairly standard CTA methods, targeting the process by which ad hoc teams quickly developed trust in one another, and attempting to understand the precursors that led to its development. Although this research dealt with interpersonal trust, a similar approach could be used for human-machine trust by changing the basic incident selection and targets of the prompts.

Task analysis assessments of trust can also be used to improve design by improving system trustworthiness. Given a typical trust process and a trust-cue taxonomy, de Visser et al. (2014) give a procedure for identifying trust cues for a specific case:

1. Select a scenario that involves trust in a cognitive agent.
2. Conduct a task analysis to identify critical trust-related tasks.
3. Identify key pieces of information for operator decision making.
4. Verify pieces of information against the trust-cue taxonomy.
5. Construct a visual display representing this information.

This procedure provides a potential avenue for designing or redesigning an information display that is able to capture the important factors that lead to trust in users.

6.4.1.1 Using Critical Decision Method for Understanding Trust

The Critical Decision Method (CDM) Klein et al. (1989) is a popular CTA approach commonly used to understand the challenges for important decisions,

typically focusing on expert decisions. It works by selecting one or more incidents that serve as 'critical' decisions, establishing a timeline of events prior to and following the incident, and then engaging in several passes and retellings of the incident that focus on different factors, each time deepening the understanding of the incident. Klein et al. (1989) originally described a number of probes for these passes (e.g., cues, knowledge, aiding, time pressure, situation assessment, etc.), and trust issues often emerge during these additional passes. For example, focusing on knowledge and cues might uncover unreliability that would signal either trust or distrust, and the proper interpretation might depend on background knowledge. Klein et al. (2010) used this method to study developers of a complex city-wide simulation model that incorporated social, infrastructure, and economic factors. The simulation showed periodic dips in economic activity that would recover after a few days. When novices saw those dips, they inferred it as a cue that the model was broken and maybe should to be trusted. Experts who understood the modeling system better knew these dips were linked to simulated power outages that were caused by sabotage or disrepair. The same cue that appeared to novices to indicate unreliability demonstrated to experts that the system was working properly.

Although trust can emerge in some of the common CDM deepening passes, Brown et al. (2025) suggested using "How did you know to trust the information?" as a probe that would focus on trust specifically. This could be phrased in many ways depending on the person and the system being interviewed. To tailor the CDM to eliciting trust, we recommend the following as a starting point:

1. The incident should be selected to focus on a case where trust was involved, perhaps within routine operation of an automation system, intelligent machine, or AI.
2. Following the initial timeline, at least one pass can focus on how trust or reliance impacted the decision or outcome.
3. Rather than simply probing about 'trust', consider incorporating relevant factors from the trust models in Chapter 5 that are known to impact trust of the interviewee (past experience, expertise, personality, etc.), trustworthiness of the trustee (e.g., reliability, transparency, interactivity, capability, etc.), and the context (e.g., the team, organization, policies, etc.).
4. Rather than trying to elicit a degree of trust, recognize the contextual, componential, categorical, and multidimensional nature of trust. Ask "What about it did you trust?", "What did you trust it for?", "What part or function were you trusting?", and "When would you have trusted or not trusted it?".
5. To consider how the trust or distrust was justified, ask "Why did you trust/distrust the system?".

6.4.1.2 Lighter-weight CTA methods

The CDM can be quite time-intensive to use, and it produce long transcripts that require substantial resources to organize, code, and extract themes. Other methods are lighter weight, including Applied Cognitive Task Analysis (ACTA) (Militello and Hutton, 1998), which incorporates an approach called the *Knowledge Audit*. One of the standard probes of a knowledge audit is *Equipment difficulties* which is intended to target “Not trusting misleading instruments” by asking “Has the equipment pointed one way while your judgment said another?”. This is a fairly narrow focus on trust, and it would be feasible to incorporate several additional trust-related probes here, such as “Did you consider not using the system and doing it yourself?”, “What part of the system could you trust completely, and what would you be unsure about?”, “In what ways would you trust the system at this point?”. If trust is critical, a comprehensive “Trust audit” might be developed along those lines

Overall, these interview methods can be powerful ways to understand how and why, and in what context people trust the systems they use. They are not without drawbacks. They require access to specialized users, trained interviewers, and can require manual data coding. Also, they typically involve small numbers of observations, and so can be subject to sampling issues and interviewee biases. There are several other verbal-based methods for analyzing trust that bear discussion. These have some of the same strengths and weaknesses, but also offer distinct approaches.

6.4.2 Other Verbal Measures

Interview methods and many CTA approaches typically involve verbal information that is analyzed and processed for clues regarding trust.

Think-aloud protocols

A think-aloud protocol is a self-report method oftentimes considered a task-analysis interview method, but it is somewhat distinct in that it asks a user to verbalize their thoughts, goals, and actions as they perform a task. It can be revealing for many aspects of performance, although there are also limits into what people are really able to verbalize about their performance (Nisbett and Wilson, 1977; Ericsson and Simon, 1980). Although this is a potentially useful method, it must (like most behavioral measures) be accompanied by appropriate tasks that are likely to influence trust. Furthermore, to use such an approach, it is probably useful to provide instructions and prompts for users to focus on their feelings of trust or trustworthiness in a system, to help surface those attitudes.

This method sheds light on the person’s thought process and could be used to identify leverage points of trust in real-time as the user encounters them. A

retrospective protocol analysis can also be used, where participants complete a task and explain their thought process afterwards, possibly prompted by a video recording of the event. However, just like interviews, think-aloud protocols can generate a large amount of qualitative data that need to be analyzed and summarized—typically with similar qualitative research methods used in interviews—to identify issues related to trust.

This approach was used to evaluate theories of trust within a risk management setting by Earle (2004), as well as by Gautam et al. (2026) to investigate researcher trust in LLMs to perform different early-stage research tasks.

Discourse analysis

There is a body of work in the domain of pragmatic linguistics that has used discourse analysis to study trust in written or spoken language. Rather than examining transcripts or responses generated by users of a tool, discourse analysis often works on written public documents, and it describes a number of related qualitative research approaches where linguistic form and usage in those documents is examined for insights into a construct (e.g., trust). For example, Fuoli and Paradis (2014) examined CEO letters following the BP oil spill accident that were intended to repair trust in BP after the spill. Saragih et al. (2026) used this approach to study how AI was represented in online media, and identified themes related to trust. This approach can be used for both archival transcripts (e.g, cockpit communications logs, after-action review reports) as well as newly-designed studies where communication is logged incidentally or as the purpose of the study (for example, in a focus group, unstructured user interview, or user feedback).

Natural Language Processing and Sentiment Analysis

In contrast to discourse analysis which is typically performed by trained experts, there are a number of natural language processing approaches that use classification and clustering to extract attitudes and sentiment. Sentiment analysis is a technique that tries to characterize positive or negative attitudes within text, by counting key words, using latent semantic analysis, or other machine learning or AI tools. Although this does not assess trust per se, it can provide an understanding of whether social media posts, user reviews, traditional media, or small-scale communication involves positive or negative opinions, and further language classification might be used to narrow down trust and distrust specifically. For example, Guo et al. (2024) used sentiment analysis to infer trust relationships from online review data, and Niu et al. (2020) used a similar approach to identify trust in lenders in online comments.

6.4.3 Behavioral Measures

Human behavior can be impacted by their trust in a system in a number of ways—both direct and indirect. Here, the term appears as sort of a catch-all that is contrasted with non-behavioral measures (neurophysiological measures, text analysis) as well as verbal behaviors (interviews, questionnaires are technically behaviors). The most obvious behavior indicating trust is whether a person uses a system—which might be called adoption or reliance instead of trust, but there are other less direct behavioral indicators of trust.

Observational methods

A number of observational methods can be used, in which behavior is observed in real or simulated environments, and potentially without the awareness of the observed party. Although this class of methods is very broad, it includes some very useful techniques such as observing a first-time user of a tool (Mueller et al., 2009), or an observational usability study. Observational methods often occur in healthcare settings to understand many aspects of trust, including patient trust in processes, advice, physicians (Zondag et al., 2024; George et al., 2019; Paul et al., 2022; Riedl et al., 2023; Jongerius et al., 2022; Aiken et al., 2021; Orrange et al., 2021), AI patient-oriented tools (Stein and Brooks, 2017), and AI decision support (Sakamoto et al., 2024).

Defensive Monitoring

Muir and Moray (1996) argued that trust is negatively related to the need for monitoring automation. This “defensive” monitoring refers to a trust-related behavior that is undertaken to reduce the likelihood of negative outcomes if automation deviates from expectations and predictions. When automation is highly reliable, the need for monitoring reduces, but unreliable systems may require constant monitoring. For example, one might use an eyetracker to count eye movements to the roadway for an operator of a vehicle in autonomous mode, while the operator is doing some other task such as reading. Even if the operator never takes over driving from the autonomy, when a person is looking at traffic or the road ahead, we might infer that they trust the system less than periods when they are not looking at traffic. In other settings, this can be examined more directly through recording specific sensors displays that need to be activated by the user, or other means.

If using a monitoring behavioral paradigm, the “optimal sampling rate” normally needs to be established. Most forms of automation need a certain level of monitoring, and deviations from that might indicate overtrust or undertrust of the automation. Therefore, self-report scales and behavioral measures of defensive monitoring, such as the number of checks on automation, can be used to assess the required frequency of monitoring. Analyzing video logs or

digital records of instrumented systems can also be effective ways to measure defensive monitoring (Adams et al., 2003).

Use or reliance

The use of automation can be assessed based on the frequency of automation use, the length of each period of automation use, or the total proportion of time that the system is under automatic control versus manual control (Adams et al., 2003). In many situations, the use of automation may signify trust in the automated system. The method for measuring use of automation is through system logs that collect data on automation vs. manual control. Additionally, an alternative method is to analyze audio/visual logs to decide the use of automation. By observing and analyzing them, researchers will know whether individuals trust automation.

However, use and reliance should be interpreted carefully as a measure of trust. First, there can be many reasons people use or rely on a tool they do not trust. They may have no other choice, or the consequences of failure may be low, or they may simply trust the system more than themselves (the “authority hypothesis” proposed by Mosier and Skitka (1996)). Moreover, reliance or adoption is often understood as distinct from trust, which is often considered more prospective and about future behavior, not immediate performance.

Non-verbal communication, social cues, other gestures

Nonverbal communication and nonverbal social cues can include facial expressions, a set of expressions, gestures, and mimicry, posture, and other behaviors. For example, an eye roll in conjunction with a large grin may more accurately convey humor. In the context of measuring trust, these can be both deliberate (a gesture communicating to a team-mate) and subconscious (posture, agitation, expressions of anger). Although analysis of these cues have been used in many contexts, they have occasionally been used to specifically study trust, such as interpersonal ratings of trust ((Wood, 2006)), who found that non-verbal cues impacted trustworthiness ratings of salespersons. Similarly, Burgoon et al. (2021) provided a very comprehensive examination of nonverbal signals (including glare, voice, facial expression, body movement, speaking tempo, turn-taking, interruption), and found that longer and slower turn-taking was positively associated with perceptions of trust, but many non-verbal behaviors were not.

These cues have also been used to assess trust in social robotics. Lee et al. (2013) found certain nonverbal gestures like face touching, crossed arms, leaning backward, and hand touching were associated with lower levels of trust in a robot partner during a social game.

Interaction and social graph analysis

Another way to infer trust between entities is through interactions, or explicit surveys about trust relationships within a team (possibly including AI/technology entities). Interactions could be measured from communication logs, video logs or location sensors, and equipment usage records. For example, Krackhardt (1987) inferred ‘cognitive social structures’ by asking individuals within a team who they went to for advice; Zhou et al. (2023) used an interactive hybrid task analysis approach to create flowchart diagrams describing different team members in a human-autonomy team in a remotely piloted aircraft scenario. Mohammadi and Hashemi Golpayegani (2021) proposed a social-network–based model that infers trust. By using these task and social representations characterize interaction, a number of well-established network metrics can be deployed to understand team effectiveness, interdependence, and reliance.

6.4.4 Physiological and Neuropsychological Methods

Kohn et al. (2021) identified four groups of neuropsychological trust measures—electrodermal activity, heart-rate change and variability, eye-gaze tracking, and neural measures. They must be interpreted in the context of the task being performed, because the same signal also tracks load, arousal, and surprise, but they can be useful real-time or post-event measures that could either indicate timepoints where trust breaks down, or physiological precursors to trust or distrust. Typically, they should be correlated with other direct measures of trust to establish their validity in the context of study. Often, these measures are used together, along with both behavioral paradigms and subjective self-report scales of trust, alongside stress, workload, and mental effort. Gupta et al. (2020) described more than 10 studies that have used combinations of heart rate, GSR, eyegaze, EEG, and other physiological measures as proxies for trust. They note that cognitive load and trust are often related experimentally, and in many of the tasks that they reviewed (VR and other games), this may make sense—immediate cognitive load may prevent assessing the trustworthiness of a partner and so lead the two measures to co-vary.

Eye Tracking. Eye-tracking has been used to establish trust in automation as shown by Lu and Sarter (2019) in ways similar to defensive monitoring discussed earlier. They found that fixations increased in both duration and length in a low-reliability condition when compared to a high-reliability condition. Furthermore, participants who were primed to expect the system was high or low reliability altered their eyegaze behavior, which demonstrates that eyegaze can track not only trust level, but change in trust or trust calibration as experience is gained.

Heart monitoring. Heart rate changes are known to be sensitive to many aspects of skilled performance, as can the secondary measure of heart rate variability (HRV). These measures have also been linked to trust states, but these are often correlational results, and work must be done to establish both sensitivity (the measure is related) and specificity (it is NOT related to other explanatory factors). For example, Leichtenstern et al. (2011) found that when a website was not designed to elicit trust in the user, heart rate increased from baseline and post-test subjective surveys found that users were stressed when interacting with the low trust website. Gupta et al. (2020) used both heart rate and HRV as measures of trust and cognitive load, and established that a higher cognitive load while engaged with the VR indicated lower user trust in a voice-assist agent. This sort of research normally relies on manipulations of trustworthiness and subjective measures to establish sensitivity of heart rate. They likely cannot be used alone as a measure of trust, but need to be incorporated alongside other validating measures.

Galvanic skin response (GSR). GSR is another physiological measure that has been used to measure levels of trust in AI. But like heart rate monitoring, it also measures arousal, anxiety, stress and cognitive load, so specific features of GSR may be needed to distinguish among these psychological states. For example, Akash et al. (2018) used GSR together with EEG to find associations with human trust in machines. They found that mean GSR as well as GSR peak value were significantly affected by the amount of cognitive load required to interact with the machine and trust in the machine. However, the patterns of load and trust differed, even though higher load was associated with greater levels of trust. High cognitive load was associated with higher *peaks* in GSR while low amounts of human trust are associated with a higher *mean GSR*. Together with EEG, GSR could estimate a user's reported level of trust in the machine. Similarly, Hald et al. (2020) determined that GSR could be used to measure trust in a robot during human/robot team collaboration. This research found that GSR spiked when robots made unexpected movements and sounds and when measured alongside user's changes in physical activity to the robot, indicated loss of trust in the robot.

EEG and fNIRS. EEG (electroencephalograph) and fNIRS (functional near-infrared spectroscopy) are relatively non-intrusive methods for measuring brain activity. EEG uses sensors attached to the scalp and measures electrical activity, whereas fNIRS attaches a light sensor to the skull to measure blood flow. Research conducted by Eloy et al. (2022) found that using fNIRS to examine the human prefrontal cortex can elicit information about trust in human-agent teams (HAT) to such a degree that decisions in relation to trust can be predicted up to 15 seconds beforehand. This research found that the

dorsal medial prefrontal cortex, which is responsible for social impressions, activates when the user is determining default trust in the agent. During a condition designed to elicit suspicion in the human team member there was fronto-polar area (strategy), temporoparietal junction (moral decision making), and dorsolateral prefrontal cortex (planning and abstract reasoning) activation. Palmer et al. (2020) elicit mistrust from a human user in an automated driving scenario in which the autonomous system made poor decisions. An fNIRS sensor detected ventrolateral prefrontal cortex activation, which is associated with decision uncertainty.

Finally, EEG has been utilized to determine human controller trust in a situation where they are monitoring an automated system. In their condition designed to elicit mistrust, Oh et al. (2020) found increased gamma wave power, associated with increased focus and concentration, when interacting with the automated system. They also found that when the system is designed to increase trust, EEG detected increased alpha, associated with relaxation and contentment, and beta waves, associated with active thinking and wakefulness. These techniques can take advantage of spatial specialization of human brain regions to provide a more complex signal for detecting trust.

Overall, these physiological measures have advantages—they can record brain state without interfering in a user's activity, they can have high temporal accuracy to pinpoint exactly when trust changes occur, and they can provide additional evidence of different brain subsystems responsible for trust (e.g., emotional vs cognitive). However, they have limitations, including the cost and hassle of using them in comparison to other methods, as well as the typical lack of specificity in measuring trust. The measures developed may also be closely linked to experimental scenarios and manipulations, and may not easily generalize to other paradigms, or field evaluation of an actual work system.

6.4.5 Assessing Trust with Surveys or Questionnaires

Questionnaires are used extensively to measure trust in decision aids, automation, AI assistants, and the like. They are also often used in conjunction with other methods. As we saw earlier, they can be used to help validate behavioral and physiological measures. They are also sometimes coupled with interviews or focus groups to provide more detailed assessments. Although most trust scales involve repeated Likert-scale (1 to 7) responses, they can also include open-ended items that provide information more like one would obtain from interviews.

Serious attempts to measure trust as a subjective rating emerged by the 1990s, as research on trust in automation became popular and started to investigate the emerging problems and accidents occurring because of automation. These were sometimes just direct ratings of specific constructs, such as Muir and Moray (1996), who used self-report assessments of trust

and related constructs on a 1-100 scale to test a regression-style model of trust during a pasteurization simulator.

Kohn et al. (2021) noted that custom single-item “how much do you trust. . .” scales are the most common method in practice. Moreover, many studies develop ad hoc surveys targeting their own specific research questions. For example, Xiang et al. (2020) studied the attitude of the public towards AI in the medical domain using a questionnaire, designed to elicit different attitudes toward the use of AI in medical settings, including perceptions (efficiency, safety, validity, trust, and expectations), receptivity (willingness to use AI services), and demands (which health care functions need medical AI).

However, systematic psychometrically-sound scales of trust have been developed that are intended to be used generally (see Table 6.3 and 6.5). Distinct scales are often intended for different purposes (e.g., prospective, dispositional, or system evaluation), or are targeting different kinds of systems (technology in general, robots, AI). In Section 6.5, we will discuss these in much greater detail.

6.4.6 Summary of Trust Measurement Approaches

Many researchers and developers appear to view trust as simply a quantifiable subjective report that is obtained via a questionnaire. As we saw in Chapter 5, there are many perspectives on what trust is, and so it is not surprising that there are also many distinct ways of measuring trust. Questionnaire-based trust can be helpful, but other ways of evaluating the nature of trust may be more helpful for system design and evaluation. Furthermore, many studies have benefited from using complementary approaches to establish coherence and validity of measures. In all cases, understanding basic psychometrics is critical—is the measure *valid* (does it actually measure trust) and *reliable* (is it likely to measure the same thing twice), and also important to understand both its *sensitivity* (how well can it distinguish between trust and non-trustworthy states) and *specificity* (is it measuring global performance or other related constructs and not trust specifically). Scales that measure trust have often established these psychometric properties of their measures, and we will cover many of the existing scales in the next sections.

6.5 Questionnaires and Scales Used to Measure Trust in Automation

The methods above treat trust as something that can be inferred from interviews, behavior, or physiology. In practice, many empirical studies operationalize trust with a questionnaire: participants rate agreement with statements about a system, a class of systems, or people in general. Chapter 5 used

one such item—"I trust the system," item 9 of Körber (2019a)'s Trust in Automation (TiA) questionnaire—to illustrate how one perspective on trust is that it IS the outcome of a questionnaire. It also shows a common problem faced by trust scale developers—to answer the question, we must assume a user already knows what trust is. Many questionnaires include items such as this, but also other items with related constructs that are either related to trust definitions (e.g., faith, reliability, accuracy), or predictors or consequences of trust (reliance, understanding). But these related concepts are not identical to trust—a user can rely on a system they do not trust, or they can have faith in a system for unjustified reasons. Consequently, it may be important to be careful when interpreting a global trust score, as it may be composed of a user's own interpretation, and related but not identical concepts.

6.5.1 Scales that Measure General Trust and Trust in Automation or Technology

A number of trust scales and questionnaires appear in Table 6.3, and we will discuss some of the background research for several of the scales next.

Table 6.2: Trust assessment and measures. Items marked * appear in Chapter 10 of Adams et al. (2003); items marked † are described by Kohn et al. (2021). Interviews/CTA and think-aloud protocols are included here even though neither review treats them as a developed measurement method.

Assessment / measure	Description	Strength / weakness
Interviews (and CTA)*	Semi-structured interviews, often early in a program (e.g., naval data-fusion work reviewed in Aldern, 1995). A CTA of the work itself can expose trust issues indirectly (Matthews et al., 1999).	Explains <i>why</i> and <i>what</i> is trusted; non-numeric.
Think-aloud / verbal protocol	Concurrent or retrospective spoken thoughts while the person works with the system (Crandall et al., 2006).	Although a useful direct indicator of user thought process while in the situation, aspects related to trust may be rare.
Questionnaires and scales*†	Ranges from informal ratings to validated scales with subscales: (see Table 6.3).	Repeatable and easy to field. Risk that an easy measure replaces in-depth inquiry into system.
Neuropsychological measures†	Secondary indices: electrodermal activity, heart rate / HRV, eye gaze and pupil, and neural measures (EEG, fNIRS, fMRI) (Kohn et al., 2021).	Objective and precisely timed. Does not measure trust directly, and may incorporate load, arousal, risk, and other unrelated constructs.
Defensive monitoring / verification*†	Examining user monitoring frequency while in supervisory control.	Continuous and behavioral, but coding audio/video logs is labor-intensive.
Use of / reliance on automation*†	Frequency, duration, and automatic vs. manual control (Adams et al., 2003). Also discussed by Kohn et al. (2021) in terms of reliance, compliance, intervention, delegation, and decision time.	Objective, but is at best a proxy for trust, because many users rely on tools they do not trust.
Economic or strategic games / markets†	Money, tokens, points, or contract value placed at risk by cooperating or investing. Examples include the Give-Some game after a conversation with a robot (DeSteno et al., 2012; Lee et al., 2013), investment and social-dilemma payoffs in computer-mediated teams (Zheng et al., 2002), loans to a robot partner (Jessup et al., 2019), and repeated Prisoner's Dilemma, Chicken, and multi-arm trust games (Juvina et al., 2015, 2019; Collins and Juvina, 2021).	Incentivized behavior adds realism; often limited to understanding interpersonal trust with a machine as the partner.

Table 6.3: Selected questionnaire measures of trust and related constructs, excluding scales and questionnaires used for measuring *trust in AI* which are covered in Table 6.5.

Instrument	Source	Focus
General interpersonal trust		
Mayer and Davis trust and trustworthiness	Mayer and Davis (1999)	Interpersonal items for ability, benevolence, integrity, and trust. Sometimes adapted to automation.
Merritt trust scale	Merritt (2011)	Six trust items plus ability and benevolence; often adapted for automation.
ESS / Breyer Social Trust Scale	Breyer (2015)	Three items on <i>people in general</i> : whether most people can be trusted.
ESS Institutional Trust Scale (ESS-ITS)	European Social Survey (2023)	Seven 0–10 ratings of trust in legal and political systems—not trust in machines.
Trust in automation: component and general scales		
Trust vs. self-confidence items	Lee and Moray (1992, 1994)	Parallel ratings of trust in automation and confidence in manual control.
Complacency-Potential Rating Scale (CPRS)	Singh et al. (1993)	20 items on attitudes toward automation that includes a trust.
Component ratings (competence, predictability, etc.)	Muir and Moray (1996)	Subjective ratings of trust's predictors / components.
Trust in Automated Systems Survey (TIAS)	Jian et al. (2000)	12 empirically clustered items; A later 3-item short form (S-TIAS) was validated for AI (McGrath et al., 2025).
Human–Computer Trust (HCT)	Madsen and Gregor (2000)	25 items; reliability, competence, understandability, faith, personal attachment.
Dynamic reporting of trust	Desai (2012)	Repeated increase / same / decrease reports relative to the last rating; low interruption, for tracking change over a trial.
Propensity to Trust Automation (PTA)	Jessup et al. (2019)	Six pre-task items with “automated agent” as referent. Used for initial trust propensity, not in-task trust.
Trust in Automation (TiA)	Körber (2019a)	19 items, six subscales; CC BY-SA 4.0 (Table 6.4).
TOAST	Wojton et al. (2020)	9 items; Understanding and Performance scored separately.
Propensity to Trust in Automated Technology (PTT-A)	Scholz et al. (2025)	Validation of scale for propensity to trust automated technology, compared with propensity to trust humans (PTT-H).
Special-purpose measures (robots, HRI, XAI)		
SHAPE Automation Trust Index (SATI)	Goillau et al. (2001)	General questions (trust, usefulness, etc.) tailored to air-traffic management simulator
Trust Perception Scale–HRI	Schaefer (2016)	Robot-directed trustworthiness ratings.
MDMT	Malle and Ullman (2021)	Multidimensional human–robot trust (performance and moral facets).

These instruments do not all measure the same construct. The first section of the table covers commonly used interpersonal trust measures (e.g., Mayer and Davis, 1999) or social trust measures (Breyer, 2015). These generally are not suitable as a drop-in measure for AI or automation, but they are sometimes adapted for this purpose. Furthermore, two scales that appear in the European Social Survey (ESS) characterize general trust toward others and institutions, which may be useful in understanding how an individual views AI within society or their organization. The next section includes ten scales intended to be used for measuring trust in automation, computers, or technology. Instruments such as Human–Computer Trust (HCT) sit alongside system-state and propensity measures. Some provide ratings of the trust *state* in a specific system during or after use. Some rate a *propensity* or disposition to trust automation (much like the ESS survey items). Many of these have short version (i.e., the 12-item TIAS vs the 3-item S-TIAS) or subscales (TiA includes six subscales) that can allow for more efficient measurement or narrowing in on specific aspects of trust. The last section includes trust scales developed for specific kinds of automation domains or applications (air traffic control, robotics).

There are several caveats to be considered when selecting a scale. Adams et al. (2003) warned that many early measures were theoretically rather than empirically derived, which suggests they have unknown or may have poor psychometric properties. Furthermore, using a single numerical rating derived from one or more items falls prey to all of the mistaken assumptions about trust discussed in Chapter 5 (a numeric, uni-dimensional, holistic, and universal concept). This is partly borne out empirically—Jessup et al. (2019) found poor reliability and no prediction of the Jian ratings (a state-based measure) for predicting robot-investment behavior, while propensity to trust was predictive. Also, asking trust items during a session can influence how people rely on the recommendations of a system (Schrills et al., 2026), as it presumably encourages them to focus on the trustworthiness of the system in ways they wouldn't otherwise consider. Finally, Perrig et al. (2023) showed that trust in automation scales can fail to transfer to AI contexts, and so new measures and new validations may be needed. We will examine a number of trust-in-AI scales in a later section of this chapter.

6.5.2 Körber's Trust in Automation (TiA) Questionnaire

The dozen or so scales in Table 6.3 overlap, but are often targeting different things—disposition versus current state, or targeting automation versus robotics versus humans. It is useful to examine one scale to understand what it specifically measures. Körber (2019b) released his TiA scale as a CC-BY-SA, which allows us to reproduce it here in Table 6.4. Körber (2019a) built TiA from Mayer and colleagues' organizational-trust model and Lee and See (2004)'s account of trust in automation. An initial pool of 57 items was reduced to 19

Table 6.4: Trust in Automation (TiA) questionnaire (Körber, 2019a). Items marked * are reverse-coded. Response scale: 1 = strongly disagree to 5 = strongly agree. Reproduced under CC BY-SA 4.0.

#	Item	Subscale
1	The system is capable of interpreting situations correctly.	Reliability/Competence
2	The system state was always clear to me.	Understanding/Predictability
3	I already know similar systems.	Familiarity
4	The developers are trustworthy.	Intention of Developers
5*	One should be careful with unfamiliar automated systems.	Propensity to Trust
6	The system works reliably.	Reliability/Competence
7*	The system reacts unpredictably.	Understanding/Predictability
8	The developers take my well-being seriously.	Intention of Developers
9	I trust the system.	Trust in Automation
10*	A system malfunction is likely.	Reliability/Competence
11	I was able to understand why things happened.	Understanding/Predictability
12	I rather trust a system than I mistrust it.	Propensity to Trust
13	The system is capable of taking over complicated tasks.	Reliability/Competence
14	I can rely on the system.	Trust in Automation
15*	The system might make sporadic errors.	Reliability/Competence
16*	It is difficult to identify what the system will do next.	Understanding/Predictability
17	I have already used similar systems.	Familiarity
18	Automated systems generally work well.	Propensity to Trust
19	I am confident about the system's capabilities.	Reliability/Competence

items on six subscales: Reliability/Competence, Understanding/Predictability, Familiarity, Intention of Developers, Propensity to Trust, and Trust in Automation. Responses use a five-point agreement scale (plus “no response”), and several items (5, 7, 10, 15, and 16) are reverse-framed so that higher values indicate more distrust.

Like many of the other trust scales, this includes questions about reliability, predictability, clarity, and user knowledge. This illustrates one of the assertions in Chapter 5 that trust is *multidimensional*. Like most scales, it treats trust as a value that can be quantified, and refers to ‘the system’ as a whole, ignoring the componential nature of trust. Finally, the questions tend to ignore context—they neither ask nor qualify ratings to understand *when* trust is appropriate. Nevertheless, this can easily provide a user assessment of trust that is comprehensive and has strong psychometric properties.

Körber suggests individual subscales should be examined separately and not as a full scale. It is feasible to collect individual subscales as well, such

as the Trust in Automation items (9 and 14) which include “I trust the system”. Körber (2019a) also discusses when a single trust item is enough. Note that the 3-item S-TIAS (McGrath et al., 2025) is a short form of *Jian’s* scale, not of TiA.

6.5.3 Measures of Trust in Artificial Intelligence

Most scales of trust in automation were adapted from trust theories rooted in interpersonal trust, in the context of automation. These often found that the principles used to understand interpersonal trust do not translate directly to automation. So it is also important to know whether constructs related to trust in automation translate to trust-in-AI. We might suspect that the additional autonomy, risks, increased cognitive domain, and anthropomorphic properties will make measuring trust in AI more like measuring interpersonal trust—but it may also involve other new factors. Perrig et al. (2023) showed that several popular trust questionnaires do not easily translate to AI— their factor structures change when used to evaluate AI. Scharowski et al. (2025) ran a pre-registered test of Jian et al.’s Trust between People and Automation (TPA) scale and of the Hoffman Trust Scale for the XAI Context (TXAI). TXAI proved useful as a short state measure, but TPA did not support Jian’s original single factor: items 1–5 behaved as *distrust* and items 6–12 as *trust*. They advised scoring those composites separately without reverse-coding the negative items into one total. McGrath et al. (2025) similarly revalidated TPA for AI and published a 3-item short form (S-TIAS).

Table 6.5: Questionnaire measures developed or validated specifically for trust in AI.

Instrument	Source	Nature and lineage
Trust in XAI Context (TXAI)	Hoffman et al. (2021, 2023b)	8 Likert items plus an open reason for a <i>named</i> tool after use (Table ??). Independent validation: Scharowski et al. (2025).
Jian TPA used for AI; S-TIAS	Jian et al. (2000); McGrath et al. (2025); Scharowski et al. (2025)	Same 12 items as the automation scale, now with an AI referent. Scharowski recommends using two scores (distrust 1–5, trust 6–12). Adaptation of Jian.
TAIS (Trust in AI Scale)	Wischnewski et al. (2026)	30 items; bifactor: global trust plus ability, integrity, transparency, unbiasedness, and vigilance. New scale adapted from Muir, Schaefer, MDMT, TOAST, and others.
Semantic-differential AI trust	Shang et al. (2024)	New adjective-pair scale for an AI <i>agent</i> , scoring emotional and cognitive trust separately.
Human-like vs. functionality trust	Choung et al. (2023)	11 items (6 human-like benevolence/integrity; 5 functionality/competence). Built for a TAM path model of AI acceptance, starting with Mayer-style interpersonal trust constructs.
TrAAIT	Stevens and Stetson (2023)	Clinician trust <i>and</i> acceptance of AI in medical/health-care contexts.
Big Six AI trust	Dickens et al. (2026)	20-item scale with 6 factors for knowledge-work settings.
Gulati HCT (HCI)	Gulati et al. (2019); Pinto et al. (2022)	Human–computer trust for interactive systems; adapted by Pinto et al. (2022) for robotics.
GAAIS	Schepman and Rodway (2023)	General <i>attitudes</i> toward AI (positive/negative), extends beyond trust.
TAIGHA / TAIGHA-s	Kopka et al. (2026)	New scale (and short form) for trust in AI-generated <i>health advice</i> . Domain-specific.
Human–generative AI trust	Wang et al. (2026); Alon and Levkovich (2026)	Validation and adaptation of multidimensional trust items for generative AI / “black box” tools.
Feeling trusted by AI	Xie et al. (2025b)	New construct measuring perception that an agent trusts respondent. Not a measure of trust <i>in</i> the AI.

The measures in Table 6.5 focus on trust in AI in general or in broad contexts (XAI, health advice, generative AI). There are a number of other surveys and questionnaires that have an even narrower focus, including: distance-learning trust in AI (Üstün et al., 2026), hospital follow-up systems (Xie et al., 2025a), social-service robots (Cai et al., 2024), a child-centered scale (Ragone et al., 2026), and a human-and-AI trust-attitude scale (Larasati et al., 2023; Larasati, 2025).

There are also a number of attitude measures (single-item AI attitude; AI-interaction positivity scales) that have been developed, and these have been shown to correlate with ChatGPT trust, but are not measures of trust alone (Montag and Ali, 2025; Montag and Elhai, 2025). Finally, dispositional propensity to trust technology (akin to the PTA and PTT-A discussed earlier) was found to mediate AI fear and acceptance (Montag et al., 2023b).

For the Human Factors Practitioner

This chapter demonstrates that the notion of trust seems like a simple intuitive concept, but in reality it is nuanced and difficult to pin down. Some common practical lessons from this include:

- Measuring trust should recognize the nature of trust identified in Chapter 5: it is a process, it is multidimensional, contextual, componential, and can be categorical.
- Before using a questionnaire, consider some qualitative approaches such as interviews or think-aloud protocols.
- Most questionnaires measuring trust involve a broad set of associated concepts. These will often be highly correlated because of careless responding.
- Trust-in-AI, trust-in-automation, and trust-in-people are distinct constructs.
- Complementary methods for measuring trust can be deployed—combining physiological, behavioral, and subjective measures to provide stronger converging evidence.

Conclusions

Trust has been measured in many ways. The first concern is not the metric, but the context and paradigm trust is measured under. These vary from real-world settings to simulations, wizard-of-oz studies, and lightweight scenarios.

Once a setting is determined, measuring trust frequently starts and ends with short questionnaires or surveys. However, there are many other measures, including interview and linguistic approaches, behavioral measures and

neurophysiological systems. Many of these are useful in ways and contexts that questionnaires are not.

Within scale-based trust measures, there are also many options, measuring different aspects and senses of trust. Research has shown that automation-based trust measures do not necessarily translate to AI systems, so care should be used when selecting which instruments, and how they are used.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction, Michigan Technological University, Spring 2024.

Author Contributions

YW, BF, JW, LS: Background research and writing for specific subtopics; STM: Conceptualization; background research, additional writing in all sections; general editing.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How We Understand AI and Automation: Mental Models and Transparency

Blade Frisch, Katie Ulinski, Erin Matas, Christopher Cischke, and Shane T. Mueller

Michigan Technological University

Abstract

The notion of a "Mental model" has had incredible influence in cognitive science and human factors in understanding how we represent the world, interact with complex systems and devices, and forecast expected behavior. The idea, originating in the 1940s by Craik (1943), grew in interest during the 1960s and 1970s in the study of cybernetics and ergonomics to describe how users understand complex mechanical and control systems. Its importance for human-automation and AI interaction is illustrated by *good regulator* theorem, which suggests that mental models are necessary (or at least a satisfactory mental representation) for controlling any complex system. The notion of a mental model is central to many applied human-centered research domains (human factors, human-AI interaction, Human-computer interaction), and has been shown to be useful for improved system design, performance, and user training. We discuss its use in AI research as well, where the notion of an AI "Theory of mind" corresponds to an AI system's model of humans. We also cover methods used to elicit and characterize mental models in several research areas, the related issues of transparency and observability, and conclude with a discussion of some of the practical uses for mental models in research and design.

7.1 Introduction: Are internal models *necessary*? The Good Regulator Theorem

A famous paper in cybernetics and systems science by Conant and Ashby (1970) argued that “every good regulator of a system is a model of that system.” This “good regulator” principle is sometimes explained with an analogy like “a good key must be a model of the lock it opens.” They offered a mathematical proof grounded in information theory for this. Whether the proof actually maps onto the claim has been disputed, as it depends on what “good” and what “model” is interpreted to mean. For example, if one argues that a thermostat is a good controller of room temperature, the “model” in a thermostat can be two thin pieces of metal that bend differentially in response to the temperature, causing a circuit to close based on the position of a control dial. We either need to accept that this is not a good regulator (after all, as shown in Chapter 1, it can oscillate), or concede that the physical arrangement of metal is a model. Because these kinds of thermostats are among the most common and useful control devices in history, it is almost silly to insist it is NOT a good regulator, so we are left with the conclusion that the thermostat is a model. This is not too different from certain machine learning schemes: Q-learning (see also Chapter 1) has been shown to learn optimal control of a system through updating weights in a policy, and if it achieves optimal control, the weights in the policy must be the model identified by Conant and Ashby—even if they do not easily represent the system being controlled. Although this control-theory perspective on human-machine interaction was foundational in the field of cybernetics and has been developed more systematically (Francis and Wonham, 1976), perspectives on human control have gone far beyond this abstract approach to understand many of the specific ways humans control systems (e.g., Jagacinski and Flach, 2003), to include internal models, but also to focus on other important aspects of control.

The implication of the good regulator theorem is that for a human to control or regulate any system (including an AI system), they must have an internal model of that system. However, the *mental model* required to satisfy the theorem may not be the kind that humans typically use. Consequently, the good regulator theorem might be difficult to justify for mental models. However, maybe a related version of the regulator theorem does hold: *an adequate mental model permits good regulation of a system*. Furthermore, because a mental model of the world is a natural way for humans to build understanding of a complex system, a mental model may be the best regulator for human control of intelligent systems. It lets us predict outcomes, understand limitations, debug problems, and operate the system effectively—even when the system involves other people or intelligent agents.

7.2 Mental Models: Definitions and Uses

The term “mental model” was popularized by Johnson-Laird (Johnson-Laird, 1980) as a mental structure useful for understanding human reasoning, although he attributes the concept to Craik (1943), who discussed how humans form small-scale models of the external reality. Rouse and Morris (1986) cited a number of prior uses dating to the 1960s (e.g., Rasmussen, 1979) in cybernetics, ergonomics, and systems control research, and noted that its use in psychology was relatively recent. Its use in the study of reasoning was one of the early cross-disciplinary constructs in the field of fundamental cognitive science, but it was actually imported from the applied field of cybernetics, rather than being a psychological construct applied to the study of human-machine interaction, as many may assume.

A “mental model” is often associated minimally with knowledge and understanding of a system. However, it typically goes beyond this, and implies a coherent mental structure that is a representation of a system—one that can be introspected, permits mental simulation, and allows forecasting the results of changes to inputs. A mental model is considered more than just knowledge: Carroll and Olson (1988) proposed a hierarchy of knowledge structures with simple rules at the lowest level, general methods at a higher level, and mental models higher still. However, researchers often use the concept of a mental model when simpler knowledge structures may really be at play. In the extreme, a mental model might be contrasted with a Gibsonian approach that suggests the complexity of the system lives largely in the environment (Flach and Holden, 1998), with tight coordination of perception and action, and complex inner representations are not necessary. However, even the Ecological Interface Design paradigm rooted in this tradition considers the mental models of operators (Borst et al., 2014), which illustrates the power of the concept.

Rouse and Morris (1986) examined the similarities and differences among different usages of mental models, and (based on Rasmussen (1979)’s taxonomy) defined mental models as “mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states” (p. 7 of DTIC report). This notion, redrawn from the original in Figure 7.1, illustrates that mental models allow operators to *describe*, *explain*, and *predict* the purpose, form, function, and state of a system.

7.3 Kinds of Mental Models

As shown in Table 7.1, many different researchers have proposed distinct definitions or taxonomies of mental models. They tend to share more in

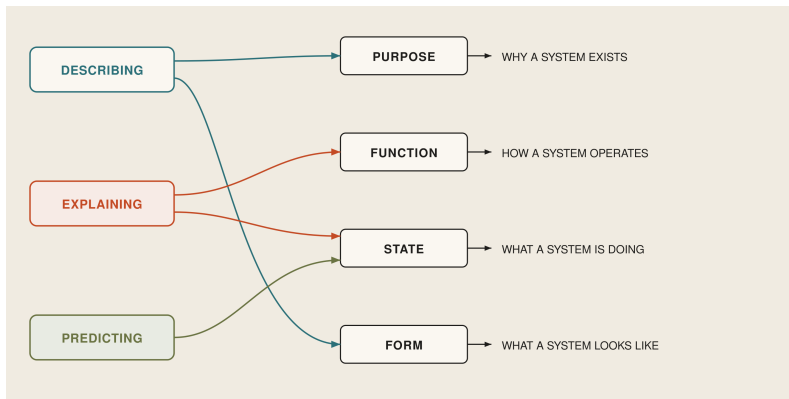


Figure 7.1: Purposes of mental models, redrawn after Figure 1 of Rouse and Morris (1986). Mental models allow one to describe, explain, and predict a system's purpose, form, function, and state.

common than they disagree on, but many are essentially views on how mental models are used within specific domains of research.

For example, Schumacher and Czerwinski (1992) identified three main categories of mental models described in the literature: (1) collections of knowledge structures; (2) descriptions of mental models as metaphors and analogies; and (3) process descriptions of how users interact with complex systems. The first group is commonly used for describing physical systems via mechanistic mental models (Carroll and Olson, 1988; de Kleer and Brown, 1983). The second group is common both in the basic cognitive science literature (e.g., Gentner and Stevens, 1983), but also in human-computer interaction research, where mental-model-metaphors include the “desktop”, the “file tree” and “menu”. The third is a more functional description, common in cognitive systems engineering (Borders et al., 2024), and other applications where the inner workings of a system are less important than a functional understanding of how a system is used. Next, we will discuss each of these, in addition to other kinds of mental models that have been described in the literature.

7.3.1 Mechanistic and Structural Mental Models

Williams et al. (1983) thought of mental models as a sort of runnable mental device: a collection of “autonomous objects” and their connections, where an autonomous object is a mental object with a state, a set of parameters, and connections to other objects. They illustrated this definition using a heat exchange system whose function is to cool down a hot liquid. The hot fluid could be an autonomous object with its temperature as a parameter and a

Table 7.1: Description of major “mental model” conceptions.

Reference	Description	Primary functions
Craik (1943)	Internal small-scale model of external reality	Anticipate; try actions “in the head”
Johnson-Laird (1980, 2004)	Possibilities for a reasoning / discourse problem	Infer; deduce; evaluate consistency
Gentner and Stevens (1983); Collins and Gentner (1987)	Physical / conceptual systems via analogy & structure-mapping	Explain; transfer; learn new domains
Williams et al. (1983)	Device as autonomous objects, parameters, links	Predict; explain by “running” connections
de Kleer and Brown (1983)	Mechanistic device models (assumptions & ambiguities)	Causal simulation; diagnose
Norman (1983)	Limited, simplified, and useful notions of how a system works	Incomplete and unstable, but useful
Rouse and Morris (1986)	Operator’s model of system	Describe/Explain/Predict
Carroll and Olson (1988)	Model of interactive software (HCI)	Learn; operate; troubleshoot
Wilson and Rutherford (1989)	HF applications of construct	Theory ↔ performance applications
Schumacher and Czerwinski (1992)	Knowledge structures, metaphors, and interaction	Supports expertise development
Staggers and Norcio (1993)	HCI concept review of competing MM definitions	Maps, causal reasoning, user-system models
Cannon-Bowers et al. (1990)	Shared Mental Model	Team’s understanding of roles and functions
Crandall et al. (2006)	Expert knowledge in natural settings; Richer with expertise	Decide; adapt; enrich with experience

connection to the heat exchange medium. One can then mentally imagine this object and connect it to the other components of the heat exchange system (e.g., the cooling medium, the heat exchange mechanism, etc.) to create a mental model of the system. The user can then “run” the mental model by following the connections between the autonomous objects as the state of an object changes. Williams et al. suggested that humans make use of mental models to perform such tasks as predicting the behavior of physical systems or providing explanations for the phenomena encountered. The mechanics of this definition are defined solely by the autonomous objects, their internal properties, and their connections to other autonomous objects. (de Kleer and Brown, 1983) used the term “mechanistic” mental models to specifically describe physical devices,

These might be considered a subset of *structural mental models* (Carroll and Olson, 1988; Schumacher and Czerwinski, 1992; Preece et al., 1994),

de Kleer and Brown (1983) specifically considered how such mental models allow quantitative simulation akin to running the system virtually inside the “mind’s eye.” This usage maps onto how many researchers in cybernetics and cognitive systems engineering have used the term, including Williams et al. (1983) (who studied a heat exchanger), Bainbridge (1992) (who examined a nuclear power plant) and Moray and Inagaki (1999) (who used a pasteurization plant simulator). However, whether they are considered mechanistic or structural, these same sorts of mental models of physical systems or objects were commonly used within basic cognitive science to study reasoning and analogy. For example, Gentner and Stevens (1983) used mental models as a means of representing processes and mechanisms in order to understand how analogical reasoning could be used for problem solving. (Collins and Gentner, 1987) extensively studied the mental models of water molecules and evaporation, which is not exactly a mechanical system, but involves mental simulations of colliding particles that go beyond simple structural understanding.

In contrast, the *probabilistic mental model* proposed by Gigerenzer et al. (1991) to understand confidence in judgments based on faulty memory retrieval is more properly considered a structural mental model. He used the term mental model to describe ad hoc models used to support judgment and reasoning for tasks such as estimating the number of inhabitants of cities in Germany. In this case, either a local mental model is retrieved from memory (that gives a high-confidence estimate), or one is constructed via analogy and probabilistic information to support the judgment.

Both mechanistic and structural mental models can describe abstract but mechanistic or structural accounts of intelligent systems. For example, many accounts of how a deep neural network work start by describing the architecture, the operations of layers, the training mechanisms, and tracing information through the network, which Hoffman et al. (2018a) discussed as a “global” explanation. Although the system is completely virtual, a mental model of that system still requires understanding both the structure and the mechanisms of that system.

7.3.2 Functional Mental Models

A number of researchers have distinguished between structural mental models and more function-oriented mental models (Sanchez et al., 2026; Kieras and Bovair, 1984; Young, 1981). Sanchez et al. (2026) defined functional models as models that capture a system’s behavior at a surface level and how to operate a system. Functional models also describe the competence envelope, the relationship between inputs and outputs, and when/where to trust or rely on the system. For example, most people have no idea how the mechanisms in their anti-lock brakes work (i.e., they have no accurate mechanistic mental model). The actual physical system can be relatively simple, and includes

sensors, actuators, and other logic, and there are several different mechanisms that have been used. Nevertheless, they may have a good idea of when the system will turn on, when it will be helpful, when it might fail, and how to deal with it if it does fail. Importantly, a mechanistic mental model is likely not very helpful to a driver when they hit an icy spot—even though it may be critical for a mechanic who needs to fix the system.

This notion dates back at least to Young (1981, 1983), who contrasted surrogate models (a runnable stand-in for the device) with a task/action mapping model that included goals/tasks mapped onto actions and procedures. Furthermore, Rouse and Morris (1986) identified that mental models cover describing and explaining the *function* and purpose of a system. One of the three conceptualizations of mental models described by Schumacher and Czerwinski (1992) was “process descriptions of how users interact with complex systems”, and Preece et al. (1994) distinguished between structural and functional mental models in interaction design. Seong and Bisanz (2000) identified input-output mappings, and Bansal et al. (2019) used the notion to describe how a human understands the *error boundary* of a classifier system—akin to the competence envelope we have discussed in Chapters 5–6. Also, these “functional” mental models are distinct from how-to-operate procedures that were frequently contrasted with mental-model based instruction (Carroll and Carrithers, 1984; Halasz and Moran, 1983; Kieras and Bovair, 1984) in early tests of the notion of mental models. This research which often studying pocket calculator users compared learned routines to deeper understanding via mental models, but the learned routines were usually rote step-by-step procedures.

The distinction between functional and structural mental models is critical, especially for applications of design and training. It is probably true that users of a system require a functional model for normal operation, and may benefit from a structural model in some cases, but in many situations, a structural model is not needed and might be a distraction. On the other hand, a developer or a mechanic might require a structural mental model in order to successfully diagnose and fix problems. For example, users of LLMs normally understand the hallucination problem—it makes up sources and ideas and fabricates pointers to non-existent provenance. If they understand and can predict when this is likely to happen, then they can successfully implement workarounds and checks. They do not need to understand the transformer architecture of the large language model, the back-propagation process, or other learning schemes such as parameter tuning, retrieval-augmented generation (RAG), and HITL-ML. However, an engineer who is attempting to improve their system will need to know this. This issue impacts the field of Explainable AI, which we cover in Chapter ??.

7.3.3 Shared mental models

Cannon-Bowers et al. (1990); Cannon-Bowers and Salas (2001) popularized the notion of mental models as it applied to team cognition, in terms of “shared mental models”. They used it to describe how, in high-performing teams, common understanding of goals meant a team could coordinate behavior with minimal communication. This notion is typically used to describe how teams understand shared goals, understanding of the task, and the roles of other team members. The term is a central component of a broader set of constructs called shared or team cognition (DeChurch and Mesmer-Magnus, 2010a), and although the term appears to describe many different senses of team cognition, a meta-analysis showed that shared mental models were positively related to team performance, and only research that used mental models to represent the structure of knowledge was predictive of the team process (DeChurch and Mesmer-Magnus, 2010b). These demonstrate the utility of specific structural kinds of mental models in mediating team performance.

As automation and AI become “members” of the team (see Chapter 12), this same shared understanding can be critical for team performance in human-AI teams (Andrews et al., 2023). Some research has shown how it is important for the human team member to have a good mental model of the AI (Bansal et al., 2019). However, the other direction is also possible: user models and student models have been successful in recommender systems and intelligent tutoring systems for many years, and these represent the equivalent of a “mental model” representation within an AI system about humans. As we discuss in Section 7.4.6, these notions have been developed to help AI systems reason about people who are using it. For example, Schelble et al. (2022) and Andrews et al. (2023) used shared mental models in the context of human-AI teaming.

7.4 Mental Models across Disciplines

7.4.1 Mental Model as a Central Construct in Human-Automation Research

Mental models of automation and AI serve as the “glue” that binds together many of the topics covered in this handbook. As illustrated in Figure 7.2, almost all of human-automation and human-AI interaction involves factors that either lead to better mental models, or are improved by having better mental models. First, to form an accurate mental model, you may need to understand the system architecture, algorithms, and purpose (Chapter 1). Next, mental models are supported by transparency (Section 7.6), explanation (Chapter ??) and instruction (Chapter 10). Work setting factors such as task demands and teaming with other people and machines (Chapters 3, 12, and 13) also shape

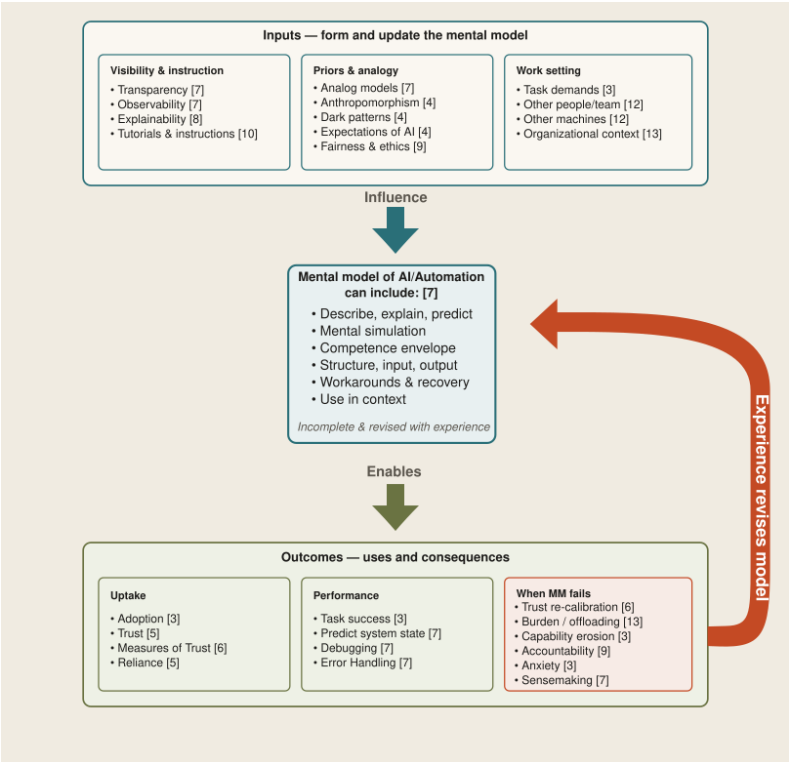


Figure 7.2: Mental models at the center of human–AI interaction. Topical inputs form and update the model; the hub lists what the model contains and affords; outcomes include uptake, performance uses, and failure modes when the model is inadequate. Bracketed numbers indicate the primary handbook chapter for each topic.

what must be modeled. An important aspect of the mental model is what the tool is good for and what it is not (the Mental Model Matrix discussed in Chapter 3 and in this chapter), which is useful for understanding justified and calibrated trust and mistrust (Chapters 5–6). Accurate and warranted trust further supports adoption and reliance (Chapters 3 and 5), which can help develop even more nuanced mental models. Anthropomorphic features or assumptions (Chapter 4) and fairness or ethics priors (Chapter 9) may shape initial expectations and mental models for how the system works, with those models revised through experience.

Mental models also form central components of a number of other conceptual and scientific models. For example, in her diagram illustrating situation awareness, Endsley (1995) suggested that mental models are the central information that mediates goals, plans, situation models, outcomes, and the

environment. Similarly, the trust model of Muir (1994) (see Figure 5.7) places the operator mental model as the central construct between perceptions of trustworthiness and behavior. Klein et al. (2006) uses the notion of a simplified mental model they call a “frame” as the core knowledge that is being used and deployed in sensemaking. Not only are mental models central to human-automation research, they are important components in many other models relevant to the field.

To summarize the different conceptualizations, mental models permit:

- Explaining and describing the system (its function, purpose, state, form) (Rouse and Morris, 1986)
- Mental simulation (Rouse and Morris, 1986; Klein, 1998)
- Understanding what the system is good at and what it is not good at (competence envelope) (Borders et al., 2024)
- Predicting direction and degree of association between input parameters and a system output (Seong and Bisantz, 2000)
- Workarounds, debugging (Kulesza et al., 2013), and error recovery.
- Use in context with other people and machines, including how work is allocated (Ahmed et al., 2026), interaction, and feedback.

7.4.2 Mental models in cognitive science

As discussed earlier, mental models have a prominent role in the basic study of cognitive processes within cognitive science (Johnson-Laird, 2004), especially with a focus on how they support logical reasoning (Johnson-Laird, 1980) and analogical reasoning (Gentner and Stevens, 1983). Collins and Gentner (1987) explored how such mental models are constructed, and Chi et al. (1994); Chi (2000) explored how self-explanation can generate and repair mental models. Although these typically describe basic research on human mental representation, reasoning, and learning, they form a core experimental and theoretical literature for establishing the concept. It should be recognized that these researchers were focused on basic problems, but with the emergence of cognitive science as a multi-disciplinary field involving psychology, computer science, AI, philosophy, and linguistics, there was an interest and capability to study more complex representations and processes than had previously been feasible. So these researchers appear to have been inspired by the earlier applied use of mental models in more practical settings, rather than the reverse.

7.4.3 Mental models in HCI research

The notion also found strong uptake in the 1980s with research focused on human-computer interaction—driven in part by the development of personal computers. This came from the recognition that devices and software that the

nascent computer industry often created were unusable by consumers without training, and hostile to users who eventually came to depend on the systems. This usability problems sometimes represented a mismatch between how the software or product was designed to work in the mind of the designer, and how the user thought it worked—a mismatch of mental models.

For example, Norman (1983) offered a much more limited view on the use of mental models for human-computer interaction. Through observations of users of a calculator, he found that their understanding was limited, parsimonious, and unstable, and evolving as they learned and interacted with a system. They did not need to be accurate but they must be functional for the user. He also found that their use for mental simulation was limited in his studies—users appeared to have little ability to “run” their model in their head, and would prefer to write down intermediate answers on paper.

The notion was used in this context extensively by Carroll and Olson (1988), who considered both design and training as possible applications of mental models in HCI. Much of this work on mental models was carried out on studying calculators (Young, 1981, 1983; Kieras and Bovair, 1984; Norman, 1983; Halasz and Moran, 1983), which required a specific understanding of its workings that did not always match users’ initial mental models. As the field expanded to design of desktop computer software, the software became more intuitive and the mental models that came to be used were often metaphorical (a window, desktop, and filing system) rather than the mechanistic models required for mentally simulating calculators.

7.4.4 Mental models in human factors

The field of human factors is broader than HCI, but has found use for mental models across a wide number of applications. A number of comprehensive reviews exist that cover mental models in human factors and ergonomics. Wilson and Rutherford (1989), who discussed the use of the term across cognitive science and linguistics, noted confusion about the term and its relationship to and distinction from other terms (such as schemata). They argued that in human factors, the term has no single agreed-upon definition, while largely ignoring the more basic research conducted in cognitive science research. Staggers and Norcio (1993) recognized that mental models play three roles, for design, for user performance, and for training. They also posed several critical questions regarding mental models—how can we tell they exist, why users can be relatively capable even while appearing to have minimal or poor mental models, and how mental models are developed.

Although these reviews were published during the heyday of mental models research, it is fair to say that in the intervening years, the notion has remained influential, while the variety of interpretations and limitations have remained. Mental models have found special resonance within research on

team cognition, with the highly influential notion of “shared” mental models (discussed earlier in this chapter). As much of this work was rooted in the study of automation, the notion has also gained new currency in the study of AI-as-automation and human-AI interaction, both from the perspective of how users form mental models of AI systems, and (maybe controversially) how AI forms mental models of humans.

7.4.5 Mental Models of Chatbots and other AI systems

Just as we form models of physical and software systems (Ehrlich, 1996), so do we form mental models of AI systems. Mental models of AI have been studied in many contexts, and many of these correspond to classic mental models of automation, including human-AI decision making (Bansal et al., 2019), home virtual assistants (Cho, 2018), image classifiers (Bos et al., 2019), and reinforcement learning (Anderson et al., 2020). These sometimes focus on underlying mechanisms, but often are more functional—they describe how a user understands the system’s strengths, limitations, and competencies, often with expectations set by human performance in similar tasks.

As one example, Cho studied the mental models people form of home virtual assistants (HVAs) like Google Home. They had participants in their study interact with the HVA to carry out tasks or answer questions. The interaction patterns revealed aspects of the users’ mental models: some would try to prime the HVA with the information they thought was necessary for answering their question, such as telling the HVA that the Detroit Pistons are a basketball team before asking a question about the team. Others would back off of specific questions to more general questions if the HVA failed to answer the specific questions satisfactorily, becoming unsure of how much they believed the HVA knew or how well it could understand their questions. This revealed users’ mental models of the assistant, with an expectation of it working like human language and memory.

This anthropomorphic expectation (Chapter 4) appears to strongly influence mental models of AI users—especially systems such as chatbots or “conversational agents” with anthropomorphic features. These will often import the expectations, interaction patterns, and perceived abilities and weaknesses from our mental models of humans doing similar tasks. For example, Gero et al. (2020) found that people tended to overestimate the abilities of the AI by assuming the AI would be acting in a way similar to humans (e.g., providing hints that can all make a sentence with the target word). However, Grimes et al. (2021) found that the initial expectations in a customer service agent could influence engagement. Participants’ expectations about capability (i.e., being told they were going to be chatting with a human versus chatbot) were manipulated; they found that a low-capability chatbot maintained the initial engagement score, higher than if they were initially led to believe they would

be talking with a human. Here, the mental model involved expectations about poor conversational ability from chatbots. Villareale and Zhu (2021); Villareale et al. (2022) used word games involving AI systems as a test domain to elicit mental models.

Brachman et al. (2025) elicited mental models (via semi-structured interviews) of users who interacted with a chatbot to understand a topic or get help with the health of a relative. They found that resulting mental models after interaction were limited and had gaps in understanding how the model worked, performed actions, used data, and so forth. This finding is similar to what Norman (1983) concluded when interviewing calculator users decades before. Furthermore, although the chatbot's responses were mostly inaccurate (ranging from 0% to 38% correct), confidence in the answers was uniformly high (around 6 on a 7-point scale). So although their knowledge was minimal, it was also not functional because it misled them into thinking the system was more accurate than it really was.

As chatbots gain capability and become more widespread, it is likely that the expectations and initial mental models will change. The impact of anthropomorphic features is likely to persist, but expectation that they work like humans may be replaced with expectations that they work like the last generation of chatbot, or how they have seen others claim to use it. It is also clear that many features in communication that we used to judge competence of other people can be easily replicated by chatbots, even when the actual competence is low. Correct spelling, grammar, detailed links to sources, and bullet-lists that look like complex reasoning chains can often signal that a human has carefully and attentively created a document, suggesting accuracy of the content—this is subverted in AI-generated chats and documents, because these surface features are almost always stellar, even if the accuracy or wisdom of the content is not.

7.4.6 Mental Models in and of AI: Theory of Mind and Game Theory

Mental models involving other people are studied in two main contexts: within psychology (especially developmental psychology) in terms of a “Theory of Mind” (ToM), and within economic decision making in terms of strategic and cooperative “game theory”. The focus of these disciplines differs: ToM research often attempts to understand how children gain the awareness that other people think and have their own goals. In strategic games research, we consider different actions one agent can take, and responses a partner can take, with different costs and benefits associated with each outcome. Optimal strategies must often force one to look at the problem from the partner or adversary's perspective, and sometimes look at how the adversary looks at the problem from one's own perspective. Although this work is often couched

SimpleToM example

Story.

“The can of Pringles has moldy chips in it. Mary picks up the can in the supermarket and walks to the cashier.”

Mental state. “Is Mary aware of the mold?”

Correct: No

Behavior. “Will Mary pay for the chips or report the mold?”

Correct: Pay

Judgment. “Mary paid for the chips. Was that reasonable?”

Correct: Yes

Figure 7.3: An illustrative SimpleToM item (Gu et al., 2024). Commercial LLMs often answer the mental-state question correctly but fail more often on behavior prediction and judgment.

in terms of theory of mind, it often uses notions from game theory (for example, second-order reasoning).

In both cases, simple or complex mental models are formed of other intelligent agents' information processing. Thus, these form natural baseline models for human mental models of intelligent systems. As we argued in the last section, our initial mental models of AI systems are often influenced by expectations regarding how humans perform tasks, which in turn can be shaped by anthropomorphic features of agents.

The terms have also been adopted by AI developers (Garcia-Lopez, 2024; Schossau and Hintze, 2023; Mao et al., 2024), advocating for various approaches to embedding theory of mind into agents and determining whether the AI has ToM capabilities. There are theory-of-mind benchmarks allowing for testing ToM against mostly text-based scenarios. Gu et al. (2024) developed the SimpleToM benchmark, which involves a number of short narratives that highlight awareness, behavior, and judgments about behavior that can indicate LLM awareness of the mental state of a character in the narrative. Figure 7.3 shows one such item: the model is asked about the character's mental state, predicted behavior, and a judgment of that behavior.

Gu et al. (2024) showed that commercial LLMs tend to do well at mental state, but tend to fail at accurately predicting the behavior or making judgments about it, highlighting a current weakness.

Other related research has also focused on frameworks for “mutual” theory of mind (MToM), supporting human and AI representations of one another (Chakraborti and Kambhampati, 2018; Ashktorab et al., 2025; Wang et al., 2024). Others have argued that AI created with a theory of mind will allow

for social intelligence, promote integration into society, allow for decisions aligned with human benefit, and overall support ethical decision making by AIs (Cuzzolin et al., 2020; Williams et al., 2022; Langley et al., 2022; Nebreda et al., 2024; Matta, 2026; Walsh et al., 2026).

Nevertheless, there are some limitations of the overall project to embed ToM into AI. Yin et al. (2025) argued that despite AI systems performing well on ToM laboratory tasks, their performance is really sophisticated pattern matching rather than true awareness. A sophisticated classifier can be trained to respond correctly to many such tests, even without a coherent theory of mind. Similarly, Marchetti et al. (2025) argued that although LLMs can often perform well on ToM tasks, this is an “illusion of understanding”, as these systems do not possess the embodiment or developmental and cognitive mechanisms necessary for true Theory of Mind, and that results should be doubted because of methodological limitations.

7.4.7 Summary of Mental Models use in Research

The notion of a mental model has found purchase in many areas. It provides a powerful metaphor for understanding how we work with and learn complex devices, technology, social and socio-technical systems. This provides justification for Conant & Ashby’s Good Regulator theorem: we have mental models of these systems so that we can learn to use, control, and regulate them better.

However, the term is often used more informally to simply represent knowledge or understanding. Researchers may use it but not be committed to a strong interpretation of mental models, which requires a coherent mental structure that models a system or process in the world, and supports the ability for mental simulation and interpretation of the system’s state. It is difficult to distinguish between the simple and inconsistent mental models many people have of systems they use, and the hypothesis that people do not use a mental model to reason about the system. But even without a deep commitment to the defining properties that were discussed at the start of this chapter, we can attempt to uncover mental models (or their fragmentary parts) with a number of methods, which we will discuss next.

7.5 Eliciting and Measuring Mental Models

In this section, we will cover two aspects of methods for eliciting and characterizing mental models. First, we will discuss approaches for how information about mental models is elicited from users and research participants. Then, we will discuss frameworks that either guide this elicitation, or can be used to represent the mental-model-based information obtained during elicitation.

7.5.1 Empirical methods for eliciting mental-model information

Approaches for eliciting mental models cover many of the methods for measuring and characterizing trust that we discussed in Chapter 6. Many early studies of mental models adopted think-aloud verbal protocols to understand mental models (e.g., Norman, 1983). While a completely open-ended interview can be used, it is usually helpful to have a structured or semi-structured interview to guide the elicitation process. Semi-structured interviews allow for some unplanned questions or for questions to be asked out-of-order from the script. One obvious risk in interview techniques is that the mental model is constructed by the interviewer and may reflect more of the interviewer than the interviewee. Borders et al. (2024) describe aspects of Cognitive Task Analysis and semi-structured interview techniques like the Critical Decision Method that are popular in the fields of human factors and human-computer interaction for extracting tacit knowledge, with a focus on reporting specific incidents and thought processes, that can help guard against either implicit influences of the interviewer or confabulation by the interviewee.

However, elicitation techniques go beyond asking people to describe their mental model, or inferring it from interviews or think-aloud protocols. The best approach depends on the goals of the study. Sometimes, we may assume that a mental model is affected by a manipulation and simply want to characterize whether a user has a more accurate understanding. In this case, it may be enough to simply measure knowledge with a test or quiz (Kulesza et al., 2013). But a number of more formal approaches have been used that give different views on the mental model itself.

Several researchers have provided comprehensive overviews or novel methods for eliciting mental models in different domains. For example, Sasse (1997) identified three categories (performance data, concurrent protocols, and verbal protocols); DeChurch and Mesmer-Magnus (2010b) examined 23 studies in team cognition; Grenier and Dudzinska-Przesmitzki (2015) proposed a multi-method approach in human resources including a survey, interview, and concept mapping exercise; Hoffman et al. (2023b) described a comprehensive set of methodologies dating back to the early 20th century, in the context of explanation; and Sanchez et al. (2026) conducted a review of methodologies eliciting mental models in human-AI interaction. Table 7.2 combines these taxonomies, and organizes it into two groups: one related to the actual elicitation method, and a second related more to how the mental model is represented either as a coding scheme or a outcome.

Table 7.2: Methods for eliciting mental models and structural frameworks used to organize elicited content.

Category	Techniques / focus
Elicitation methods	
Researcher-led verbalization ^{[a],[c],[k],[l]}	Open-ended survey questions; structured interviews; metaphor or analogy elicitation; scaffolded task retrospection / reflection (including replay); other verbal protocols (including retrospective).
Participant-led verbalization ^{[a],[c],[k],[l]}	Think-aloud / on-line (concurrent) protocols; think-aloud with concurrent question answering; retrospective commentary; self-explanation / teach-back; diary or ethnographic accounts.
Structured judgments and ratings ^{[a],[b],[k],[m]}	Knowledge or comprehension tests; prediction or decision tasks (with justification); nearest-neighbor selection of explanations or diagrams; glitch / accident–error detection; ShadowBox comparison to an expert; vignettes or scenarios; selection or ranking; attitude or trait scales; similarity ratings; questionnaires.
Artifact-based elicitation ^{[a],[b],[k],[n]}	Free drawing; card sorting / affinity diagrams; concept mapping / diagramming; cooperative design; mockups, collages, or moodboards.
Content analysis of cognitive maps ^[b]	Coding verbal protocols or diagrams into cognitive maps (shared-team literature).
Performance data ^[c]	Time; errors; task completion; task retention—useful, but not sufficient alone to characterize a mental model.
Multi-method Mental Model Elicitation (MMME) ^{[j],[k]}	Explicitly combines survey, interview, and concept mapping in one protocol.
Structural models to guide elicitation	
Concept maps ^[n]	Concepts linked by labeled relations; often co-constructed with the interviewee (see Figure 7.4).
Pathfinder networks ^[d]	Similarity ratings among concepts yield a numeric network in place of linking phrases.
Causal / graphical models ^[e]	Causal relations identified in interview, then rendered as a graph.
Intention–attention ^{[f],[g]}	Coding framework for initial formation of mental models.
Assimilation / accommodation ^[g]	Coding framework for how mental models change with experience.
Automated map extraction ^[h]	Algorithmic parsing of text corpora into concept diagrams.
Mental Model Matrix ^[i]	Structure focused on capabilities and limitations of the system and the user.

Notes: ^[a] Sanchez et al. (2026); ^[b] DeChurch and Mesmer-Magnus (2010b); ^[c] Sasse (1997); ^[d] Schvaneveldt et al. (1988); ^[e] Ford and Sterman (1998); ^[f] Waern (1987); ^[g] Villareale et al. (2022); ^[h] Carley and Palmquist (1992); ^[i] Borders et al. (2024); ^[j] Grenier and Dudzinska-Przesmitzki (2015); ^[k] Hoffman et al. (2023b); ^[l] Illustrative verbalization methods: Ericsson and Simon (1984); Rasmussen et al. (1994); Gentner and Stevens (1983); Dodge et al. (2021); Cañas et al. (2003); ^[m] Illustrative judgment / selection methods: Muramatsu and Pratt (2001); Hardiman et al. (1989); Klein and Wright (2016); St-Cyr and Burns (2002); ^[n] Illustrative artifact / diagramming methods: van der Veer and Melguizo (2003); Novak and Gowin (1984); Moon et al. (2011); Chi et al. (1981).

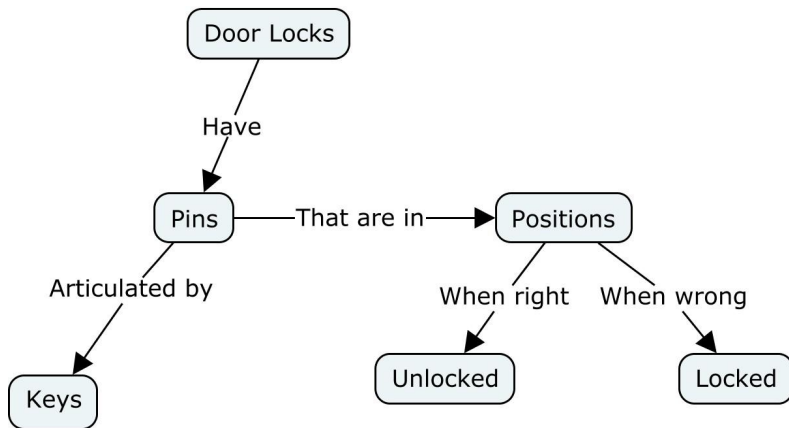


Figure 7.4: Example elicited concept map of a mechanistic mental model of a door lock: components and causal relations that can be “run” to explain locked versus unlocked states.

7.5.2 Structural approach to representing mental models

Concept mapping for mental model elicitation Many elicitation techniques rely on the creation of a diagram, as it reduces the amount of interpretation by the interviewer and immediately creates an artifact. A method frequently used here is concept maps (Novak and Gowin, 1984; Moon et al., 2011; Hoffman et al., 2023b). In these diagrams, ideas or “concepts” are connected by arrows and linking phrases, as shown in Figure 7.4.

This shows how a particular subject may understand the parts of a system (the pins, the key) and how they believe they interact with one another. Developing a concept map in conjunction with the user can be useful to ensure nothing is missing from the model and to perform several iterations on the map punctuated by semi-structured interview questions.

Other frameworks Regardless of the method by which information about a mental model is elicited, there have been a number of approaches to coding specific structures to provide a more comprehensive or consistent view of a mental model. For example, Villareale et al. (2022) focused on the *intention and attention* events proposed by Waern (1987) to describe initial formation of mental models, and also adopted the Piagetian distinction between *assimilation* and *accommodation* to characterize how mental models changed. Ford and Sterman (1998) described a process of identifying causal relationships first in an interview session and then representing them graphically. Pathfinder networks (Schvaneveldt et al., 1988) replace the linking phrases of concept maps with numeric similarity measures between the concepts to represent a

mental model using the “pathfinder” network approach. Finally, if you have a corpus of text, Carley and Palmquist (1992) describe algorithmic ways to parse the text and automatically generate diagrams relating concepts. The Mental Model Matrix (Borders et al., 2024) is a different structural approach, focusing on capabilities and limitations of the system and the user.

In summary, there are many well-established approaches to both eliciting mental models and describing them in structured frameworks. The choices depend on many factors (echoing the same issues we discussed for eliciting trust). In particular, the method might be influenced by:

- Distinct goals of elicitation (for design/redesign, for training, for improved performance, or for assessment).
- The kind of system (mechanical, virtual, socio-technical, etc.).
- The type of user (a developer, debugger, or an end-user).
- The level of expertise among the participants.
- Experience and practices among the research team.
- Time and resource availability (see minimum necessary rigor in Section 3.9).

The choice of method and approach depends on the context. This section reveals that there are often many alternatives that can be considered, and that choice is likely to make a difference in what is learned.

7.6 Transparency, Observability, and Explainability

Researchers studying mental models have frequently discussed how “transparency” can support better and more accurate mental models. This sense of the word is used because many systems were (and still remain) so-called “black box” architectures in which the users had no ability to observe or understand the inner workings. Researchers described hypothetical alternatives, such as the “glass box” system (Carroll and Carrithers, 1984), which is an analogy to educational models such as demonstration internal combustion engines with glass casings that allow non-experts to understand how the system works. This allows users to “see inside” the system, and terms such as “transparency”, “observability” and “explainability” have been used to describe this.

These terms appear to be used interchangeably at times. They each describe ways that a system can be designed to help users understand how the system works, and potentially develop better mental models of its workings. Each has taken on distinct connotations within different communities. From a cognitive and human-factors approach, these are typically used to describe design elements or interface features that permit a user to develop a better understanding or more accurate mental model. Transparency often indicates interface elements that show underlying data; observability reveals the state factors important for the user to know; explainable algorithms typically add

post-hoc data visualization to help understand how input features influenced a decision or classification.

These terms are often linked to mental-model development. For example, Ramaraj et al. (2019) showed how “transparent” interfaces that showed a robot’s visual processing of the world and the logical processing of its task led to better mental models of human observers. Similar informational uses appear in robot-based transparency (Matthews et al., 2019), recommender and search systems (Tsai and Brusilovsky, 2019; Ngo et al., 2020; Degachi et al., 2025), and human–autonomy teaming (Scali and Macredie, 2019; Parakh et al., 2025; Bhaskara et al., 2020). This research typically uses the term transparency in purely an informational sense, looking at its influence on mental models and understanding of a system. Observability is sometimes distinguished from transparency insofar as it focuses on surfacing important information rather than just removing ‘the lid’. One can visualize activation in a neural network (transparency), but a better approach might be to identify patterns that are used for specific detections, and make those observable (Zeiler and Fergus, 2014). Explainable AI (which we cover more extensively in Chapter ??) describes a series of approaches that aim to offer “explanations” for why an AI system behaves or makes a particular choice or decision. In general, explanations in this context are not usually text-based narratives describing causal reasoning, but often involve visualization, highlighting, providing examples, creating counterfactual situations, etc. In these cases, good explanations help build better mental models, leading to better understanding, trust, and performance (Hoffman et al., 2017, 2018b,c,a; Mueller et al., 2019).

However, each of these terms—especially transparency—is also used in describing the social and societal impacts of automation and AI. Transparency is often used in the sense of accountability and disclosure—revealing hidden motivations, obligations, risks, and consequences of using the system, which might include data privacy, the sources of training data, the potential environmental effects, and other risks. Andrada et al. (2022) described a taxonomy of transparency notions (which we will describe in Chapter 9), and most common uses address social and societal issues of the use of automation and AI. They include informational and reflexive transparency, which is the notion most similar to how this has been used traditionally to describe understanding a system’s mechanism and inner workings.

Similarly, *observability* is often used interchangeably with transparency, but Rieder and Hofmann (2020) distinguish between the transparency and observability for improving accountability of platforms. The rationale for this is because the algorithms themselves are often unable to be made intelligible—even to the developers—so that a better approach is a more pragmatic one where we make clear the important aspects of the decision. Walmsley (2021) suggested that in those cases in which transparency is not feasible because the inner workings of the system are either inscrutable or hard to obtain, an

alternative to adopt is *contestability*: the right to challenge the outcome of an algorithmic decision.

Finally, explainability is sometimes considered as a way of understanding fairness: regulations sometimes require that algorithmic decisions need to provide explanations to those impacted, so that if someone is denied a loan, they can know the reasons, and especially whether the reasons were justified. These uses are related, and better understanding of an algorithm is usually a consequence, but these are not intended to improve or develop a more accurate mental model, which is normally reserved for frequent users of a system who rely on it for daily work.

These are distinct senses of the terms used when discussing the societal impacts of AI. Although one may argue that a mental model *should* encompass the social impacts and risks of a technology, this dilutes the terms substantially so that it no longer has a clear meaning outside of “understanding the system and its context.” It is sometimes useful to qualify the use of these terms when discussing informational aspects that lead to more complete or more accurate mental models, so that others senses of transparency, observability, and explainability are not conflated.

Understanding and mental models also interact with other concepts in human-AI interaction, such as trust, usability, and acceptance. Transparent AI systems contribute to the development of trust by providing users with evidence of reliability, fairness, and accountability. Moreover, transparent interfaces and explanations enhance the usability of AI technologies, allowing users to effectively interact with and control AI systems (Borders et al., 2024). Ultimately, fostering transparency in human-AI interaction may promotes the acceptance and integration of AI technologies into various domains of society, improving adoption and use. It may also help potential users understand the drawbacks and limitations of the system, and so lead to justified distrust and disuse.

7.7 Practical Application of Mental Models

The notion of a mental model has been used extensively to explain how users understand systems, engage in reasoning, learn about the system, and test hypotheses. In this chapter alone, we have cited several comprehensive reviews and discussions of mental models in cognitive science and human factors—each of which go into much more detail than the present chapter. But some questions arise: what is our understanding of mental models good for? If we were to abandon the notion of a mental model and just consider “knowledge” or “understanding”, would it make a difference? There are some very practical ways we can use the notion of mental models, regardless of whether they truly “exist” or are just an associated set of knowledge about a system. Carroll and

Olson (1988) offered two main benefits: better design, and better user training, and Staggers and Norcio (1993) echoed these and offered a third: better performance. These are strategies for dealing with a mismatch between the user's mental model and the underlying system logic—you can either change the system, or change the user, or change the interaction (Mueller, 2020).

Using mental models for improved interface design Carroll and Olson (1987) described several examples of early software systems that were designed to support the mental models of users, and identified design approaches that were sensitive to user mental models. One is the “naive model”: build the system to match the naive user's mental model. A second approach is the “training wheels” approach (Carroll and Carrithers, 1984), using a simplified system for novice users that maps onto a sufficient but simplified mental model. There are advantages and challenges to any such approach, especially in scaling from novice users to expertise—they note that naive users do not necessarily make for good design, and the impact of a mental model on design can be obscured by other design aspects. Furthermore, Wilson and Rutherford (1989) argued that using mental models for design must increase the total effort, because the system must be developed and used before you can elicit mental models. Thus, it may be most useful for redesigning existing systems.

Mental models for better user training Carroll and Olson (1987) also reviewed existing research on attempting to directly train users with a mental model—either via a diagram or analogy (desktop metaphor). In comparison to more procedural training, results were equivocal—rarely did direct training on a mental model provide an advantage over other forms of learning. However, benefits were seen when an active learning approach was taken whereby the user developed a mental model through experience. In all of these examples, the mental model forms the basis for the learning goals, but is not provided directly (Mayer, 1980; Mack et al., 1983; Carroll and Mack, 1985). Consequently, for routine procedural tasks, learning a mental model may not provide substantial benefits over learning the procedure.

Other work has shown benefits for mental-model instruction versus rote learning. Kieras and Bovair (1984) tested device-model versus rote procedural learning in understanding an electronic device. Across three experiments, the mental model group learned procedures faster, retained them better, executed faster, and used shortcuts far more, and were able to infer untaught procedures. Fein et al. (1993) similarly showed an advantage for mental models in both recall and transfer learning, and Halasz and Moran (1983) showed that although there was no advantage of mental model training over procedure-based training on routine calculator problems, on novel “invention” problems there was a large advantage. These all suggest that there can be advantages to

mental model training, but it depends on what is being learned, and how the mental model is presented to the user.

There has also been evidence for the benefit of direct instruction about a system via diagrams. Butcher (2006) showed that a simplified diagram of the circulatory system produced the best mental-model development, in comparison to text-only training, or text accompanying a complex diagram. The correspondence between a spatial diagram and the spatial layout of the mechanical system likely provides benefits that direct instruction about more abstract mental models cannot.

Mental models can help you understand users If you use the mental model definitions to understand what users know about a system, this can provide an encapsulated look at their knowledge. Asking “how do you think the system works”, “what is the purpose of the system”, “what is the system doing right now?”, “what is the system good at?”, and “what is the system NOT good at” can reveal their clarity and misunderstandings.

Mental models can help identify pathways to expertise Using a concept-map elicitation can help you (and the user) identify missing areas of knowledge, and a comparative study with experts can help identify pathways to expertise. Some evidence from this comes from Chi’s self-explanation research (Chi et al., 1994; Chi, 2000), which showed people repair models via explanation, which in turn predicts learning. Furlough and Gillan (2018) showed that experience changes structure of models, not just amount of facts, and Bayman and Mayer (1984) showed that instructional models shifted novices toward expert-like conceptions, and Schumacher and Czerwinski (1992) argued mental models are critical for the acquisition of expert knowledge.

Mental models provide leverage for simulation-based training Many researchers view mental models as a system that can run inside your head to simulate future outcomes. This can be improved by providing simulation-based training—allowing users to test how changes to “input” impact the outcomes.

Mental models permit users to encapsulate structure rather than learn rules or patterns Conant and Ashby’s Good Regulator theorem suggests that an internal model is necessary for controlling a system. It allows users to make accurate predictions of system behavior in unseen circumstances, and it involves understanding the logical and causal mechanisms that control behavior, rather than rules or patterns.

7.8 Conclusions

Craik (1943) originated the notion of internal models of the world we use for reasoning. It has been a prominent hypothetical construct within cybernetics, human factors, cognitive science, and human-AI interaction research since that time. Although the strong version of Conant and Ashby's *good regulator* theorem is probably not justified, we still understand that control of a system is supported and simplified when we have an internal "model" that helps us explain, interpret, and predict the target.

The rich body of research on mental models frequently takes the existence of mental models as given, and tries to use the concept to structure or interpret data about the interaction between humans and intelligent machines. There are theoretical questions about how they differ from similar knowledge constructs such as "schemas" or even simple knowledge, understanding, and rules. These questions are rarely tested by researchers who use the notion of mental models. They certainly do not map directly onto neural structures in the same way that many other mental structures are known to (e.g., object perception and identification, location representation via grid cells, mirror neurons, emotional reasoning centers, etc.). Even researchers who are working within the framework have admitted that mental models can be limited, are often minimal and incorrect. Nevertheless, the notion is powerful for capturing how humans understand intelligent and complex systems.

Moreover, they can be useful in practical ways—providing a framework for measuring what users know about a system, for improving design, for understanding performance with a system, and for developing training. Furthermore, as the use of the term has expanded, it has begun to encompass many practical aspects (limitations, work-arounds, functional understandings) that further promote these applications. Although mental models may not be "real", they are undoubtedly useful.

For the Human Factors Practitioner

This chapter covers the widely-accepted notion that user mental models underlie their understanding and ability to use automation and AI. However, it remains a fundamentally theory-based concept. In practice, there are a number of ways one can use the idea.

- Mental models can be useful for a number of goals: for understanding and improving human-machine performance, for improving training, and for improving system design. Although it can be helpful to know the outcome when attempting to characterize a mental model, the same interview, think-aloud protocol, or concept-mapping exercise can create artifacts that will benefit many different outcomes.
- Mental models are often fragmented, have gaps, may have inconsistencies, and can be incomplete—especially for novice users but even for experienced ones. It is often more helpful to characterize those gaps than to infer missing information to develop a clearer picture of the user mental model. These gaps can be an important source for improved design, tutorials, and training.
- Consider both functional and mechanistic mental models as a means for helping users. They may not need to know the mechanisms of a system, if they can predict outcomes, boundary conditions, and competency envelopes.
- Eliciting the mental model of the developer and expert user who “know” how a system works can be contrasted with inadequate mental models of novices in order to identify learning objectives and training methods. Showing this contrast can help avoid designer-centered design that is adequate for the developer but hostile to normal users.
- The methods proposed for eliciting mental models are very similar to those discussed in Chapter 6 on measuring trust. These can be combined for more effective and complete pictures of human-automation or human-AI interaction.

Acknowledgments

This chapter was originally developed as part of a course project for HF 5430 Human-AI Interaction, Michigan Technological University, Spring 2024.

Author Contributions

BF, KU, EM, CC: writing and research of individual sections; STM: Conceptualization; background research, writing in all sections; general editing.

Conflicts of Interest

The authors declare no conflict of interest.

AI Usage Statement

Generative AI models were used for additional research, to identify missing concepts, to support better organization, and for editorial tasks such as formatting, evaluating grammar/clarity and citation collation, glossary development, and creation of figures. Content, text, and ideas are otherwise original to the human author contributors.

How to Cite This Chapter

Frisch, B., Ulinski, K., Matas, E., Cischke, C., & Mueller, S. T. (2026). How we understand AI and automation: mental models and transparency. In Shane T. Mueller (Ed.), *A Handbook of Human-AI and Human-Automation Interaction*. https://pages.mtu.edu/~shanem/human_ai/

Glossary

A* search Best-first graph search using actual path cost plus a heuristic estimate toward the goal. (Hart et al., 1968; Jones, 2008)

Absolute Trust Also known as complete trust. The trust that is put into an AI system that it will perform as intended without any verification or supervision, and will behave as expected and make proper choices consistently. (Laplante and Voas, 2022)

Absolute Trusting Absolute Trusting is when the user takes the computer's assertions (data, claims) as valid and true in all circumstances. (Hoffman et al., 2021)

Accountability the giving or demanding of reasons for conduct (León et al., 2020)

ACT-R A computational cognitive architecture (Adaptive Control of Thought–Rational) used to model human performance; in trust research, applied to simulate strategic trust decisions in interaction.

Adaptive cruise control Cruise control that adjusts speed using sensors to maintain distance from a lead vehicle.

Agent The idea of an agent originated with John McCarthy in the mid-1950's, and the term was coined by Oliver G. Selfridge a few years later, when they were both at the Massachusetts Institute of Technology. They had in view a system that, when given a goal, could carry out the details of the appropriate computer operations and could ask for and receive advice, offered in human terms, when it was stuck. An agent would be a "soft robot" living and doing its business within the computer's world. (Kay, 1984; Bradshaw et al., 2017)

AGI (Artificial General Intelligence) The original use of the term "artificial general intelligence" was in a 1997 article about military technologies by Mark Gubrud (Gubrud, 1997), which defined AGI as "AI systems that rival or surpass the human brain in complexity and speed, that can

acquire, manipulate and reason with general knowledge, and that are usable in essentially any phase of industrial or military operations where a human intelligence would otherwise be needed.” (Morris et al., 2023)

Agile development Iterative software development approach emphasizing short cycles, working product increments, and inspect-and-adapt reviews. (Beck et al., 2001)

AI alignment (values) The goal that AI or automation has goals, actions, and values aligned with human intentions—supporting human consideration without necessarily behaving like humans. (Lutkevich, 2023)

AI Bias Because of prejudiced assumptions in the process of developing the model, the output of a machine-learning model may cause discrimination against individuals or specific groups based on gender, social class, sexual orientation or race. However, bias means a deviation from the standard and does not necessarily lead to discrimination. (Belenguer, 2022)

AI-assisted decision making Systems in which AI supports or augments human judgment rather than replacing it outright. (Zhang et al., 2020)

Algorithm Any well-defined computational procedure that takes some value, or set of values, as input and produces some value, or set of values, as output.

Algorithm aversion the reluctance of humans to use algorithms, which are superior but imperfect, in favor of their own judgment (Burton et al., 2020)

Algorithmic decision-making the process of using predefined rules, procedures, or computational algorithms to analyze data and make decisions or predictions (Kochenderfer et al., 2022)

Alignment problem (AI safety) The technical challenge of ensuring AI systems pursue objectives that match human intent and values, especially under uncertainty, misspecified rewards, or unanticipated situations. (Strickland, 2023; Lutkevich, 2023)

Altruistic punishment *[Definition pending.]*

Amershi HAI design guidelines Eighteen evidence-based guidelines for human-AI interaction in four lifecycle contexts (Amershi et al., 2019). (Amershi et al., 2019)

ANI An single-goal oriented AI element specialized to perform a specific task such as a chess engine, facial recognition software, or an email spam filter (Kuusi and Heinonen, 2022)

Anthropomorphic design Designing AI to behave like humans—for predictability, acceptance, or accessibility. (Deshpande et al., 2023)

Anthropomorphic expectations/attributions Users attribute human-like motivations or properties to AI, or expect human-like behavior from systems. (Mueller, 2020)

Anthropomorphism The belief that a non-human entity is human, often through the use of human-like features by the entity. (Nass and Moon, 2000)

Anti-trusting *[Definition pending.]*

Applied Cognitive Task Analysis (ACTA) A lightweight cognitive task analysis package that includes Knowledge Audit probes; useful for eliciting trust-related difficulties with equipment and systems. (Militello and Hutton, 1998)

Appropriate trust *[Definition pending.]*

APT (Argument-based probabilistic trust) *[Definition pending.]*

Artificial Intelligence A machine with the capacity to adapt given that proper data/information is accessible. (Wang, 2019)

Artificial Neural Network a mathematical model that attempts to simulate the structure and functionalities of biological neural networks (Krenker et al., 2011)

ASI (Artificial Super Intelligence) A hypothesis of artificial intelligence, ASI, matches or surpasses the human mind and possesses capabilities exceeding humans. They are even capable of experiencing emotions, relationships, and generating their own code without the need for human intervention. (Keary, 2024)

Auditability terms to a human

Authority Hypothesis Humans perceive computational systems as being better skilled at specific tasks than them (Mosier and Skitka, 1996)

Authority-based trusting *[Definition pending.]*

Automation The process of structuring a task/environment/system to reduce labor and decision-making time through offloading labor and decision-making to a human-created system (Goldberg, 2011)

Automation Bias When errors occur due to the overreliance on automated cues rather than observant seeking and processing

Back Propagation a method for calculating all derivatives of a single target quantity with respect to a large set of input quantities (Werbos, 1990)

Backtracking (search) Search that extends partial solutions and retreats when constraints or goals fail. (Jones, 2008)

Backward chaining Symbolic inference working from goals backward to find supporting facts and rules.

Bag-of-words model Document representation as a multiset of word tokens, ignoring order; basis for many search indices.

Bayes Network A concise and visual representation of the dependencies between variables in a joint probability distribution. (Russell and Norvig, 2021)

Bayesian inference Updating beliefs using Bayes' rule given prior probabilities and observed evidence. (Russell and Norvig, 2021)

- Behavior Tree** A way for an autonomous system to switch between tasks, which allows for multiple decisions as data is collected. (Colledanchise and Ögren, 2018)
- Beneficence** The ought to assist others for their own benefit, which includes prevention of harm. (Halsband and Heinrichs, 2022)
- Bias in computer systems** Systematic and unfair discrimination against certain individuals or groups of individuals in favor of others within the design, implementation, or use of computational technologies. This bias can manifest in various forms, including preexisting bias, which originates from social institutions, practices, and attitudes; technical bias, which arises from technical constraints or considerations; and emergent bias, which emerges in the context of system use. (Friedman and Nissenbaum, 1996)
- Black box** *[Definition pending.]*
- Blackboard System** An AI approach where a central knowledge base (also known as the “blackboard”) is updated continuously with new data and information and applied to new problems. (Nii, 1986)
- Blind trust** *[Definition pending.]*
- Boundary conditions (automation)** The dividing line between situations where an automated or AI system performs reliably and situations where it fails or should not be used; closely related to the competence envelope. (Hoffman et al., 2013)
- Boundary conditions (theory)** While theories provide answers to the “what,” “how,” and “why” questions, boundary conditions refer to the “who, where, when” aspects of a theory. (Busse et al., 2015)
- CAPTCHA** Completely Automated Public Turing test to tell Computers and Humans Apart; challenge tests used to distinguish humans from bots.
- CASA** A framework for understanding how humans treat computers as social agents when interacting with them. (Nass et al., 1994)
- Case-based reasoning (CBR)** Reasoning from a library of prior cases: retrieve, reuse, revise, retain. (Schank, 1983; Watson and Marir, 1994)
- Chain-of-Thought Prompting** Providing an AI model with a series of intermediate reasoning steps to improve the model’s accuracy (Wei et al., 2023)
- Charrette** Intensive participatory planning workshops that engage stakeholders in joint problem-solving to build ownership and reduce conflict. (Lennertz, 2003)
- Chess engines** A type of ANI that specializes in analyzing chess moves and pieces based on predefined moves and strategies to determine the move that will provide the best outcome (Newborn, 2014)
- Classifier or Classification System** A supervised learning technique that uses predetermined classes or groups to assign an object based on its similarity to the groups; can also be used to predict the likelihood of an object fitting into a particular group. (Vilone and Longo, 2021)

- Clumsy automation** Automation that helps during low-workload routine periods but increases human burden during critical peaks. (Wiener, 1989)
- Clustering** Unsupervised grouping of similar data points into clusters. (Halkidi et al., 2001)
- CO-STAR Method** A prompt template developed by GovTech Singapore's Data Science & AI team that considers (C) Context, (O) Objective, (S) Style, (T) Tone, (A) Audience, and (R) Response (Teo, 2024)
- Coactive Design** An approach to the design of Human-Robot Interaction that uses interdependence as the central principle among people and robots who work together in a joint activity (Johnson et al., 2014)
- Coffee test** Wozniak's AGI challenge: enter an average house and make coffee using common-sense knowledge of objects and affordances.
- Cognition-based trust** *[Definition pending.]*
- Cognitive anthropomorphism** The expectation that AI will perform cognitive work in the same ways humans do. (Mueller, 2020)
- Cognitive architecture** AI-hybrid models that embed known human cognitive capabilities and limitations to predict and explain human performance. (Newell, 1990; Byrne, 2003)
- Cognitive Decathlon** A battery of varied knowledge and motor tasks matched to human performance without requiring human-like reasoning underlying performance. (Mueller and Minnery, 2008; Mueller, 2010)
- Cognitive digital twin** A model used to predict human cognitive performance; closely related to computational cognitive models.
- Cognitive Load Theory** Framework for designing instruction that manages intrinsic, extraneous, and germane cognitive load. (Sweller and Chandler, 1991)
- Cognitive Systems Engineering (CSE)** Engineering discipline treating humans as part of automated systems rather than merely operators of them. (Woods and Roth, 1988)
- Cognitive Task Analysis (CTA)** Structured methods for eliciting how people perform tasks and make decisions in context. (Crandall et al., 2006)
- Cognitive Tutorial** An interactive learning tool that leads users through a sequence of interactive steps or activities to help them understand complex topics or processes. (Mueller et al., 2021)
- Cognitive Work Analysis (CWA)** Ecological framework for analyzing work domains to guide tool and automation design. (Vicente, 1999)
- Cognitive Task Analysis** The determination of the cognitive skills, strategies, and knowledge required to perform tasks using specific methods and guidance to probe the cognitive processes. (Militello and Hoffman, 2008)
- Collaboration lifecycle** Stage model of human-AI work spanning task allocation, interaction, feedback, and adoption risks as tools are fielded in teams. (Ahmed et al., 2026)

- Collaborative Filtering** Collaborative filtering is a type of recommender system that can provide recommendations to users based on the ratings of similar users. (Su and Khoshgoftaar, 2009)
- Competence (fidelity)** Whether an AI can accomplish a task at all (Mueller fidelity dimension). (Mueller and Minnery, 2008)
- Competence envelope** The range of situations in which a system performs reliably; users need to know when to trust or distrust the system.
- Componential trust** *[Definition pending.]*
- Computers in-the-loop** Computers act only as helpful tools to the humans who are tasked with accomplishing some goal. The technology is not the driving factor nor an equal party to the humans in the system. (Shneiderman, 2020)
- Concept map** *[Definition pending.]*
- Constraint Satisfaction Problem (CSP)** Problem with variables, domains, and constraints; goal is to find assignments satisfying all constraints. (Jones, 2008)
- Contestability** *[Definition pending.]*
- Context window** *[Definition pending.]*
- Contextual Design** User-centered process using ethnographic field studies to understand work practices. (Beyer and Holtzblatt, 1999)
- Contextual trust** *[Definition pending.]*
- Contingent Trusting** Contingent Trusting is when the user can take some of the computer's presentations or assertions as valid and true under certain circumstances. (Hoffman et al., 2021)
- Contractual Trust** when a trustor has a belief that the trustee will stick to a specific contract; trust between a user and an AI model is trust that some implicit or explicit contract will hold, and a formalization of 'trustworthiness' (Jacovi et al., 2021)
- Controller** *[Definition pending.]*
- Cost Function** Cost functions are a measure of undesirable a particular path is within a graph. The less desirable (due to distance, financial cost, complexity), the higher the cost. (Jones, 2008)
- Counter-trusting** *[Definition pending.]*
- Critical Decision Method** Interview method used in naturalistic decision making to elicit expert judgments about critical incidents. (Klein, 1998)
- Critical periods (HCAI lifecycle)** Stages in AI development when specific human-centered methods have maximal leverage and cost-benefit.
- Cybernetics** Study of control and communication in animals and machines; precursor to cognitive systems engineering. (Wiener, 1948)
- Dark patterns** User-hostile design patterns intended to deceive, confuse, or extract money from users; also transferred to conversational AI. (Lacey and Caudwell, 2019; Stockman and Nottingham, 2024; Traubinger et al., 2024; Dubiel et al., 2024)

- Data Card** Data Cards are structured summaries of essential facts about various aspects of ML datasets needed by stakeholders across a dataset's lifecycle for responsible AI development. These summaries provide explanations of processes and rationales that shape the data and consequently the models—such as upstream sources, data collection and annotation methods; training and evaluation methods, intended use; or decisions affecting model performance. (Pushkarna et al., 2022)
- Deep learning** Neural networks with many layers trained on large datasets, often via backpropagation on GPUs. (LeCun et al., 2015)
- Default trusting** *[Definition pending.]*
- Defensive monitoring** Supervisory checking of automation when reliability is uncertain; treated as a behavioral indicator inversely related to trust. (Muir and Moray, 1996; Adams et al., 2003)
- Dependability** In interpersonal and automation trust models (e.g., Rempel; Muir), a mid-stage basis for trust: confidence that the trustee will behave consistently and can be counted on, beyond short-term predictability. (Rempel et al., 1985; Muir, 1994)
- Design thinking** Human-centered innovation process that cycles through empathize, define, ideate, prototype, and test. (Brown, 2008)
- Desirable difficulty** Learning design that introduces manageable challenge to improve long-term retention and transfer. (Bjork and Bjork, 2020)
- Desirements** Stakeholder wants and requirements for AI systems, distinct from formal specifications or regulations. (Hoffman et al., 2023a)
- Digital Twin** *[Definition pending.]*
- Digressive Trusting** Digressive Trusting is when the user takes fewer of the machine's presentations or assertions as valid and true over time and across experiences. (Hoffman et al., 2021)
- Dijkstra's Algorithm** An algorithm that takes a graph as an input and finds the minimum cost between all pairs of nodes within the graph. (Dijkstra, 1959)
- Dimensionality reduction** Transforming high-dimensional data into fewer dimensions while preserving essential structure. (Baruah, 2023)
- Directability** In coactive design, the ability of teammates to influence or redirect one another's behavior. (Johnson et al., 2014)
- Distributed Cognition** Thinking spread across people, tools, representations, and environmental structures rather than confined to individual minds. (Hutchins et al., 2000; Hutchins, 1995)
- Distributional verisimilitude** Whether the distribution of AI performance matches the distribution of human performance (Mueller fidelity dimension). (Mueller and Minnery, 2008)
- Distrust** *[Definition pending.]*
- Disuse (automation)** Failure to use automation that would be beneficial, often linked to distrust or undertrust relative to system capability. (Parasuraman and Riley, 1997; Lee and See, 2004)

- Domination (fidelity)** Whether an AI can perform a task better, faster, or cheaper than a human (Mueller fidelity dimension). (Mueller and Minnery, 2008)
- Ecological Interface Design** *[Definition pending.]*
- Effective anthropomorphism** Users initially apply human capabilities to understand an unfamiliar AI system—a useful strategy even when the system is not human-like. (Bos et al., 2019)
- ELIZA** Early pattern-matching chatbot that played a Rogerian therapist, mirroring the client rather than giving direct advice. (Weizenbaum, 1966)
- Emotional Connection** A social-anthropomorphic design dimension describing commonly reported emotional engagement with AI. (Gibbons et al., 2023)
- Emulation (human-like AI)** The design goal of creating AI that emulates human capability rather than leveraging distinctive machine strengths. (Shneiderman, 2020)
- Encapsulated trust** Trust limited to known conditions within a system's competence envelope rather than blanket acceptance. (Hoffman et al., 2013)
- Ethopoeia** responding to a non-human entity as if it were human while simultaneously knowing the entity does not require human-style interaction. (Nass and Moon, 2000)
- Evidence-based licensure** *[Definition pending.]* (Tate et al., 2016)
- Evolutionary algorithm** Search inspired by natural selection: evaluate fitness, select, recombine, iterate. (Russell and Norvig, 2021)
- Experiential User Guide** Interactive training letting users learn AI capabilities through guided experience rather than static documentation. (Mueller et al., 2009, 2021)
- Expert System** A high-performing AI program that can generate domain-specific expert level solutions, which might aid in problemsolving in a given area. (Buchanan, 1981)
- Explainability** *[Definition pending.]*
- Explainable AI** XAI can explain their decision or predictions to human users to increase transparency, trustworthiness and accountability of system. (Saranya and Subhashini, 2023)
- Faith-based trusting** *[Definition pending.]*
- Feed-Forward Neural Network** A neural network whose connections only go in one direction; can be thought of as a directed acyclic graph with designated input and output nodes. (Russell and Norvig, 2021)
- Feigenbaum Test** A Turing-style test targeting domain experts (e.g., engineering, physics), emphasizing reasoning over breadth of knowledge. (Feigenbaum, 2003)
- Few-shot prompting** Providing a small number of instructions or examples to an AI model to guide it in performing a task (Prompt Engineering Guide, 2024)

Filter bubble *[Definition pending.]*

Fine-tuning There are two steps in our framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. (Church et al., 2021; Devlin et al., 2019)

formative assessment Lightweight early evaluation used to steer design and development rather than only to demonstrate final effectiveness.

Forward chaining Symbolic inference applying rules to known facts to derive new conclusions (data-driven).

Functional mental models *[Definition pending.]*

Functional understanding Working knowledge of a system sufficient for effective use without deep technical mastery. (Mueller et al., 2009)

Functionality (anthropomorphism space) Capability, accuracy, or ability of AI to accomplish tasks (Gibbons anthropomorphism dimension). (Gibbons et al., 2023)

GAMs Generalized additive models for interpretable machine learning (Chang et al., 2021)

General Problem Solver (GPS) Early AI system searching a problem space via subgoals and operators. (Ernst and Newell, 1967)

Generate and test CSP approach that proposes candidate solutions and checks whether constraints are satisfied. (Jones, 2008)

Generative AI A branch of artificial intelligence that receives user input (such as a prompt) and produces an output based on training data. Outputs of generative AI are vast but include text, audio, video, images, etc. (Brynjolfsson et al., 2023)

generative model An approach to statistical classification in which the underlying patterns of the existing data are examined to create new data similar to the current data in the dataset. (Xu et al., 2015)

Genetic algorithm Evolution-inspired search using selection, crossover, and mutation over a population of candidate solutions. (Holland, 1992)

Glass Box Model Glass box models are transparent about how a conclusion was reached and provide an accurate and human-interpretable explanation. (Carroll and Olson, 1988)

Good regulator theorem Every good regulator of a system must be a model of that system (Conant and Ashby). (Conant and Ashby, 1970)

Gradient descent Iterative optimization that adjusts parameters in the direction that reduces error or cost.

Hallucination A phenomenon that LLMs occasionally generate content that diverges from the user input, contradicts previously generated context, or misaligns with established world knowledge (Zhang et al., 2023)

- heuristic** A heuristic is a rule of thumb that may help solve a given problem. Heuristics take problem knowledge into consideration to help guide the search within the domain. (Jones, 2008)
- Human Readiness Level (HRL)** Nine-level scale characterizing how thoroughly human-centered design and evaluation have matured for a system alongside technological readiness. (See et al., 2019; Human Factors and Ergonomics Society, 2021)
- Human–Computer Trust (HCT)** Madsen and Gregor’s 25-item scale covering perceived reliability, technical competence, understandability, faith, and personal attachment in human–computer trust. (Madsen and Gregor, 2000)
- Human-AI collaboration** Research and design perspective treating humans and AI as cooperative partners on shared tasks. (Wang et al., 2020)
- Human-AI Interaction (HAI)** Interdisciplinary field focused on how people interact with AI-capable systems through interfaces, workflows, and design. (Amershi et al., 2019; Xu et al., 2021)
- Human-AI teaming** Design and research framing humans and AI as teammates with complementary roles on shared goals. (Berretta et al., 2023)
- Human-automation interaction** Study of how people interact with automated systems, including supervision, monitoring, and handoff. (Sheridan and Parasuraman, 2005)
- Human-Autonomy Teaming** Team-focused approach treating automation or AI as a teammate with shared goals and complementary capabilities (also HAT/HMT). (Lyons et al., 2019; Berretta et al., 2023)
- Human-centered AI** A method of thinking about, talking about, and designing AI that focuses on people working together, in charge of technology, and benefiting from information displays; a method of designing AI that amplifies human agency and ability (Shneiderman, 2021)
- Human-in-the-loop** the AI system is able to learn from the human agent, in some capacity, to fill in holes of the knowledge base (Mosqueira-Rey et al., 2023)
- Human-in-the-loop ML** Human-in-the-loop machine learning is a way to involve human operators in the learning process, such that the AI learns in some capacity from the human operator, such as from labeling data or learning from the interactions that a human has with the system (Mosqueira-Rey et al., 2023). Many researchers are attempting to use human-in-the-loop systems to deal with sparse data, as these systems are able to use the prior knowledge that humans have to supplement data (Wu et al., 2022). As AI systems are used to make more important decisions, explainable AI focuses on a human user’s ability to understand how an AI system comes to a conclusion (Dwivedi et al., 2023). (Mosqueira-Rey et al., 2023)
- human-machine teaming** The collaboration between humans and intelligent systems to accomplish a shared goal. (Paleja et al., 2021)

- Human-Robot Interaction (HRI)** Study of how people perceive, communicate with, and work with physical robots across industrial, social, and autonomous forms. (Goodrich and Schultz, 2008)
- Humanization** Treating non-human entities as human in terms of rights, expectations, or interactions. (Spatola et al., 2022)
- Hybrid Intelligence** Hybrid intelligence attempts to utilize the strengths of human intelligence and AI, with the overall goal of creating a better system than either individually (Dellermann et al., 2019). This idea builds on the system 1 and system 2 theory of cognitive processes (Kahneman, 2011). System 1 thinking is fast, emotional, stereotypic, subconscious, and often is considered human intuition. This is where human intelligence is stronger. System 2 thinking is effortful, logical, and often follows strict rules. This is where AI is strongest. By pairing the human and AI together, both can focus on what they are strongest at and create better solutions. (Dellerman, 2019)
- ICAP framework** Learning taxonomy: Interactive, Constructive, Active, and Passive engagement modes (Chi & Wylie). (Chi and Wylie, 2014)
- IEEE P7007 Standard on Transparency** *[Definition pending.]*
- Imitation Game** Turing's original setup: a judge distinguishes two interlocutors via text; in the intelligence version a machine replaces the deceptive participant. (Turing, 1950)
- In-context learning in LLMS** *[Definition pending.]*
- Information retrieval** Systems that index, rank, and retrieve documents or content relevant to a query.
- information transparency** The disclosure of information regarding a particular AI system, especially for reasons revolving around ethics and legality. (Andrada et al., 2022)
- Intelligence** The ability to learn from and adapt to an environment. (Sternberg, 2012)
- Interactive Human-Centered AI (IHCAI)** HCAI practice enabling users to explore, manipulate, and understand AI outputs with human benefit as the design goal. (Schmidt, 2020)
- Interactivity (anthropomorphism space)** Ability to interact with the user and respond to user feedback (Gibbons anthropomorphism dimension). (Gibbons et al., 2023)
- Interpretability** ability to explain or to present in understandable (Doshi-Velez and Kim, 2017)
- Ironies of automation** Automation can leave humans with harder residual tasks (monitoring, exceptions) and erode skills needed when manual control is required. (Bainbridge, 1983)
- Jagged Technological Frontier** *[Definition pending.]* (Dell'Acqua et al., 2023)
- Joint activity** An activity involving two or more parties are participating (Johnson et al., 2014)

- Joint cognitive systems** Perspective analyzing distributed cognitive work across people, tools, and automation in a joint system. (Hollnagel and Woods, 2005)
- Justifiability** The need for an AI system to provide reasoning for its decisions/recommendations. This is especially important when determining why a decision was made by a system and whether the outcome aligns with human morals. (Muralidharan et al., 2024)
- Justified mistrusting** Appropriate skepticism toward a system when its limits or failure modes are understood. (Hoffman, 2017)
- Justified trusting** Calibrated reliance on a system based on understanding its competencies and limits (Hoffman et al.). (Hoffman, 2017)
- Knowledge elicitation bottleneck** Difficulty of extracting expert knowledge to build rule-based or expert systems. (Hayes-Roth et al., 1983)
- Knowledge transfer and exchange (KTE)** the process of exchanging relevant and meaningful knowledge between different agents (Kiefer et al., 2005)
- leakage** in AI, the use of information from outside the training data set; most often occurs when data in the training, validation, and/or testing share indirect information (Bussola et al., 2021)
- Lens model** *[Definition pending.]*
- Levels of automation** Scales distinguishing how much of information acquisition, analysis, decision selection, and action implementation is automated. (Sheridan and Verplank, 1978; Parasuraman et al., 2000)
- LIME** Local Interpretable Model-Agnostic Explanations. An algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model. (Ribeiro et al., 2016)
- LLMs** Large language models (LLMs) refer to transformer-based neural language models that contain tens to hundreds of billions of parameters, and which are pre-trained on massive text data, e.g., PaLM24, LLaMA12, and GPT-4 (Singh et al., 2024)
- Logistic regression** Classification method modeling the probability of a binary outcome rather than a continuous value.
- Loss Function** The loss function (cost function) is used for training a neural network or other machine learning models. Loss functions determine the performance of a deep neural networks. The same framework of deep CNNs with different loss functions may have different training results. According to the application scenario or properties of a loss function, we can design or select a loss function. Since an appropriate loss function will boost the whole performance of a neural network, the traditional machine learning losses are falling out of favor and their counterparts in deep learning are gaining traction. (Tian et al., 2022)
- LSA** A natural language processing technique that involves generating a set of concepts associated with documents and terms to analyze the

relationships between a set of documents and the terms they contain. (Dumais, 2004)

Machine Learning Algorithms that can learn from data, model and make predictions/decisions independently based on these models. (Kufel et al., 2023)

Macro cognition Cognitive processes used in natural settings (Crandall et al., 2006)

Markov model Statistical model where the next state depends only on the current state; basis for n-gram language models. (Shannon, 1951)

material transparency An understanding of the physical system or systems that an AI runs upon, including energy requirements, climate impact, and other socioeconomic factors. (Andrada et al., 2022)

Mayer-Hancock trust framework *[Definition pending.]*

Mechamorphism As opposed to anthropomorphism, where a human being imparts human characteristics to a non-human actor, mechanomorphism is where a human being attempts to modify their thinking to be more consistent with a non-human actor. (Glasgow et al., 2022)

Mechanistic mental models *[Definition pending.]*

Mental Model An internalized representation of a situation or individual founded on previous experiences. (Langan-Fox et al., 2004)

Mental Model Matrix Structured evaluation of user and system capabilities and limitations (Borders et al.). (Borders et al., 2024)

Mental Model of a Dynamic System "A mental model of a dynamic system is a relatively enduring and accessible but limited internal conceptual representation of an external system whose structure maintains the perceived structure of that system." (Doyle and Ford, 1998)

Mentalism Attributing an agent's behavior to internal mental, emotional, or cognitive states. (Spatola et al., 2022)

Metaphor Model *[Definition pending.]* (Carroll and Olson, 1988)

Micro cognition Fundamental cognitive processes underlying other, more complex forms of cognition (Crandall et al., 2006)

Microworld A small closed-loop simulation that captures some complexity of a work domain (e.g., process control) without fully simulating one plant or operational system. (Adams et al., 2003)

Minimum necessary rigor *[Definition pending.]* (Klein et al., 2023)

Mistrusting *[Definition pending.]*

Misuse (automation) Inappropriate reliance on automation, often linked to overtrust—using a system beyond what its capabilities warrant. (Parasuraman and Riley, 1997; Lee and See, 2004)

Mixed Initiative Interaction/Planning A method of planning that involves a general understanding of contributing sub-tasks between constituents to reach a shared outcome. (Novick and Sutton, 1997)

Model Card *[Definition pending.]*

- Monte Carlo Sampling** A computational technique used to predict probabilities of various outcomes relying on repeated random sampling (Kroese et al., 2014)
- Moravec's paradox** It is easy to give machines adult-level performance on tests and games, but hard to give them the perception and mobility skills of a one-year-old. (Moravec, 1988)
- Mutual Theory of Mind (MTOM)** *[Definition pending.]*
- N-gram** Sequence of N consecutive tokens used in language modeling and predictive text. (Shannon, 1951)
- natural language processing** Natural language processing (NLP) is a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory. (Chowdhary, 2020)
- Naturalistic Decision Making** Study of how experts make judgments under time pressure, uncertainty, and real operational constraints. (Klein, 1998)
- Observability** The ability to understand how complex machine learning systems make decisions and interpret data by observing their external outputs. (Watts, 2023)
- One-Shot Prompting** *[Definition pending.]*
- Optimization** the systematic and iterative process of improving the accuracy of a machine learning model or algorithm's output through testing and refinement (Surianarayanan et al., 2023)
- Oracle-based testing** *[Definition pending.]* (Tate et al., 2016)
- Organizational trust** *[Definition pending.]*
- Overfitting** Model parameters fit training data so closely that performance degrades on new cases outside the training set.
- Overtrust** *[Definition pending.]*
- PageRank** an algorithm used to measure the "weight" of nodes within an artificial neural network; weight is based on the number of incoming relationships the node has, as well as the importance of the corresponding sources nodes (Langville and Meyer, 2006)
- Participatory design** Design approach embedding users in the design team or designers in the work context.
- Pathfinder networks** *[Definition pending.]*
- PCA** Principal Component Analysis (PCA) is used for dimensionality reduction in multivariate data and is a widely applied technique in data analysis and machine learning (Baruah, 2023). Its objective is to transform high-dimensional data into a lower-dimensional representation, capturing the most important information. (Baruah, 2023)
- Perceived trustworthiness** *[Definition pending.]*
- PID controller** Proportional-Integral-Derivative controller regulating a process using three tuned feedback components. (Johnson, 2005)

Pre-training There are two steps in our framework: pre-training and fine-tuning. During pre-training, the model is trained on unlabeled data over different pre-training tasks. For fine-tuning, the BERT model is first initialized with the pre-trained parameters, and all of the parameters are fine-tuned using labeled data from the downstream tasks. Each downstream task has separate fine-tuned models, even though they are initialized with the same pre-trained parameters. (Church et al., 2021; Devlin et al., 2019)

Predictability In coactive design, the ability of teammates (human or machine) to anticipate one another's actions and needs. (Johnson et al., 2014)

Premortem Planning exercise in which a team imagines that a project has already failed and generates reasons why, in order to surface risks early. (Veinott et al., 2010; Kannan, 2026)

Procedural Justice fair resolution of disputes and application of administration (Ötting and Maier, 2018)

Production-Rule System A set of rules for fetching data from a database. These production rules are bound to precondition and postcondition that gets checked by the database. If the input content satisfies the conditions, the rules will be successfully applied, resulting in corresponding outcomes. (Khan, 2024)

Progressive Trust a positive evolution of trust between two subjects based on interactions. An initial trust level evolves based on interactions between the two subjects, starting with a test period, trusting each other in situations in which the risks and payoffs are relatively large, and based on the feedback the trust between parties increases or decreases. (Huerta-Canepa et al., 2011)

Progressive Trusting Progressive Trusting is when the user takes more of the machine's presentations or assertions as valid and true over time or across experiences. (Hoffman et al., 2021)

Prompt An instruction from the user that the LLM follows (?)

Prompt Engineering The systematic process of generating and enhancing prompts with the intention of eliciting targeted and favorable responses from large language models LLMs (Aljanabi et al., 2023)

Propensity to trust A trustor disposition or trait-like willingness to trust others (or systems) across situations; an antecedent in Mayer-style organizational trust and Hancock human-machine trust models. (Mayer et al., 1995; Hancock et al., 2011)

Q-learning Model-free reinforcement learning algorithm estimating action values in states. (Li, 2023)

RAG in LLMs [Definition pending.]

rational agent For each possible percept sequence, a rational agent should select an action that is expected to maximize its performance measure,

given the evidence provided by the percept sequence and whatever built-in knowledge the agent has. (Russell and Norvig, 2021)

Recognition-Primed Decision (RPD) *[Definition pending.]*

Recommender System A system (generally machine learning or algorithm-based) that reads prior usage data of a platform or environment (Amazon, Netflix, Spotify) and compares it against demographic data to provide future predictions of behavior (?)

Recurrent Neural Network A neural network in which cycles between nodes are allowed, allowing the network to have a form of memory where inputs from earlier in time can impact the current work being performed. (Russell and Norvig, 2021)

reflective transparency The ability of a user to have understanding, oversight, or ability to intervene in the decision-making processes of an AI system (Andrada et al., 2022)

Reinforcement learning Learning policies through trial-and-error interaction with an environment and reward signals. (Sutton and Barto, 2018)

Reinforcement Learning from Human Feedback (RLHF) A procedure to produce or ensure AI alignment through human feedback on outputs from the AI. (Strickland, 2023)

Reliability Systems are based on proven methods and practices; appropriate technological processes support human responsibility, fairness, and explainability (Shneiderman, 2020)

Reliance Making decisions and acting through AI. Over-reliance on AI can lead to a lack of professional skills or AI errors that are difficult to correct. Thus, it is important to have appropriate supervision and human intervention when relying on AI. (Schemmer et al., 2023)

Resemblance (fidelity) Whether AI behavior reproduces qualitative trends of human performance (Mueller fidelity dimension). (Mueller and Minnery, 2008)

Resilience Engineering Discipline focused on how people and organizations monitor, adapt, and recover under surprise, stress, or failure. (Hollnagel et al., 2006)

Reversal Curse *[Definition pending.]* (Marcus, nd)

Reverse Turing test The user's prompts and subjective intelligence make an LLM appear more intelligent. (Sejnowski, 2023)

Rule-based system An expert system that records data as rules to solve problems (Atanasova and Karashtranova, 2016)

SAE J3016 (Levels of Driving Automation) Six-level taxonomy from fully human-driven to fully autonomous vehicle operation. (SAE On-Road Automated Vehicle Standards Committee, 2014)

Safety *[Definition pending.]*

Scalable Human Oversight integrating human judgment and intervention in the development, deployment, and continuous monitoring of AI systems (Shneiderman, 2020)

Scrum Agile framework characterized by short sprints, defined roles, and regular inspect-and-adapt reviews to deliver incremental improvements. (Schwaber and Sutherland, 2020)

Self-Exfiltration *[Definition pending.]*

Sentiment analysis Classification of text by overall tone or sentiment. (Medhat et al., 2014)

SHAP An algorithm that generates values that attribute to each feature the change in the expected model prediction when conditioning on that feature. (Lundberg and Lee, 2017)

Shapley Value A method based on game theory for assigning payouts to players depending on their contribution to the total payout (Molnar, 2023)

Shared mental models *[Definition pending.]*

Shared responsibility Division of control and accountability between human operators and automated or AI components. (Shneiderman, 2020)

Shifting bottlenecks When automation displaces work, new constraints and workload peaks often emerge elsewhere in the system. (Sarter et al., 1997; Woods, 1996)

Simulacra Elaborate moving sculptures depicting animals, birds, or humans; often meant to amaze rather than replace humans. (Graefe and Bischoff, 2009; Carrier, 2017)

Situated cognition View that cognition arises from interaction among agents and their environment, not isolated information processing. (Clancey, 1997)

Situation awareness Perception, comprehension, and projection of elements in the environment relevant to goals (Endsley). (Endsley, 1995)

Skeptical trusting *[Definition pending.]*

SLAM Simultaneous Localization and Mapping: robot builds a map while estimating its position. (Huang, 2023)

Social Actor Any individual, artificial or physical, that can establish a general understanding with humans while carrying out a conversation. (Nass et al., 1994)

Social Robotics A field of research that seeks to improve robot abilities to mimic human behavior, for example gaze behaviors. (Ruhland et al., 2015)

Socio-technical system Work system comprising people, technology, organization, and culture as an integrated whole. (Clancey, 1997)

Sociotechnical systems Work settings jointly shaped by technical and social subsystems that must be designed together for automation or AI to succeed. (Baxter and Sommerville, 2011)

Soft-max confidence Classifier output scaled to resemble probabilities; often misread as true confidence by users.

Spiritualism Attributing a spiritual nature to non-human agents. (Spatola et al., 2022)

- SQA (Social Q&A)** the process of people asking questions, answering questions, and rating the content of other answers (Gazan, 2011)
- Stable trusting** *[Definition pending.]*
- Stakeholder alignment (adaptive management)** Adaptive management approach aligning diverse stakeholder needs during AI development, deployment, and assessment (Wixom et al.). (Wixom et al., 2020)
- Stakeholder analysis** Identifying and eliciting needs, goals, and constraints of people affected by or involved in an AI system. (Hoffman et al., 2023a; Plass et al., 2022)
- Stakeholder Playbook** Structured prompts and interview guide for stakeholder desirements in human-AI explainability (Hoffman et al.). (Hoffman et al., 2023a)
- Stochastic parrot** Metaphor for language models that produce fluent text via statistical pattern matching without deeper understanding. (Bender et al., 2021)
- Strong AI** Strong AI (also known as Artificial General Intelligence or AGI) is a concept where artificial intelligence possesses a mind similar to humans and can think and reason like humans. The robust and adaptive behavior of Strong AI allows it to perform tasks in various different situations, achieving a level of comprehension or execution of complex tasks that Narrow AI cannot accomplish. (Ng and Leung, 2020)
- Structural mental models** *[Definition pending.]*
- Supervised learning** Machine learning from labeled examples to predict outputs for new cases. (El Naga and Murphy, 2015)
- Supervisory control** Human role overseeing automated operation: monitoring, troubleshooting, and intervening when needed. (Sheridan, 1992)
- Surrogate Model** *[Definition pending.]* (Carroll and Olson, 1988)
- SVD** Singular Value Decomposition (SVD) is a common technique to decompose a matrix into several component matrices, exposing many of the valuable properties of the original matrix (Chengwang L., 2010) (Chengwang, 2010)
- Swift Trust** In temporary groups, people need to work on same tasks. However, they don't have prior history of collaboration, experiences, or interactions to judge each other's. In this situation, they will develop swift trust. Traditionally, swift trust has been applied to human-human trust. Now, there is a need to understand how swift trust is developed for human-robot teams given the proliferation of robots for team tasks. (Meyerson et al., 1996; Haring et al., 2021)
- Symbolic AI** AI using explicit symbolic tokens, rules, and logical pattern matching.
- Technological Readiness Level (TRL)** Nine-level scale of technology maturity used in R&D acquisition, from basic research through operational system use. (Mankins, 1995)

Technology Acceptance Model (TAM) Model of technology adoption in which perceived usefulness and ease of use shape attitude toward using and actual system use.

Tentative trusting *[Definition pending.]*

Theory of Mind *[Definition pending.]*

Thera-Turing test Human behavioral therapists evaluate a medical chatbot's output, requiring correct reasoning and compassion. (Bunge and Desage, 2024)

Think-aloud protocol A method in which participants verbalize thoughts concurrently or retrospectively while performing a task, used to surface trust and trustworthiness attitudes during interaction. (Ericsson and Simon, 1980)

Three R's of trust *[Definition pending.]*

Toulmin model A layout for practical arguments (data, warrant, backing, qualifier, rebuttal, claim) used in argument-based accounts of trust and probabilistic trust (APT). (Toulmin, 1958)

Training vs. deployment Distinction between offline parameter estimation or rule creation and online use of a trained system.

Transparency *[Definition pending.]*

transparency, transformational the ability to understand and assess the broader social, cultural, and ethical implications of AI systems (Andrada et al., 2022)

transparency-in-use When a tool or device becomes sufficiently well understood that using it requires little to no cognitive effort by the user. (Andrada et al., 2022)

Tree of knowledge A conceptual representation that all knowledge expands from micro to macro levels, and can be separated into four key categories: matter, life, mind and, culture (Henriques, 2008)

Trust Trust is the strong belief in someone or something. In AI, trust is the belief the user has in the system and its ability to be trustworthy and align with the characteristics of trustworthiness in AI (accuracy, reliability, resiliency, objectivity, security, explainability, safety, accountability, and privacy). (Stanton and Jensen, 2021)

Trust calibration *[Definition pending.]*

Trust in Automation (TiA) Körber's 19-item, six-subscale self-report questionnaire for trust in automation (Reliability/Competence, Understanding/Predictability, Familiarity, Intention of Developers, Propensity to Trust, Trust in Automation); released under CC BY-SA 4.0. (Körber, 2019a)

Trust in XAI Context (TXAI) Hoffman et al.'s short state trust scale for a named AI tool after use in an explainable-AI setting (eight Likert items plus an open reason; one reverse-coded item). (Hoffman et al., 2021, 2023b)

- Trustworthiness** Regular, honest, and cooperative behavior, based on shared norms and expectations from all agents within a system; using respected independent oversight to create systems that allow users to believe in the design, operation, and maintenance of the system (Ikenberry and Fukuyama, 1996; Shneiderman, 2020)
- Trustworthy AI** For AI to be considered trustworthy, it must be valid and reliable, safe to use, fair and transparent, have accountability, be explainable and interpretable, and protect the privacy of its users. (Tabassi, 2023)
- Turing Test** A test designed originally to answer the question, "Can machines think?" The colloquial understanding is whether or not a computer or piece of software can actually sufficiently human so that another human cannot distinguish the artificial from the truly human. (Turing, 1950)
- Universal Intelligence Measure** An information-theoretic measure of intelligence in reinforcement-learning environments capturing simplicity, adaptation, and generalization. (Legg and Hutter, 2007)
- Unjustified mistrusting** *[Definition pending.]*
- Unjustified trusting** *[Definition pending.]*
- Unsupervised learning** Machine learning that identifies structure in unlabeled data. (El Naqa and Murphy, 2015)
- Unwarranted trust** *[Definition pending.]*
- User expectations of AI** Anticipated performance, legality, safety, transparency, and anthropomorphic behavior of AI systems. (Linja et al., 2022; Zhang et al., 2021)
- Verisimilitude (fidelity)** Whether AI behavior is indistinguishable from a human's (Mueller fidelity dimension). (Mueller and Minnery, 2008)
- Vignette (research)** A written or video scenario that asks participants to imagine themselves in a described situation so attitudes such as trust can be elicited without a live system.
- Warranted trust** *[Definition pending.]*
- Weaponized anthropomorphism** Using anthropomorphic design or interaction to manipulate users (trust, emotion, compliance) rather than support them.
- Wizard of Oz (WOZ)** Evaluation method simulating AI behavior with a hidden human operator to test concepts before full implementation.
- Work domain analysis** Studying how work is performed in context to inform design of tools and automation. (Rasmussen et al., 1990; Vicente, 1999)
- Work-Integrated Learning (WIL)** An educational approach that combines traditional classroom learning with hands-on, real-world work experience (Zegwaard et al., 2023)
- Zero-Shot Prompting** Asking an AI model to perform a task without giving instructions or examples (Prompt Engineering Guide, 2024)

Bibliography

- Abercrombie, G., Curry, A. C., Dinkar, T., Rieser, V., and Talat, Z. (2023). Mirages: On anthropomorphism in dialogue systems. [arXiv:2305.09800](#).
- AbuMusab, S. (2023). Generative AI and human labor: Who is replaceable? *AI & Society*, 39:3051–3053. [doi:10.1007/s00146-023-01773-3](#).
- Adams, B. D., Bruyn, L. E., Houde, S., and Angelopoulos, P. (2003). Trust in automated systems: Literature review. Technical Report CR-2003-096, Defence Research and Development Canada.
- Adams, S. S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., and Sowa, J. F. (2012). Mapping the landscape of human-level artificial general intelligence. *AI Magazine*, 33(1):25–41. [doi:10.1609/aimag.v33i1.2322](#).
- Adjabi, I., Ouahabi, A., Benzaoui, A., and Taleb-Ahmed, A. (2020). Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188. [doi:10.3390/electronics9081188](#).
- Afroogh, S., Akbari, A., Malone, E., Kargar, M., and Alambeigi, H. (2024). Trust in ai: progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11(1):1568. [doi:10.1057/s41599-024-04044-8](#).
- Ahmed, M. F., Anni, I. K., and Rony, M. M. U. (2026). Toward resilient human-AI collaboration: A lifecycle taxonomy of sociotechnical risks and cascading failures. *arXiv preprint*. [arXiv:2608.05614](#).
- Aiken, L. H., Sloane, D. M., Ball, J., Bruyneel, L., Rafferty, A. M., and Griffiths, P. (2021). Patient satisfaction with hospital care and nurses in england: an observational study. *BMJ open*, 8(1):e019189. [doi:10.1136/bmjopen-2017-019189](#).
- Akash, K., Hu, W.-L., Jain, N., and Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, 8(4):1–20. [doi:10.1145/3132743](#).
- Akula, A. R. et al. (2019). Explainable AI and trust.
- Alam, L., Mamun, T. I., and Mueller, S. T. (2025). Application of cognitive empathy elements into ai chatbots: an interview study exploring patient-physician interaction. *Journal of Cognitive Engineering and Decision Making*, 19(2):135–153. [doi:10.1177/15553434241310500](#).
- Aldern, T. (1995). Intelligent system operational support requirements. In *The Human/Electronic Crew: Can We Trust the Team? Proceedings of the 3rd International Workshop on Human-Computer Teamwork*.
- Aljanabi, M., Yaseen, M. G., Ali, A. H., and Mohammed, M. A. (2023). Prompt engineering: Guiding the way to effective large language models. *Iraqi Journal for Computer Science and Mathematics*, pages 151–155. [doi:10.52866/ijcsm.2023.04.04.012](#).

- Allen, J. E., Guinn, C. I., and Horvitz, E. (1999). Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23. doi:10.1109/5254.796083.
- Alon, L. and Levkovich, I. (2026). Trusting the black box: Adapting a multidimensional measure of trust in generative ai. *Computers in Human Behavior: Artificial Humans*, page 100295. doi:10.1016/j.chbah.2026.100295.
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., and Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. doi:10.1145/3290605.3300233.
- Anderson, A., Dodge, J., Sadarangani, A., Juozapaitis, Z., Newman, E., Irvine, J., Chattopadhyay, S., Olson, M., Fern, A., and Burnett, M. (2020). Mental models of mere mortals with explanations of reinforcement learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(2):1–37.
- Anderson, J. R. (1996). Act: A simple theory of complex cognition. *American psychologist*, 51(4):355. doi:10.1037/0003-066x.51.4.355.
- Andersson, J. (2008). Levels of automation and user control — evaluation of a turbine automation interface. Technical report, Chalmers University of Technology. <https://www.osti.gov/etdeweb/biblio/943611>.
- Andrada, G., Clowes, R. W., and Smart, P. R. (2022). Varieties of transparency: exploring agency within AI systems. *AI & Society*, 38(4):1321–1331. doi:10.1007/s00146-021-01326-6.
- Andrews, R. W., Lilly, J. M., Srivastava, D., and Feigh, K. M. (2023). The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175.
- Arai, K. (2009). Defining trust using expected utility theory. *Hitotsubashi Journal of Economics*, 50(2):99–118. doi:10.15057/18045.
- Arbib, M. A., editor (2002). *The Handbook of Brain Theory and Neural Networks*. The MIT Press, Cambridge, MA. doi:10.7551/mitpress/3413.001.0001.
- Arora, A., Arora, A., Sivakumar, K., and Taras, V. (2024). The role of anthropomorphic, xenocentric, intentional, and social (AXIS) robotics in human–robot interaction. *Computers in Human Behavior: Artificial Humans*, 2(1):100036. doi:10.1016/j.chbah.2023.100036.
- Ashtorab, Z., Bouneffouf, D., Brimijoin, K., Bellamy, R., Campbell, M., Goldberg, A., Gonzalez, G. E., Houde, S., Liu, M., Moran, D. A. S., et al. (2025). Bridging the gap: Unifying HCI & ML perspectives on mutual theory of mind. In *International Joint Conference on Artificial Intelligence*.
- Atanasova, I. and Karashtanova, E. (2016). A novel approach for quality of life evaluation: Rule-based expert system. *Social Indicators Research*, 128(2):709–722. doi:10.1007/s11205-015-1052-0.
- Baer, M. D. and Colquitt, J. A. (2018). Why do people trust? In *The Routledge Companion to Trust*, pages 163–182. Routledge. doi:10.4324/9781315745572-12.
- Baier, A. (1986). Trust and antitrust. *Ethics*, 96(2):231–260. doi:10.1086/292745.
- Bainbridge, L. (1983). Ironies of automation. In *Analysis, Design and Evaluation of Man–Machine Systems*, pages 129–135. Pergamon. doi:10.1016/B978-0-08-029348-6.50026-9.
- Bainbridge, L. (1992). Mental models in cognitive skill: the example of industrial process operation. In *Models in the Mind*, pages 119–143. Academic Press.
- Balcázar, J. L., Gavalda, R., and Siegelmann, H. T. (1997). Computational power of neural networks: A characterization in terms of Kolmogorov complexity. *IEEE Transactions on Information Theory*, 43(4):1175–1183. doi:10.1109/18.605580.
- Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., and Horvitz, E. (2019). Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11.
- Barber, B. (1983). *The Logic and Limits of Trust*. Rutgers University Press, New Brunswick, NJ. doi:10.2307/2077058.
- Baroni, I., Re Calegari, G., Scandolari, D., and Celino, I. (2022). AI-TAM: A model to investigate user acceptance and collaborative intention in human-in-the-loop AI applications. *Human*

- Computation*, 9(1):1–21. doi:10.15346/hc.v9i1.134.
- Baruah, I. D. (2023). Dimensionality reduction techniques—PCA, LCA and SVD. Nerd For Tech. <https://medium.com/nerd-for-tech/dimensionality-reduction-techniques-pca-lca-and-svd-f2a56b097f7c>.
- Bassett, C. (2019). The computational therapeutic: Exploring Weizenbaum's ELIZA as a history of the present. *AI & Society*, 34(4):803–812. doi:10.1007/s00146-018-0825-9.
- Baxter, G. and Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. *Interacting with Computers*, 23(1):4–17. doi:10.1016/j.intcom.2010.07.003.
- Bayman, P. and Mayer, R. E. (1984). Instructional manipulation of users' mental models for electronic calculators. *International Journal of Man-Machine Studies*, 20(2):189–199.
- Beck, K., Beedle, M., van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R. C., Mellor, S., Schwaber, K., Sutherland, J., and Thomas, D. (2001). Manifesto for agile software development. <https://agilemanifesto.org>.
- Belenguer, L. (2022). AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. *AI and Ethics*, 2(4):771–787. doi:10.1007/s43681-022-00138-8.
- Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623. doi:10.1145/3442188.3445922.
- Berka, P. (2015). Expert (knowledge-based) systems. In *Encyclopedia of Information Science and Technology*, pages 4555–4563. IGI Global, 3 edition. doi:10.4018/978-1-4666-5888-2.ch446.
- Berretta, S., Tausch, A., Ontrup, G., Gilles, B., Peifer, C., and Kluge, A. (2023). Defining human-AI teaming the human-centered way: A scoping review and network analysis. *Frontiers in Artificial Intelligence*, 6. doi:10.3389/frai.2023.1250725.
- Beyer, H. and Holtzblatt, K. (1999). Contextual design. *interactions*, 6(1):32–42. doi:10.1145/291224.291229.
- Bhaskara, A., Skinner, M., and Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, 50(3):215–224.
- Bjork, R. A. and Bjork, E. L. (2020). Desirable difficulties in theory and practice. *Journal of Applied Research in Memory and Cognition*, 9(4):475–479. doi:10.1016/j.jarmac.2020.09.003.
- Boehm, B. W. (1988). A spiral model of software development and enhancement. *Computer*, 21(5):61–72. doi:10.1109/2.59.
- Boon, S. D. and Holmes, J. G. (1991). The dynamics of interpersonal trust: Resolving uncertainty in the face of risk. In Hinde, R. A. and Groebel, J., editors, *Cooperation and Prosocial Behaviour*, pages 190–211. Cambridge University Press, Cambridge.
- Borders, J., Klein, G., and Besuijen, R. (2024). Mental model matrix: Implications for system design and training. *Journal of Cognitive Engineering and Decision Making*. doi:10.1177/15553434231226317.
- Borenstein, J. and Howard, A. (2021). Emerging challenges in AI and the need for AI ethics education. *AI and Ethics*, 1(1):61–65. doi:10.1007/s43681-020-00002-7.
- Bornstein, A. (2019). AI search algorithms every data scientist should know. Towards Data Science. <https://towardsdatascience.com/ai-search-algorithms-every-data-scientist-should-know-ed0968a43a7a>.
- Borst, C., Flach, J. M., and Ellerbroek, J. (2014). Beyond ecological interface design: Lessons from concerns and misconceptions. *IEEE Transactions on Human-Machine Systems*, 45(2):164–175.
- Borum, R. (2010). The science of interpersonal trust.
- Bos, N., Glasgow, K., Gersh, J., Harbison, I., and Paul, C. L. (2019). Mental models of AI-based systems: User predictions and explanations of image classification results. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 184–188.

- SAGE. doi:10.1177/1071181319631392.
- Brachman, M., Kunde, S., Miller, S., Fucs, A., Dempsey, S., Jabbour, J., and Geyer, W. (2025). Building appropriate mental models: what users know and want to know about an agentic AI chatbot. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 247–264.
- Bradshaw, J. M., Feltoovich, P. J., and Johnson, M. (2017). Human-agent interaction. In *The Handbook of Human-Machine Interaction*, pages 283–300. CRC Press. doi:10.1201/9781315557380-14.
- Branley-Bell, D., Whitworth, R., and Coventry, L. (2020). User trust and understanding of explainable AI: Exploring algorithm visualisations and user biases. In *International Conference on Human-Computer Interaction*, pages 382–399. Springer. doi:10.1007/978-3-030-49108-6_28.
- Breazeal, C., Dautenhahn, K., and Kanda, T. (2016). Social robotics. In *Springer Handbook of Robotics*, pages 1935–1972. Springer. doi:10.1007/978-3-319-32552-1_72.
- Breyer, B. (2015). Soziales vertrauen [Social trust scale]. Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS). doi:10.6102/zis235.
- Brown, O., Power, N., and Gore, J. (2025). Cognitive task analysis: Eliciting expert cognition in context. *Organizational Research Methods*, 28(3):375–404. doi:10.1177/10944281241271216.
- Brown, T. (2008). Design thinking. *Harvard Business Review*, 86(6):84–92.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., et al. (2020). Language models are few-shot learners. *arXiv:2005.14165*.
- Brugha, R. and Varvasovszky, Z. (2000). Stakeholder analysis: A review. *Health Policy and Planning*, 15(3):239–246. doi:10.1093/heapol/15.3.239.
- Brunswik, E. (1943). Organismic achievement and environmental probability. *Psychological Review*, 50(3):255–272. doi:10.1037/h0060889.
- Brynjolfsson, E., Li, D., and Raymond, L. R. (2023). Generative AI at work. NBER Working Paper 31161, National Bureau of Economic Research. doi:10.3386/w31161.
- Buchanan, B. G. (1981). Research on expert systems. *Stanford University California Department of Computer Science*.
- Bunge, E. and Desage, C. (2024). Thera-turing test: A framework for evaluating mental health artificial intelligence-based chatbots. *PsyArXiv*. doi:10.31234/osf.io/wdn2b.
- Burgoon, J. K., Wang, X., Chen, X., Pentland, S. J., and Dunbar, N. E. (2021). Nonverbal behaviors “speak” relational messages of dominance, trust, and composure. *Frontiers in psychology*, 12:624177. doi:10.3389/fpsyg.2021.624177.
- Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239. doi:10.1002/bdm.2155.
- Busse, C., Kach, A. P., and Wagner, S. M. (2015). Boundary conditions: What they are, how to explore them, why we need them, and when to consider them. *SSRN Electronic Journal*. doi:10.2139/ssrn.2713980.
- Bussola, N., Marcolini, A., Maggio, V., Jurman, G., and Furlanello, C. (2021). AI slipping on tiles: Data leakage in digital pathology. In *Lecture Notes in Computer Science Pattern Recognition. ICPR International Workshops and Challenges*, pages 167–182. Springer International Publishing. doi:10.1007/978-3-030-68763-2_13.
- Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology*, 98(1):182–197.
- Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *arXiv:1604.03605*.
- Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., and Durand, F. (2018). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757. doi:10.1109/TPAMI.2018.2815601.
- Byrne, M. D. (2003). Cognitive architecture. *The human-computer interaction handbook: Funda-*

- mentals, evolving technologies and emerging applications, pages 97–117.
- Cabiddu, F., Moi, L., Patriotta, G., and Allen, D. G. (2022). Why do users trust algorithms? a review and conceptualization of initial trust and trust over time. *European Management Journal*, 40(5):685–706. doi:10.1016/j.emj.2022.06.001.
- Cai, J., Fu, X., Gu, Z., and Wu, R. (2024). Validating social service robot interaction trust (ssrit) scale in measuring consumers' trust toward interaction with artificially intelligent (ai) social robots with a chinese sample of adults. *International Journal of Human–Computer Interaction*, 40(16):4319–4334. doi:10.1080/10447318.2023.2212224.
- Calonia, J. (2020). What is tone? Grammarly. <https://www.grammarly.com/blog/tones/>.
- Cañas, A. J., Coffey, J. W., Carnot, M. J., Feltovich, P., Hoffman, R. R., Feltovich, J., and Novak, J. D. (2003). A summary of literature pertaining to the use of concept mapping techniques and technologies for education and performance support. Technical report, Institute for Human and Machine Cognition, Pensacola, FL.
- Cannon-Bowers, J. A. and Salas, E. (2001). Reflections on shared cognition. *Journal of Organizational Behavior*, 22(2):195–202. doi:10.1002/job.82.
- Cannon-Bowers, J. A., Salas, E., and Converse, S. A. (1990). Cognitive psychology and team training: Training shared mental models of complex systems. *Human Factors Society Bulletin*, 33(12):1–4.
- Capiola, A., Baxter, H. C., Pfahler, M. D., Calhoun, C. S., and Bobko, P. (2020). Swift trust in ad hoc teams: A cognitive task analysis of intelligence operators in multi-domain command and control contexts. *Journal of Cognitive Engineering and Decision Making*, 14(3):218–241. doi:10.1177/1555343420943460.
- Card, S. K., Moran, T. P., and Newell, A. (1986). The model human processor- an engineering model of human performance. *Handbook of perception and human performance.*, 2(45–1):1–35.
- Care, C. (2010). *Technology for Modelling: Electrical Analogies, Engineering Practice, and the Development of Analogue Computing*. Springer, London. doi:10.1007/978-1-84882-948-0.
- Carley, K. and Palmquist, M. (1992). Extracting, representing, and analyzing mental models. *Social Forces*, 70(3):601–636. doi:10.1093/sf/70.3.601.
- Carrier, R. (2017). *The Scientist in the Early Roman Empire*. Pitchstone Publishing.
- Carroll, J. M. (1997). Human–computer interaction: Psychology as a science of design. *Annual Review of Psychology*, 48(1):61–83. doi:10.1146/annurev.psych.48.1.61.
- Carroll, J. M. and Carrithers, C. (1984). Training wheels in a user interface. *Communications of the ACM*, 27(8):800–806.
- Carroll, J. M. and Mack, R. L. (1985). Metaphor, computing systems, and active learning. *International journal of man-machine studies*, 22(1):39–57.
- Carroll, J. M. and Olson, J. R. (1987). Mental models in human-computer interaction: Research issues about what the user of software knows. Technical report, National Research Council, Committee on Human Factors, Washington, DC. <https://files.eric.ed.gov/fulltext/ED292465.pdf>.
- Carroll, J. M. and Olson, J. R. (1988). Mental models in human-computer interaction. In Helander, M., editor, *Handbook of Human-Computer Interaction*, pages 45–65. North-Holland, Amsterdam. doi:10.1016/B978-0-444-70536-5.50007-5.
- Cave, S. and Dihal, K. (2018). Ancient dreams of intelligent machines: 3,000 years of robots. *Nature*, 559(7715):473–475. doi:10.1038/d41586-018-05773-y.
- Chakraborti, T. and Kambhampati, S. (2018). Algorithms for the greater good! on mental modeling and acceptable symbiosis in human-AI collaboration. *arXiv preprint*. arXiv:1801.09854.
- Chamola, V., Hassija, V., Sulthana, A. R., Ghosh, D., Dhingra, D., and Sikdar, B. (2023). A review of trustworthy and explainable artificial intelligence (XAI). *IEEE Access*. doi:10.1109/access.2023.3294569.
- Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021). How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 95–105. ACM. doi:10.1145/3447548.3467453.
- Chang, V., Eniola, R. O., Golightly, L., and Xu, Q. A. (2023). An exploration into human–computer

- interaction: Hand gesture recognition management in a challenging environment. *SN Computer Science*, 4(5):441. doi:10.1007/s42979-023-01751-y.
- Chapman, A. and Shigetomi, Y. (2018). Developing national frameworks for inclusive sustainable development incorporating lifestyle factor importance. *Journal of Cleaner Production*, 200:39–47. doi:10.1016/j.jclepro.2018.07.302.
- Chen, J. Y. and Barnes, M. J. (2018). Human-autonomy teaming in military settings. *Theoretical Issues in Ergonomics Science*, 19(3):255–258. doi:10.1080/1463922X.2018.1428998.
- Chengwang, L. (2010). Singular value decomposition in active monitoring data analysis. In *Handbook of Geophysical Exploration: Seismic Exploration*, volume 40, pages 421–430. Elsevier. doi:10.1016/s0950-1401(10)04027-9.
- Cherney, M. A. (2023). AI quietly reshapes apple iphones, watches. Reuters. <https://www.reuters.com/technology/ai-quietly-reshapes-apple-iphones-watches-2023-09-12>.
- Chi, M. T. H. (2000). Self-explaining: The dual processes of generating inference and repairing mental models. In Glaser, R., editor, *Advances in Instructional Psychology*, volume 5, pages 161–238. Lawrence Erlbaum Associates, Mahwah, NJ.
- Chi, M. T. H., De Leeuw, N., Chiu, M.-H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, 18(3):439–477.
- Chi, M. T. H., Feltovich, P. J., and Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science*, 5(2):121–152. doi:10.1207/s15516709cog0502_2.
- Chi, M. T. H. and Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4):219–243. doi:10.1080/00461520.2014.965823.
- Cho, J. (2018). Mental models and home virtual assistants (HVAs). In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6. ACM. doi:10.1145/3170427.3180286.
- Choung, H., David, P., and Ross, A. (2023). Trust in AI and its role in the acceptance of AI technologies. *International Journal of Human–Computer Interaction*, 39(9):1727–1739. doi:10.1080/10447318.2022.2050543.
- Chowdhary, K. R. (2020). Natural language processing. In *Fundamentals of Artificial Intelligence*, pages 603–649. Springer India. doi:10.1007/978-81-322-3972-7_19.
- Chowdhury, S., Dey, P., Joel-Edgar, S., Bhattacharya, S., Rodriguez-Espindola, O., Abadie, A., and Truong, L. (2023). Unlocking the value of artificial intelligence in human resource management through AI capability framework. *Human Resource Management Review*, 33(1):100899. doi:10.1016/j.hrmr.2022.100899.
- Church, K. W., Chen, Z., and Ma, Y. (2021). Emerging trends: A gentle introduction to fine-tuning. *Natural Language Engineering*, 27(6):763–778. doi:10.1017/s1351324921000322.
- Clancey, W. J. (1991). Situated cognition: Stepping out of representational flatland. *AI Communications*, 4(2/3):109–112. doi:10.3233/aic-1991-42-309.
- Clancey, W. J. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge University Press, Cambridge.
- Clancey, W. J. (2006). Observation of work practices in natural settings. In *The Cambridge Handbook of Expertise and Expert Performance*, pages 127–145. Cambridge University Press. doi:10.1017/cbo9780511816796.008.
- Clancey, W. J., Sachs, P., Sierhuis, M., and Van Hoof, R. (1998). Brahms: Simulating practice for work systems design. *International Journal of Human–Computer Studies*, 49(6):831–865. doi:10.1006/ijhc.1998.0229.
- Cohen, M. S., Parasuraman, R., Serfaty, D., and Andes, R. (1997). Trust in decision aids: A model and a training strategy. Technical report, Cognitive Technologies, Inc., Arlington, VA.
- Cohen, P. R. (2005). If not Turing's test, then what? *AI Magazine*, 26(4):61–67.
- Colledanchise, M. and Ögren, P. (2018). *Behavior Trees in Robotics and AI: An Introduction*. CRC Press, Boca Raton, FL. doi:10.1201/9780429489105.
- Collins, A. and Gentner, D. (1987). How people construct mental models. In *Cultural Models in*

- Language and Thought*, pages 243–265. Cambridge University Press.
- Collins, M. G. and Juvina, I. (2021). Trust miscalibration is sometimes necessary: An empirical study and a computational model. *Frontiers in Psychology*, 12:690089. doi:10.3389/fpsyg.2021.690089.
- Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97. doi:10.1080/00207727008920220.
- Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press, Cambridge, UK.
- Crandall, B., Klein, G. A., and Hoffman, R. R. (2006). *Working Minds: A Practitioner's Guide to Cognitive Task Analysis*. MIT Press, Cambridge, MA. doi:10.7551/mitpress/7304.001.0001.
- Crockett, M. J., Clark, L., Hauser, M. D., and Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences*, 107(40):17433–17438. doi:10.1073/pnas.1009396107.
- Cuzzolin, F., Morelli, A., Cirstea, B., and Sahakian, B. J. (2020). Knowing me, knowing you: theory of mind in AI. *Psychological Medicine*, 50(7):1057–1061.
- Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of Oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*, pages 193–200. doi:10.1145/169891.169968.
- Dantzig, G. B. (1982). Reminiscences about the origins of linear programming. *Operations Research Letters*, 1(2):43–48. doi:10.1016/0167-6377(82)90043-8.
- Dasgupta, A. R., Kumari, P., Sahu, H., Amarnath, S., and Nanda, S. (2022). Smart-home automation using AI assistant and IoT. In *2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCTST)*, pages 329–333. doi:10.1109/ICRTCTST54752.2022.9781919.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340. doi:10.2307/249008.
- de Fine Licht, K. and Brölde, B. (2021). On defining “reliance” and “trust”: Purposes, conditions of adequacy, and new definitions. *Philosophia*, 49(5):1981–2001. doi:10.1007/s11406-021-00339-1.
- de Graaf, M. M. A. and Malle, B. F. (2017). How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*. doi:10.1177/1541931213601786.
- de Kleer, J. and Brown, J. S. (1983). Assumptions and ambiguities in mechanistic mental models. In Gentner, D. and Stevens, A. L., editors, *Mental Models*, pages 155–190. Lawrence Erlbaum Associates, Hillsdale, NJ.
- de Visser, E. J., Cohen, M. S., Freedy, A., and Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In Shumaker, R. and Lackey, S., editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, volume 8525 of *Lecture Notes in Computer Science*, pages 251–262. Springer. doi:10.1007/978-3-319-07458-0_24.
- DeChurch, L. A. and Mesmer-Magnus, J. R. (2010a). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology*, 95(1):32–53. doi:10.1037/a0017328.
- DeChurch, L. A. and Mesmer-Magnus, J. R. (2010b). Measuring shared team mental models: A meta-analysis. *Group Dynamics: Theory, Research, and Practice*, 14(1):1–14. doi:10.1037/a0017455.
- Degachi, C., Freire, S. K., Niforatos, E., and Kortuem, G. (2025). Understanding mental models of generative conversational search and the effect of interface transparency. *arXiv preprint. arXiv:2506.03807*.
- Dell'Acqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Kraymer, L., Candelon, F., and Lakhani, K. R. (2023). Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Working Paper 24-013, Harvard Business School. doi:10.2139/ssrn.4573321.

- Dellerman, U. (2019).
- Deng, Y. et al. (2024). Contemporary research on the turing test.
- Desai, M. (2012). *Modeling Trust to Improve Human–Robot Interaction*. PhD thesis, University of Massachusetts Lowell.
- Deshpande, A., Rajpurohit, T., Narasimhan, K., and Kalyan, A. (2023). Anthropomorphization of AI: Opportunities and risks. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 1–7, Singapore. Association for Computational Linguistics. doi:10.18653/v1/2023.nllp-1.1.
- Deshpande, A. and Sharp, H. (2022). Responsible AI systems: Who are the stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pages 227–236. doi:10.1145/3514094.3534187.
- DeSteno, D., Breazeal, C., Frank, R. H., Pizarro, D. A., Baumann, J., Dickens, L., and Lee, J. J. (2012). Detecting the trustworthiness of novel partners in economic exchange. *Psychological Science*, 23(12):1549–1556. doi:10.1177/0956797612448798.
- Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human relations*, 13(2):123–139. doi:10.1177/001872676001300202.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. doi:10.18653/v1/N19-1423.
- Dickens, J., Uvet, H., Bauman, A., and Kwon, H. Y. (2026). The big six model of ai trust: a unifying framework and 20-item scale for knowledge creation. *Journal of Knowledge Management*, pages 1–36. doi:10.1108/jkm-02-2025-0222.
- Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271. doi:10.1007/BF01386390.
- Dodge, J., Khanna, R., Irvine, J., Lam, K.-H., Mai, T., Lin, Z., Kiddle, N., Newman, E., Anderson, A., Raja, S., et al. (2021). After-action review for AI (AAR/AI). In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pages 197–208. ACM. doi:10.1145/3397481.3450683.
- Doshi-Velez, F. and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. [arXiv:1702.08608](https://arxiv.org/abs/1702.08608).
- Doyle, J. K. and Ford, D. N. (1998). Mental models concepts for system dynamics research. *System Dynamics Review*, 14(1):3–29. doi:10.1002/(sici)1099-1727(199821)14:1<3::aid-sdr140>3.0.co;2-k.
- Dubiel, M., Sergeeva, A., and Leiva, L. A. (2024). Impact of voice fidelity on decision making: A potential dark pattern? In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 181–194. doi:10.1145/3640543.3645202.
- Dumais, S. T. (2004). Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230. doi:10.1002/aris.1440380105.
- Earle, T. C. (2004). Thinking aloud about trust: A protocol analysis of trust in risk management. *Risk Analysis*, 24(1):169–183. doi:https://doi.org/10.1111/j.0272-4332.2004.00420.x.
- Ehrlich, K. (1996). Applied mental models in human–computer interaction. In Oakhill, J. and Garnham, A., editors, *Mental Models in Cognitive Science: Essays in Honor of Phil Johnson-Laird*. Psychology Press.
- Eisikovits, N. (2023). Chatbots aren’t becoming sentient, yet we continue to anthropomorphize AI. Fast Company. <https://www.fastcompany.com/90867578/chatbots-arent-becoming-sentient-yet-we-continue-to-anthropomorphize-ai>.
- El Naga, I. and Murphy, M. J. (2015). What is machine learning? In El Naga, I., Li, R., and Murphy, M. J., editors, *Machine Learning in Radiation Oncology*, pages 3–11. Springer. doi:10.1007/978-3-319-18305-3_1.
- Eloy, L., Doherty, E. J., Spencer, C. A., Bobko, P., and Hirshfield, L. (2022). Using fNIRS to identify transparency- and reliability-sensitive markers of trust across multiple timescales in

- collaborative human-human-agent triads. *Frontiers in Neuroergonomics*, 3:838625. doi:10.3389/fnrgo.2022.838625.
- Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64. doi:10.1518/001872095779049543.
- Endsley, M. R. (2019). Human factors & aviation safety. Testimony to the United States House of Representatives, Committee on Transportation and Infrastructure, hearing on The Boeing 737 MAX: Examining the Federal Aviation Administration's Oversight of the Aircraft's Certification. <https://docs.house.gov/meetings/PW/PW00/20191211/110296/HHRG-116-PW00-Wstate-EndsleyM-20191211.pdf>.
- Endsley, M. R. (2023). Ironies of artificial intelligence. *Ergonomics*, 66(11):1656–1668. doi:10.1080/00140139.2023.2213404.
- Ericsson, K. A. and Simon, H. A. (1980). Verbal reports as data. *Psychological review*, 87(3):215. doi:10.1037/0033-295x.87.3.215.
- Ericsson, K. A. and Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.
- Ernst, G. W. and Newell, A. (1967). *GPS: A Case Study in Generality and Problem Solving*. Academic Press, New York.
- European Social Survey (2023). Institutional trust scale (ESS-ITS). OpenScales. <https://openscales.net/scale.php?code=ESS-ITS>.
- Everett, J. A., Claessens, S., Knöchel, T.-D., and Reinecke, M. G. (2026). Principles for understanding trust in artificial intelligence. *Nature Reviews Psychology*, pages 1–14. doi:10.1038/s44159-026-00562-1.
- Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., and Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*, 110:104743. doi:10.1016/j.engappai.2022.104743.
- Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415:137–140. doi:10.1038/415137a.
- Feigenbaum, E. A. (2003). Some challenges and grand challenges for computational intelligence. *Journal of the ACM*, 50(1):32–40. doi:10.1145/602382.602400.
- Fein, R. M., Olson, G. M., and Olson, J. S. (1993). A mental model can help with learning to operate a complex device. In *INTERACT'93 and CHI'93 Conference Companion on Human Factors in Computing Systems*, pages 157–158.
- Fernandes, G. J., Choi, A., Schauer, J. M., Pfammatter, A. F., Spring, B. J., Darwiche, A., and Alshurafa, N. I. (2023). An explainable artificial intelligence software tool for weight management experts (PRIMO): Mixed methods study. *Journal of Medical Internet Research*, 25:e42047. doi:10.2196/42047.
- Ferrario, A., Loi, M., and Viganò, E. (2020). In AI we trust incrementally: A multi-layer model of trust to analyze human–artificial intelligence interactions. *Philosophy & Technology*, 33(3):523–539. doi:10.1007/s13347-019-00378-3.
- Festerling, J. and Siraj, I. (2022). Anthropomorphizing technology: A conceptual review of anthropomorphism research and how it relates to children's engagements with digital voice assistants. *Integrative Psychological and Behavioral Science*, 56(3):709–738. doi:10.1007/s12124-021-09668-y.
- Flach, J. M. and Holden, J. G. (1998). The reality of experience: Gibson's way. *Presence*, 7(1):90–95.
- Ford, D. N. and Serman, J. D. (1998). Expert knowledge elicitation to improve formal and mental models. *System Dynamics Review*, 14(4):309–340. doi:10.1002/(SICI)1099-1727(199824)14:4<309::AID-SDR154>3.0.CO;2-5.
- Francis, B. A. and Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5):457–465. doi:10.1016/0005-1098(76)90006-6.
- Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347. doi:10.1145/230538.230561.

- Fuoli, M. and Paradis, C. (2014). A model of trust-repair discourse. *Journal of Pragmatics*, 74:52–69. doi:<https://doi.org/10.1016/j.pragma.2014.09.001>.
- Furlough, C. S. and Gillan, D. J. (2018). Mental models: structural differences and the role of experience. *Journal of Cognitive Engineering and Decision Making*, 12(4):269–287.
- García-Lopez, A. (2024). Theory of mind in artificial intelligence applications. In *The Theory of Mind Under Scrutiny: Psychopathology, Neuroscience, Philosophy of Mind and Artificial Intelligence*, pages 723–750. Springer.
- García-Peñalvo, F. and Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? a systematic mapping of the evolution, trends, and techniques involved in generative AI. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8(4):7. doi:[10.1002/ijimai.2023.07.006](https://doi.org/10.1002/ijimai.2023.07.006).
- Gautam, S., Liu, H., Choi, Y., and Lease, M. (2026). How researchers navigate accountability, transparency, and trust when using ai tools in early-stage research: A think-aloud study. In *The 2026 ACM Conference on Fairness, Accountability, and Transparency*, pages 1494–1512. doi:[10.1145/3805689.3812262](https://doi.org/10.1145/3805689.3812262).
- Gazan, R. (2011). Social q&a. *Journal of the American Society for Information Science and Technology*, 62(12):2301–2312. doi:[10.1002/asi.21562](https://doi.org/10.1002/asi.21562).
- Gentner, D. and Stevens, A. L. (1983). *Mental Models*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- George, G. C., Iwuanyanwu, E. C., Buford, A. S., Piha-Paul, S. A., Subbiah, V., Fu, S., Karp, D. D., Pant, S., Hinojosa, C. O., Hess, K. R., et al. (2019). Cancer-related internet use and its association with patient decision making and trust in physicians among patients in an early drug development clinic: a questionnaire-based cross-sectional observational study. *Journal of medical Internet research*, 21(3):e10348. doi:[10.2196/10348](https://doi.org/10.2196/10348).
- Gero, K. I., Ashktorab, Z., Dugan, C., Pan, Q., Johnson, J., Geyer, W., Ruiz, M., Miller, S., Millen, D. R., Campbell, M., Kumaravel, S., and Zhang, W. (2020). Mental models of AI agents in a cooperative game setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12. ACM. doi:[10.1145/3313831.3376316](https://doi.org/10.1145/3313831.3376316).
- Gibbons, S., Mugunthan, T., and Nielsen, J. (2023). The 4 degrees of anthropomorphism of generative AI. NN/g Nielsen Norman Group. <https://www.nngroup.com/articles/anthropomorphism/>.
- Gigerenzer, G., Hoffrage, U., and Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, 98(4):506–528. doi:[10.1037/0033-295X.98.4.506](https://doi.org/10.1037/0033-295X.98.4.506).
- Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., and Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in human behavior*, 115:106607. doi:[10.1016/j.chb.2020.106607](https://doi.org/10.1016/j.chb.2020.106607).
- Gillespie, N., Lockey, S., Curtis, C., Pool, J., and Akbari, A. (2023). Trust in artificial intelligence: A global study. Technical report, The University of Queensland and KPMG Australia. doi:[10.14264/00d3c94](https://doi.org/10.14264/00d3c94).
- Glasgow, K., Kopecky, J., Gersh, J., and Crego, A. (2022). Won't you see my neighbor? user predictions, mental models, and similarity-based explanations of AI classifiers. In *AHFE International Artificial Intelligence and Social Computing*. AHFE International. doi:[10.54941/ahfe1001440](https://doi.org/10.54941/ahfe1001440).
- Glikson, E. and Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, 14(2):627–660. doi:[10.5465/annals.2018.0057](https://doi.org/10.5465/annals.2018.0057).
- Goertzel, B. (2014). Artificial general intelligence: Concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1–48. doi:[10.2478/jagi-2014-0001](https://doi.org/10.2478/jagi-2014-0001).
- Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Doubleday.
- Goillau, P., Kelly, C., Boardman, M., and Jeannot, E. (2001). Development of a measure of trust in ATM systems. Technical report, European Organisation for the Safety of Air Navigation (EUROCONTROL).
- Goldberg, K. (2011). What is automation? *IEEE Transactions on Automation Science and*

- Engineering*, 9(1):1–2. doi:10.1109/TASE.2011.2178470.
- Goodrich, M. A. and Schultz, A. C. (2008). Human-robot interaction: A survey. *Foundations and Trends in Human-Computer Interaction*, 1(3):203–275. doi:10.1561/1100000005.
- Gould, J. D. and Lewis, C. (1985). Designing for usability: Key principles and what designers think. *Communications of the ACM*, 28(3):300–311. doi:10.1145/3166.3170.
- Graefe, V. and Bischoff, R. (2009). From ancient machines to intelligent robots—a technical evolution. In *2009 9th International Conference on Electronic Measurement & Instruments*, pages 3–418–3–421. IEEE. doi:10.1109/icemi.2009.5274297.
- Grammarly (2019). Introducing Grammarly's tone detector. Grammarly Blog. <https://www.grammarly.com/blog/tone-detector/>.
- Gratton, L. (2022). *Redesigning Work: How to Transform Your Organization and Make Hybrid Work for Everyone*. MIT Press, Cambridge, MA.
- Green, P. and Wei-Haas, L. (1985). The rapid development of user interfaces: Experience with the wizard of Oz method. In *Proceedings of the Human Factors Society Annual Meeting*, volume 29, pages 470–474. doi:10.1177/154193128502900515.
- Grenier, R. S. and Dudzinska-Przesmitzki, D. (2015). A conceptual model for eliciting mental models using a composite methodology. *Human Resource Development Review*, 14(2):163–184. doi:10.1177/1534484315575966.
- Grimes, G. M., Schuetzler, R. M., and Giboney, J. S. (2021). Mental models and expectation violations in conversational AI interactions. *Decision Support Systems*, 144:113515.
- Gu, Y., Tafjord, O., Kim, H., Moore, J., Le Bras, R., Clark, P., and Choi, Y. (2024). SimpleToM: Exposing the gap between explicit ToM inference and implicit ToM application in LLMs. *arXiv:2410.13648*.
- Guinness, H. (2023). What is artificial general intelligence (AGI)? Zapier. <https://zapier.com/blog/artificial-general-intelligence/>.
- Gulati, S., Sousa, S., and Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour & Information Technology*, 38(10):1004–1015. doi:10.1080/0144929x.2019.1656779.
- Guo, L., Zhan, J., Kou, G., and Martínez, L. (2024). A sentiment analysis and dual trust relationship-based approach to large-scale group decision-making for online reviews: A case study of China Eastern Airlines. *Information Sciences*, 667:120515. doi:10.1016/j.ins.2024.120515.
- Gupta, K., Hajika, R., Pai, Y. S., Duenser, A., Lochner, M., and Billinghurst, M. (2020). Measuring human trust in a virtual assistant using physiological sensing in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 756–765. doi:10.1109/VR46266.2020.00099.
- Halasz, F. G. and Moran, T. P. (1983). Mental models and problem solving in using a calculator. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 212–216. ACM. doi:10.1145/800045.801613.
- Hald, K., Rehm, M., and Moeslund, T. B. (2020). Human-robot trust assessment using motion tracking & galvanic skin response. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6282–6287. doi:10.1109/IROS45743.2020.9340944.
- Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). Clustering algorithms and validity measures. In *Proceedings Thirteenth International Conference on Scientific and Statistical Database Management (SSDBM 2001)*, pages 3–22. doi:10.1109/SSDM.2001.938534.
- Halsband, A. and Heinrichs, B. (2022). AI, suicide prevention and the limits of beneficence. *Philosophy & Technology*, 35(4). doi:10.1007/s13347-022-00599-z.
- Hancock, P. A., Billings, D. R., Oleson, K. E., and Chen, J. Y. C. (2015). Trust and human–robot interactions. *Human Factors*.
- Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors*, 53(5):517–527. doi:10.1177/0018720811417254.
- Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., and Szalma, J. L. (2021). Evolving trust in robots: Specification through sequential and comparative meta-analyses. *Human Factors*, 63(7):1196–1229. doi:10.1177/0018720820922080.

- Hanna, L. L., Reifegerste, D., Berg, A., Memenga, P., Baumann, E., Weber, W., Geulen, J., Müller, A., Hahne, A., and Weg-Remers, S. (2023). Understanding trust determinants in a live chat service on familial cancer: Qualitative triangulation study with focus groups and interviews in Germany. *Journal of Medical Internet Research*, 25. doi:10.2196/44707.
- Harbers, M. et al. (2009). Cited in de graaf and malle (2017) discussion of agency attributions.
- Hardiman, P. T., Dufresne, R., and Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. *Memory & Cognition*, 17(5):627–638. doi:10.3758/BF03197085.
- Hardin, R. (1996). Trustworthiness. *Ethics*, 107(1):26–42. doi:10.1086/233695.
- Hardin, R. (2002). *Trust and Trustworthiness*. Russell Sage Foundation, New York.
- Haring, K. S., Phillips, E., Lazzara, E. H., Ullman, D., Baker, A. L., and Keebler, J. R. (2021). Applying the swift trust model to human-robot teaming. In *Trust in Human-Robot Interaction*, pages 407–427. Elsevier. doi:10.1016/b978-0-12-819472-0.00017-4.
- Harnad, S. (2000). Minds, machines and turing. *Journal of Logic, Language and Information*, 9(4):425–445. doi:10.1023/a:1008315308862.
- Harnad, S. (2008). The annotation game: On Turing (1950) on computing, machinery, and intelligence. In *Parsing the Turing Test*, pages 23–66. Springer.
- Hart, P. E., Nilsson, N. J., and Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107. doi:10.1109/TSSC.1968.300136.
- Hayes-Roth, F., Waterman, D. A., and Lenat, D. B. (1983). *Building Expert Systems*. Addison-Wesley, Reading, MA.
- Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York. doi:10.2307/1418888.
- Henriques, G. R. (2008). Special section: The problem of psychology and the integration of human knowledge. *Theory & Psychology*, 18(6):731–755. doi:10.1177/0959354308097255.
- High-Level Expert Group on Artificial Intelligence (2019). Ethics guidelines for trustworthy AI. Technical report, European Commission. doi:10.2759/346720.
- Hoch, T., Martinez-Gil, J., Pichler, M., Silvina, A., Heinzl, B., Moser, B., others, and Leva, M. C. (2023). Multi-stakeholder perspective on human-AI collaboration in industry 5.0. In *Artificial Intelligence in Manufacturing*, pages 407–421. Springer, Cham. doi:10.1007/978-3-031-46452-2_23.
- Hoffman, R., Miller, T., Mueller, S. T., Klein, G., and Clancey, W. J. (2018a). Explaining explanation, part 4: a deep dive on deep nets. *IEEE Intelligent Systems*, 33(3):87–95.
- Hoffman, R. R. (2017). A taxonomy of emergent trusting in the human-machine relationship. In *Cognitive Systems Engineering*, pages 137–164. CRC Press. doi:10.1201/9781315572529-8.
- Hoffman, R. R. (2018). A taxonomy of emergent trusting in the human-machine relationship. *Cognitive Systems Engineering*. doi:10.1201/9781315572529-8.
- Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, 28(1):84–88. doi:10.1109/MIS.2013.24.
- Hoffman, R. R., Klein, G., Jalaeian, M., Tate, C., and Mueller, S. T. (2023a). Explainable AI: Roles and stakeholders, desires and challenges. *Frontiers in Computer Science*, 5:1117848. doi:10.3389/fcomp.2023.1117848.
- Hoffman, R. R., Klein, G., and Mueller, S. T. (2018b). Explaining explanation for “explainable ai”. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 197–201. SAGE Publications.
- Hoffman, R. R., Mueller, S. T., and Klein, G. (2017). Explaining explanation, part 2: Empirical foundations. *IEEE Intelligent Systems*, 32(4):78–86.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2018c). Metrics for explainable ai: Challenges and prospects. *arXiv preprint*. arXiv:1812.04608.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2021). Measuring trust in the XAI context. *PsyArXiv*. doi:10.31234/osf.io/e3kv9.
- Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2023b). Measures for explainable

- AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human–AI performance. *Frontiers in Computer Science*, 5:1096257. doi:10.3389/fcomp.2023.1096257.
- Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267(1):66–73. doi:10.1038/scientificamerican0792-66.
- Holland, J. H. (1999). *Emergence: From Chaos to Order*. Oxford University Press, Oxford. doi:10.1093/oso/9780198504092.001.0001.
- Hollnagel, E. and Woods, D. D. (2005). *Joint Cognitive Systems*. CRC Press. doi:10.1201/9781420038194.
- Hollnagel, E., Woods, D. D., and Leveson, N., editors (2006). *Resilience Engineering: Concepts and Precepts*. Ashgate, Aldershot, UK. doi:10.1136/qshc.2006.018390.
- Holstein, G., Vaughan, J. W., Daumé III, H., Dudik, M., and Wallach, H. (2019). Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 600:1–600:16. doi:10.1145/3290605.3300831.
- Hoy, M. B. (2018). Alexa, Siri, Cortana, and more: An introduction to voice assistants. *Medical Reference Services Quarterly*, 37(1):81–88. doi:10.1080/02763869.2018.1404391.
- Huang, F. (2023). The research progress of path planning algorithms. *Highlights in Science, Engineering and Technology*, 63:216–221. doi:10.54097/hset.v63i.10879.
- Huerta-Canepa, G., Lee, D., and Han, S. Y. (2011). Trust ME: A trust decision framework for mobile environments. In *2011 IEEE 10th International Conference on Trust, Security and Privacy in Computing and Communications*, pages 464–471. IEEE. doi:10.1109/trustcom.2011.60.
- Human Factors and Ergonomics Society (2021). ANSI/HFES 400-2021: Human readiness level scale in the system development process. Technical report, Human Factors and Ergonomics Society, Washington, DC. <https://www.hfes.org/publications>.
- Hutchins, E. (1995). How a cockpit remembers its speeds. *Cognitive science*, 19(3):265–288. doi:10.1207/s15516709cog1903_1.
- Hutchins, E., Hollan, J. D., and Kirsh, D. (2000). Distributed cognition. *ACM Transactions on Computer-Human Interaction*, 7(2):174–196. doi:10.1145/353485.353487.
- Ibne Mamun, T. (2023). *Investigating Collaborative Explainable AI (CXAI)/Social Forum as an Explainable AI (XAI) Method in Autonomous Driving (AD)*. PhD thesis, Michigan Technological University. doi:10.37099/mtu.dc.etrdr/1678.
- Ibne Mamun, T. and Mueller, S. T. (2024). Exploring learning paths: Understanding the learning strategies of artificial intelligence system users and their involvement in social forums. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 1501–1506. SAGE Publications Sage CA: Los Angeles, CA. doi:10.1177/10711813241260668.
- Ikenberry, G. J. and Fukuyama, F. (1996). Trust: The social virtues and the creation of prosperity. *Foreign Affairs*, 75(2):143. doi:10.2307/20047503.
- International Organization for Standardization (2019). Ergonomics of human-system interaction—part 210: Human-centred design for interactive systems. Technical Report ISO 9241-210:2019, ISO, Geneva. doi:10.3403/30388991u.
- Itti, L. and Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10–12):1489–1506. doi:10.1016/S0042-6989(99)00163-7.
- Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 624–635. doi:10.1145/3442188.3445923.
- Jagacinski, R. J. and Flach, J. M. (2003). *Control Theory for Humans: Quantitative Approaches to Modeling Performance*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Jannach, D., Zanker, M., Felfernig, A., and Friedrich, G. (2010). *Recommender Systems: An Introduction*. Cambridge University Press, Cambridge. doi:10.1007/978-0-387-85820-3_6.

- Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., and Capiola, A. (2019). The measurement of the propensity to trust automation. In Chen, J. Y. C. and Fragomeni, G., editors, *Virtual, Augmented and Mixed Reality. Applications and Case Studies*, volume 11575 of *Lecture Notes in Computer Science*, pages 476–489. Springer. doi:10.1007/978-3-030-21565-1_32.
- Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics*, 4(1):53–71. doi:10.1207/S15327566IJCE0401_04.
- Johnson, M., Bradshaw, J. M., Feltoich, P. J., Jonker, C. M., Van Riemsdijk, M. B., and Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69. doi:10.5898/JHRI.3.1.Johnson.
- Johnson, M. A. (2005). PID control technology. In Johnson, M. A. and Moradi, M. H., editors, *PID Control*, pages 1–46. Springer-Verlag. doi:10.1007/1-84628-148-2_1.
- Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science*, 4(1):71–115.
- Johnson-Laird, P. N. (2004). The history of mental models. In Manktelow, K. and Chung, M. C., editors, *Psychology of Reasoning: Theoretical and Historical Perspectives*, pages 179–212. Psychology Press, Hove, UK.
- Jones, K. (1996). Trust as an affective attitude. *Ethics*, 107(1):4–25. doi:10.1086/233694.
- Jones, K. (2012). Trustworthiness. *Ethics*, 123(1):61–85. doi:10.1086/667838.
- Jones, M. T. (2008). *Artificial Intelligence: A Systems Approach*. Jones & Bartlett Learning. doi:10.5860/choice.45-6216.
- Jongerius, C., Twisk, J. W., Romijn, J. A., Callemeyn, T., Goedemé, T., Smets, E. M., and Hillen, M. A. (2022). The influence of face gaze by physicians on patient trust: an observational study: Jongerius et al. *Journal of General Internal Medicine*, 37(6):1408–1414. doi:10.1007/s11606-021-06906-2.
- Judd, T., Durand, F., and Torralba, A. (2012). A benchmark of computational models of saliency to predict human fixations. Technical report, Massachusetts Institute of Technology.
- Juvina, I., Collins, M. G., Larue, O., Kennedy, W., de Visser, E., and de Melo, C. (2019). Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions on Interactive Intelligent Systems*, 9:1–33. doi:10.1145/3230735.
- Juvina, I., Lebiere, C., and Gonzalez, C. (2015). Modeling trust dynamics in strategic interaction. *Journal of Applied Research in Memory and Cognition*, 4:197–211. doi:10.1016/j.jarmac.2014.09.004.
- Kaelbling, L. P., Littman, M. L., and Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4:237–285. doi:10.1613/jair.301.
- Kannan, A. (2026). *Designing with Foresight: Premortem and Human-Centered Evaluation of Emerging Technologies*. Open access dissertation, Michigan Technological University. doi:10.37099/mtu.dc.etr/2053.
- Kaplan, A. D., Kessler, T. T., Brill, J. C., and Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human factors*, 65(2):337–359. doi:10.1177/00187208211013988.
- Kaufmann, T., Weng, P., Bengs, V., and Hüllermeier, E. (2023). A survey of reinforcement learning from human feedback. [arXiv:2312.14925](https://arxiv.org/abs/2312.14925).
- Kay, A. "Computer Software." *Scientific American* 251, n. . (1984). : 53-59.
- Keary, T. (2024). Artificial superintelligence (ASI). *Techopedia*.
- Kelly, C. (2001). Human-computer trust.
- Kennon, N., Howden, P., and Hartley, M. (2009). Who really matters?: A stakeholder analysis tool. *Extension Farming Systems Journal*, 5(2):9–17.
- Khan, A. (2024). Production system in ai. *ntelliPaat*.
- Kiefer, L., Frank, J., Di Ruggiero, E., Dobbins, M., Manuel, D., Gully, P. R., and Mowat, D. (2005). Fostering evidence-based decision-making in canada: Examining the need for a canadian population and public health evidence centre and research network. *Canadian Journal of Public Health*, 96(3):11–119.
- Kieras, D. E. and Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science*, 8(3):255–273. doi:10.1207/s15516709cog0803_3.
- Kiesler, S., Powers, A., Fussell, S. R., and Torrey, C. (2008). Anthropomorphic interactions with a

- robot and robot-like agent. *Social Cognition*, 26(2):169–181. doi:10.1521/soco.2008.26.2.169.
- Kim, M., Kim, S., Kim, J., Song, T.-J., and Kim, Y. (2024). Do stakeholder needs differ?—designing stakeholder-tailored explainable artificial intelligence (XAI) interfaces. *International Journal of Human-Computer Studies*, 181:103160. doi:10.1016/j.ijhcs.2023.103160.
- King, H., Williams, B., Treanor, D., and Randell, R. (2022). How, for whom, and in what contexts will artificial intelligence be adopted in pathology? A realist interview study. *Journal of the American Medical Informatics Association*, 30(3):529–538. doi:10.1093/jamia/ocac254.
- Klein, G., Hoffman, R. R., Clancey, W. J., Mueller, S. T., Jentsch, F., and Jalaeian, M. (2023). “minimum necessary rigor” in empirically evaluating human–AI work systems. *AI Magazine*, 44(3):274–281. doi:10.1002/aaai.12108.
- Klein, G., Hoffman, R. R., Mueller, S. T., and Rasmussen, L. J. (2010). Naturalistic model of causal reasoning: Developing an experiential user’s guide (EUG) to understand fusion algorithms and simulation models. Final report AFRL Contract FA8650-04-D-6546, Delivery Order #13, Air Force Research Laboratory. doi:10.21236/ada538096.
- Klein, G., Moon, B., and Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent systems*, 21(5):88–92.
- Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., and Feltovich, P. J. (2004). Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE Intelligent Systems*, 19(6):91–95. doi:10.1109/mis.2004.74.
- Klein, G. and Wright, C. (2016). Macrocognition: from theory to toolbox. *Frontiers in psychology*, 7:54. doi:10.3389/fpsyg.2016.00054.
- Klein, G. A. (1998). *Sources of Power: How People Make Decisions*. MIT Press, Cambridge, MA. doi:10.7551/mitpress/11307.001.0001.
- Klein, G. A., Calderwood, R., and Macgregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(3):462–472. doi:10.1109/21.31053.
- Klug, K. and Strang, C. (2019). The filter bubble in social media communication: How users evaluate personalized information in the facebook newsfeed. In *Media Trust in a Digital World*, pages 159–180. Springer International Publishing. doi:10.1007/978-3-030-30774-5_12.
- Kochenderfer, M. J., Wheeler, T. A., and Wray, K. H. (2022). *Algorithms for Decision Making*. The MIT Press.
- Kocienda, K. (2018). The iPhone’s autocorrect is a blessing and a curse. *Wired*. <https://www.wired.com/story/opinion-i-invented-autocorrect/>.
- Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2021). Measurement of trust in automation: A narrative review and reference guide. *Frontiers in Psychology*, 12:604977. doi:10.3389/fpsyg.2021.604977.
- Kopka, M., Majeed, A., Spinelli, G., El-Osta, A., and Feufel, M. (2026). The trust in ai-generated health advice (taigha) scale and short version (taigha-s): Development and validation study. *PLOS Digital Health*, 5(7):e0001488. doi:10.1371/journal.pdig.0001488.
- Körber, M. (2019a). Theoretical considerations and development of a questionnaire to measure trust in automation. In Bagnara, S., Tartaglia, R., Albolino, S., Alexander, T., and Fujita, Y., editors, *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018): Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics*, pages 13–30. Springer. doi:10.1007/978-3-319-96074-6_2.
- Körber, M. (2019b). TiA trust in automation questionnaire. GitHub repository. https://github.com/moritzkoerber/TiA_Trust_in_Automation_Questionnaire.
- Körber, M., Baseler, E., and Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, 66:18–31. doi:https://doi.org/10.1016/j.apergo.2017.07.006.
- Kosfeld, M., Heinrichs, M., Zak, P. J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. *Nature*, 435:673–676. doi:10.1038/nature03701.
- Kostick-Quenet, K., Lang, B. H., Smith, J., Hurley, M., and Blumenthal-Barby, J. (2023). Trust criteria for artificial intelligence in health: Normative and epistemic considerations. *Journal of*

- Medical Ethics*. doi:10.1136/jme-2023-109338.
- Krackhardt, D. (1987). Cognitive social structures. *Social networks*, 9(2):109–134. doi:10.1016/0378-8733(87)90009-8.
- Krenker, A., Bester, J., and Kos, A. (2011). Introduction to the artificial neural networks. In *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. InTech. doi:10.5772/15751.
- Kroese, D. P., Brereton, T., Taimre, T., and Botev, Z. I. (2014). Why the monte carlo method is so important today. *WIREs Computational Statistics*, 6(6):386–392. doi:10.1002/wics.1314.
- Kufel, J., Bargiel-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., and Gruszczyńska, K. (2023). What is machine learning, artificial neural networks and deep learning?—examples of practical applications in medicine. *Diagnostics*, 13(15):2582. doi:10.3390/diagnostics13152582.
- Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., and Wong, W.-K. (2013). Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10. IEEE.
- Kümmerer, M., Wallis, T. S. A., Gatys, L. A., and Bethge, M. (2017). Understanding low- and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808. doi:10.1109/ICCV.2017.513.
- Kuusi, O. and Heinonen, S. (2022). Scenarios from artificial narrow intelligence to artificial general intelligence—reviewing the results of the international work/technology 2050 study. *World Futures Review*, 14(1):65–79. doi:10.1177/19467567221101637.
- Lacey, C. and Caudwell, C. (2019). Cuteness as a 'dark pattern' in home robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, pages 374–381. doi:10.1109/HRI.2019.8673274.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240. doi:10.1037/0033-295X.104.2.211.
- Langan-Fox, J., Anglim, J., and Wilson, J. R. (2004). Mental models, team mental models, and performance: Process, development, and future directions. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 14(4):331–352. doi:10.1002/hfm.20004.
- Langley, C., Cirstea, B. I., Cuzzolin, F., and Sahakian, B. J. (2022). Theory of mind and preference learning at the interface of cognitive science, neuroscience, and AI: A review. *Frontiers in Artificial Intelligence*, 5:778852.
- Langville, A. N. and Meyer, C. D. (2006). Information retrieval and web search. In *Handbook of Linear Algebra*, pages 63–1–63–16. Chapman and Hall/CRC. doi:10.1201/9781420010572-63.
- Lankton, N. K., McKnight, D. H., and Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, 16(10):880–918. doi:10.17705/1jais.00411.
- Laplace, P. and Voas, J. (2022). Zero-trust artificial intelligence? *Computer*, 55(2):10–12. doi:10.1109/mc.2021.3126526.
- Larasati, R. (2025). Human and ai trust: trust attitude measurement instrument. *arXiv preprint. arXiv:2510.21535*.
- Larasati, R., De Liddo, A., and Motta, E. (2023). Human and ai trust: Trust attitude measurement instrument development. In *Workshop on Trust and Reliance in AI-Assisted Tasks (TRAIT) at CHI 2023*, Hamburg, Germany.
- LeCun, Y. (1987). *Modèles connexionnistes de l'apprentissage (Connectionist Learning Models)*. PhD thesis, Université Pierre et Marie Curie.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444. doi:10.1038/nature14539.
- LeCun, Y., Boser, B. E., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. E., and Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551. doi:10.1162/neco.1989.1.4.541.
- Lee, J. and Moray, N. (1992). Trust, control strategies and allocation of function in human–machine

- systems. *Ergonomics*, 35(10):1243–1270. doi:10.1080/00140139208967392.
- Lee, J. D. (2013). Trust and human–robot interaction.
- Lee, J. D. and Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40(1):153–184. doi:10.1006/ijhc.1994.1007.
- Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80. doi:10.1518/hfes.46.1.50.30392.
- Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., and DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, 4:893. doi:10.3389/fpsyg.2013.00893.
- Legg, S. and Hutter, M. (2007). Universal intelligence: A definition of machine intelligence. *Minds and Machines*, 17:391–444. doi:10.1007/s11023-007-9079-x.
- Leichtenstern, K., Bee, N., André, E., Berk Müller, U., and Wagner, J. (2011). Physiological measurement of trust-related behavior in trust-neutral and trust-critical situations. In *Trust Management V*, pages 165–172. Springer. doi:10.1007/978-3-642-22200-9_14.
- Lenneretz, B. (2003). The charrette as an agent for change. In *New Urbanism: Comprehensive Report & Best Practices Guide*, pages 12–2. New Urban Publications, Ithaca, NY, 3 edition.
- León, G. A., Chiou, E. K., and Wilkins, A. (2020). Accountability increases resource sharing: Effects of accountability on human and AI system performance. *International Journal of Human-Computer Interaction*, 37(5):434–444. doi:10.1080/10447318.2020.1824695.
- Levin, D. T. et al. (2013). Cited in de graaf and malle (2017).
- Lewis, J. A. and Crumpler, W. (2021). How does facial recognition work? Center for Strategic & International Studies. <https://www.csis.org/analysis/how-does-facial-recognition-work>.
- Li, F. and Betts, S. C. (2003). Trust relationships in business.
- Li, H. and Aral, S. (2025). Human trust in ai search: A large-scale experiment. *arXiv preprint arXiv:2504.06435*. doi:10.2139/ssrn.5210258.
- Li, S. E. (2023). *Reinforcement Learning for Sequential Decision and Optimal Control*. Springer. doi:10.1007/978-981-19-7784-8.
- Liang, C.-J., Kang, S.-C., and Lee, M.-H. (2017). RAS: A robotic assembly system for steel structure erection and assembly. *International Journal of Intelligent Robotics and Applications*, 1:459–476. doi:10.1007/s41315-017-0030-x.
- Liao, Q. V., Gruen, D., and Miller, S. (2020). Questioning the AI: Informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15. doi:10.1145/3313831.3376590.
- Linja, A., Mamun, T. I., and Mueller, S. T. (2022). When self-driving fails: Evaluating social media posts regarding problems and misconceptions about Tesla's FSD mode. *Multimodal Technologies and Interaction*, 6(10):86. doi:10.3390/mti6100086.
- Lithmee (2019). What is the difference between linear and nonlinear programming. Pediaa. <https://pediaa.com/what-is-the-difference-between-linear-and-nonlinear-programming/>.
- Liu, H. and Wang, L. (2018). Gesture recognition for human-robot collaboration: A review. *International Journal of Industrial Ergonomics*, 68:355–367. doi:10.1016/j.ergon.2017.02.004.
- Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., and Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. doi:10.1162/tacl_a_00638.
- Llinas, J., Bisantz, A., Drury, C., Seong, Y., and Jian, J.-Y. (1998). Studies and analyses of aided adversarial decision making. phase 2: Research on human trust in automation. Technical report, Center for Multisource Information Fusion, State University of New York at Buffalo. doi:10.21236/ada370937.
- Lu, L., Su, B., Jung, S., Qing, L., and Xu, X. (2025). Effects of anthropomorphic features of co-robots on users' perception and emotional responses during human–robot collaboration. *Ergonomics*, 69(6):937–951. doi:10.1080/00140139.2025.2497523.

- Lu, Y. and Sarter, N. (2019). Eye tracking: A process-oriented method for inferring trust in automation as a function of priming and system reliability. *IEEE Transactions on Human-Machine Systems*, 49(6):560–568. doi:10.1109/THMS.2019.2929712.
- Luczak, H., Roetting, M., and Schmidt, L. (2003). Let's talk: Anthropomorphization as means to cope with stress of interacting with technical devices. *Ergonomics*, 46(13-14):1361–1374. doi:10.1080/00140130310001610883.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. arXiv:1705.07874.
- Lutkevich, B. (2023). Ai alignment. *TechTarget*.
- Lyons, J. B., Wynne, K. T., Mahoney, S., and Roebke, M. A. (2019). Trust and human-machine teaming: A qualitative study. In *Artificial Intelligence for the Internet of Everything*, pages 101–116. Academic Press. doi:10.1016/b978-0-12-817636-8.00006-5.
- Mack, R. L., Lewis, C. H., and Carroll, J. M. (1983). Learning to use word processors: Problems and prospects. *ACM Transactions on Information Systems (TOIS)*, 1(3):254–271.
- Madsen, M. and Gregor, S. (2000). Measuring human–computer trust. In *Proceedings of the 11th Australasian Conference on Information Systems*.
- Malle, B. F. and Ullman, D. (2021). A multidimensional conception and measure of human–robot trust. In Nam, C. S. and Lyons, J. B., editors, *Trust in Human–Robot Interaction*, pages 3–25. Academic Press, San Diego, CA. doi:10.1016/b978-0-12-819472-0.00001-0.
- Mankins, J. C. (1995). Technology readiness levels: A white paper. Technical report, NASA Office of Space Access and Technology.
- Mao, Y., Liu, S., Ni, Q., Lin, X., and He, L. (2024). A review on machine theory of mind. *IEEE Transactions on Computational Social Systems*, 11(6):7114–7132.
- Marchetti, A., Manzi, F., Riva, G., Gaggioli, A., and Massaro, D. (2025). Artificial intelligence and the illusion of understanding: A systematic review of theory of mind and large language models. *Cyberpsychology, Behavior, and Social Networking*, 28(7):505–514.
- Marcus, G. (2018). Deep learning: A critical appraisal. arXiv:1801.00631.
- Marcus, G. (n.d.). The reversal curse.
- Marjeh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., and Griffiths, T. L. (2024). Large language models predict human sensory judgments across six modalities. *Scientific Reports*, 14(1):21445. doi:10.1038/s41598-024-72071-1.
- Markus, A. F., Kors, J. A., and Rijnbeek, P. R. (2021). The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113:103655. doi:10.1016/j.jbi.2020.103655.
- Martínez, N., Agudo, U., and Matute, H. (2022). Human cognitive biases present in artificial intelligence. *Revista Internacional de los Estudios Vascos (RIEV)*, 67(2). <https://www.eusko-ikaskuntza.eus/en/riev/human-cognitive-biases-present-in-artificial-intelligence/rart-24782/>.
- Matta, D. (2026). Artificial intelligence and theory of mind. *Journal of Psychology and AI*, 2(1):2628373.
- Matthews, G., Lin, J., Panganiban, A. R., and Long, M. D. (2019). Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems*, 50(3):234–244.
- Matthews, M. L., Webb, R. D. G., and Bryant, D. J. (1999). Cognitive task analysis of the HALIFAX-class operations room officer. Technical report, Department of National Defence, Ottawa, ON. doi:10.1037/e683312011-016.
- Mayer, R. C. and Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, 84(1):123–136. doi:10.1037/0021-9010.84.1.123.
- Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734. doi:10.5465/amr.1995.9508080335.
- Mayer, R. E. (1980). Elaboration techniques that increase the meaningfulness of technical text: An experimental test of the learning strategy hypothesis. *Journal of educational psychology*,

- 72(6):770.
- McClelland, J. L., Rumelhart, D. E., and Hinton, G. E. (1986). The appeal of parallel distributed processing. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations*, pages 3–44. MIT Press, Cambridge, MA. doi:10.1016/b978-1-4832-1446-7.50010-8.
- McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5:115–133. doi:10.1007/bf02478259.
- McGrath, M. J., Lack, O., Tisch, J., and Duenser, A. (2025). Measuring trust in artificial intelligence: Validation of an established scale and its short form. *Frontiers in Artificial Intelligence*, 8:1582880. doi:10.3389/frai.2025.1582880.
- McKnight, D. and Chervany, N. (2001). Conceptualizing trust: a typology and e-commerce customer relationships model. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, page 10. IEEE Comput. Soc. doi:10.1109/hicss.2001.927053.
- McLean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A., and Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5):649–663. doi:10.1080/0952813x.2021.1964003.
- Medhat, W., Hassan, A., and Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113. doi:10.1016/j.asej.2014.04.011.
- Meize-Grochowski, R. (1984). An analysis of the concept of trust. *Journal of Advanced Nursing*, 9(6):563–572. doi:10.1111/j.1365-2648.1984.tb00412.x.
- Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, 53(4):356–370. doi:10.1177/0018720811411912.
- Meyer, D. E. and Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. basic mechanisms. *Psychological Review*, 104(1):3–65. doi:10.1037/0033-295X.104.1.3.
- Meyerson, D., Weick, K. E., and Kramer, R. M. (1996). Swift trust and temporary groups. In *Trust in Organizations: Frontiers of Theory and Research*, pages 166–195. SAGE Publications, Inc. doi:10.4135/9781452243610.n9.
- Militello, L. G. and Hoffman, R. R. (2008). The forgotten history of cognitive task analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 52, pages 383–387. SAGE Publications Sage CA: Los Angeles, CA.
- Militello, L. G. and Hutton, R. J. B. (1998). Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics*, 41(11):1618–1641. doi:10.1080/001401398186108.
- Minsky, M. and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. The MIT Press, Cambridge, MA.
- Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37(3):311–324. doi:10.1109/TSMCC.2007.893280.
- Mohammadi, A. and Hashemi Golpayegani, S. A. (2021). Sensetrust: A sentiment based trust model in social network. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(6):2031–2050. doi:10.3390/jtaer16060114.
- Mohan, S. and Laird, J. (2012). Exploring mixed-initiative interaction for learning with situated instruction in cognitive agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 26, pages 2445–2446. doi:10.1609/aaai.v26i1.8424.
- Molnar, C. (2023). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Leanpub, 2 edition. <https://leanpub.com/interpretable-machine-learning>.
- Monett, D., Lewis, C., and Thórisson, K. R. (2020). Introduction to the JAGI special issue ‘On defining artificial intelligence’—commentaries and author's response. *Journal of Artificial General Intelligence*, 11(2):1–100. doi:10.2478/jagi-2020-0003.
- Monett, D. and Lewis, C. W. (2018). Getting clarity by defining artificial intelligence—a survey. In *Philosophy and Theory of Artificial Intelligence 2017*, pages 212–214. Springer. doi:

- [10.1007/978-3-319-96448-5_21](https://doi.org/10.1007/978-3-319-96448-5_21).
- Monroe, A. et al. (2014). Cited in de graaf and malle (2017).
- Montag, C. and Ali, R. (2025). Can we assess attitudes toward ai with single items? associations with existing attitudes toward ai measures and trust in chatgpt. *Journal of Technology in Behavioral Science*, pages 1–11. doi:[10.1007/s41347-025-00481-7](https://doi.org/10.1007/s41347-025-00481-7).
- Montag, C., Becker, B., and Li, B. J. (2024). On trust in humans and trust in artificial intelligence: A study with samples from singapore and germany extending recent research. *Computers in Human Behavior: Artificial Humans*, 2(2):100070. doi:[10.1016/j.chbah.2024.100070](https://doi.org/10.1016/j.chbah.2024.100070).
- Montag, C. and Elhai, J. D. (2025). Introduction of the ai-interaction positivity scale and its relations to satisfaction with life and trust in chatgpt. *Computers in Human Behavior*, 172:108705. doi:[10.1016/j.chb.2025.108705](https://doi.org/10.1016/j.chb.2025.108705).
- Montag, C., Klugah-Brown, B., Zhou, X., Wernicke, J., Liu, C., Kou, J., Chen, Y., Haas, B. W., and Becker, B. (2023a). Trust toward humans and trust toward artificial intelligence are not associated: Initial insights from self-report and neurostructural brain imaging. *Personality Neuroscience*, 6:e3. doi:[10.1017/pen.2022.5](https://doi.org/10.1017/pen.2022.5).
- Montag, C., Kraus, J., Baumann, M., and Rozgonjuk, D. (2023b). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, 11:100315. doi:[10.1016/j.chbr.2023.100315](https://doi.org/10.1016/j.chbr.2023.100315).
- Montague, P. R., Lohrenz, T., and Dayan, P. (2015). The three R's of trust. *Current Opinion in Behavioral Sciences*, 3:102–106. doi:[10.1016/j.cobeha.2015.02.009](https://doi.org/10.1016/j.cobeha.2015.02.009).
- Moon, B., Hoffman, R. R., Novak, J. D., and Cañas, A. J., editors (2011). *Applied Concept Mapping: Capturing, Analyzing, and Organizing Knowledge*. CRC Press, Boca Raton, FL.
- Moravec, H. (1988). *Mind Children: The Future of Robot and Human Intelligence*. Harvard University Press, Cambridge, MA. doi:[10.2307/1575314](https://doi.org/10.2307/1575314).
- Moray, N. and Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, 21(4–5):203–211. doi:[10.1177/014233129902100408](https://doi.org/10.1177/014233129902100408).
- Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. (2023). Levels of AGI for operationalizing progress on the path to AGI. [arXiv: 2311.02462](https://arxiv.org/abs/2311.02462).
- Mosier, K. L. and Kirlik, A. (2004). Brunswik's lens model in human factors research: Modern applications of a classic theory. *PsycEXTRA Dataset*. doi:[10.1037/e577272012-016](https://doi.org/10.1037/e577272012-016).
- Mosier, K. L. and Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other? In Parasuraman, R. and Mouloua, M., editors, *Automation and Human Performance: Theory and Applications*, pages 201–220. Lawrence Erlbaum Associates, Mahwah, NJ.
- Mosier, K. L. and Skitka, L. J. (1999). Automation use and automation bias. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 43, pages 344–348. SAGE. doi:[10.1177/154193129904300346](https://doi.org/10.1177/154193129904300346).
- Mosqueira-Rey, E., Hernández-Pereira, E., Alonso-Ríos, D., Bobes-Bascarán, J., and Fernández-Leal, Á. (2023). Human-in-the-loop machine learning: A state of the art. *Artificial Intelligence Review*, 56(4):3005–3054. doi:[10.1007/s10462-022-10246-w](https://doi.org/10.1007/s10462-022-10246-w).
- Mueller, S. T. (2010). A partial implementation of the BICA cognitive decathlon using the psychology experiment building language (PEBL). *International Journal of Machine Consciousness*, 2(2):273–288. doi:[10.1142/s1793843010000497](https://doi.org/10.1142/s1793843010000497).
- Mueller, S. T. (2020). Cognitive anthropomorphism of AI: How humans and computers classify images. *Ergonomics in Design*, 28(3):12–19. doi:[10.1177/1064804620920870](https://doi.org/10.1177/1064804620920870).
- Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., and Klein, G. (2019). Explanation in human-ai systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable ai. [arXiv preprint. arXiv:1902.01876](https://arxiv.org/abs/1902.01876).
- Mueller, S. T., Klein, G., and Burns, C. (2009). Experiencing the tool without experiencing the pain: Concepts for an experiential user guide. In *Proceedings of the 9th Bi-annual International Conference on Naturalistic Decision Making (NDM9)*, pages 105–112. doi:

- 10.14236/ewic/ndm2009.10.
- Mueller, S. T. and Minnery, B. S. (2008). Adapting the Turing test for embodied virtual agents: The cognitive decathlon. In *Proceedings of the Artificial General Intelligence Conference*.
- Mueller, S. T., Tan, Y.-Y., Linja, A., Klein, G., and Hoffman, R. R. (2021). Authoring guide for cognitive tutorials for artificial intelligence: Purposes and methods. Technical report, DARPA Explainable AI Program. doi:10.31234/osf.io/g7umc.
- Muir, B. M. (1994). Trust in automation: Part I. theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37(11):1905–1922. doi:10.1080/00140139408964957.
- Muir, B. M. and Moray, N. (1996). Trust in automation. Part II. experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3):429–460. doi:10.1080/00140139608964474.
- Muralidharan, A., Savulescu, J., and Schaefer, G. O. (2024). AI and the need for justification (to the patient). *Ethics and Information Technology*, 26(1). doi:10.1007/s10676-024-09754-w.
- Muramatsu, J. and Pratt, W. (2001). Transparent queries: Investigating users' mental models of search engines. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 217–224. ACM. doi:10.1145/383952.383991.
- Mustofa, R. H., Kuncoro, T. G., Atmono, D., Hermawan, H. D., and Sukirman (2025). Extending the technology acceptance model: The role of subjective norms, ethics, and trust in AI tool adoption among students. *Computers and Education: Artificial Intelligence*, 8:100379. doi:10.1016/j.caeai.2025.100379.
- Naikar, N., Hoffman, R., Roth, E. M., Klein, G., Militello, L. G., and Dominguez, C. (2026). Should we make ai more tool-like or teammate-like? *Journal of Cognitive Engineering and Decision Making*, 20(2):137–157. doi:10.1177/15553434251346904.
- Nass, C., Fogg, B. J., and Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, 45(6):669–678. doi:10.1006/ijhc.1996.0073.
- Nass, C. and Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1):81–103. doi:https://doi.org/10.1111/0022-4537.00153.
- Nass, C., Steuer, J., Tauber, E., and Reeder, H. (1993). Anthropomorphism, agency, and ethopoeia: Computers as social actors. In *INTERACT '93 and CHI '93 Conference Companion on Human Factors in Computing Systems*, pages 111–112. doi:10.1145/259964.260137.
- Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Conference companion on Human factors in computing systems - CHI '94*, page 204. ACM Press. doi:10.1145/259963.260288.
- National Institute of Standards and Technology (2019). U.S. leadership in AI: A plan for federal engagement in developing technical standards and related tools. Technical report, NIST. https://www.nist.gov/system/files/documents/2019/08/10/ai_standards_fedengagement_plan_9aug2019.pdf.
- Nave, G., Camerer, C., and McCullough, M. (2015). Does oxytocin increase trust in humans? a critical review of research. *Perspectives on Psychological Science*, 10(6):772–789. doi:10.1177/1745691615600138.
- Nebreda, A., Shpakivska-Bilan, D., Camara, C., and Susi, G. (2024). The social machine: artificial intelligence (AI) approaches to theory of mind. In *The Theory of Mind Under Scrutiny: Psychopathology, Neuroscience, Philosophy of Mind and Artificial Intelligence*, pages 681–722. Springer.
- Newborn, M. (2014). *Computer chess*. Academic Press.
- Newell, A. (1990). *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.
- Ng, G. W. and Leung, W. C. (2020). Strong artificial intelligence and consciousness. *Journal of Artificial Intelligence and Consciousness*, 07(01):63–72. doi:10.1142/s2705078520300042.
- Ngo, T., Kunkel, J., and Ziegler, J. (2020). Exploring mental models for transparent and controllable recommender systems: a qualitative study. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 183–191.
- Nielsen, J. (1993). *Usability Engineering*. Academic Press, Boston. doi:10.1016/

- b978-0-08-052029-2.50007-3.
- Nielsen, J. (2023). 4 degrees of anthropomorphism in using AI. Jakob Nielsen on UX (Substack). <https://jakobnielsenphd.substack.com/p/anthropomorphism>.
- Nii, H. P. (1986). The blackboard model of problem solving and the evolution of blackboard architectures. *AI magazine*, 7(2), 38-38.
- Nisbett, R. E. and Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological review*, 84(3):231. doi:10.1037/0033-295x.84.3.231.
- Niu, B., Ren, J., Zhao, A., and Li, X. (2020). Lender trust on the p2p lending: Analysis based on sentiment analysis of comment text. *Sustainability*, 12(8):3293. doi:10.3390/su12083293.
- Noonan, B. et al. (2008). Trust in education.
- Norman, D. A. (1983). Some observations on mental models. In Gentner, D. and Stevens, A. L., editors, *Mental Models*. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Novak, J. D. and Gowin, D. B. (1984). *Learning How to Learn*. Cambridge University Press, Cambridge.
- Novick, D. G. and Sutton, S. (1997). What is mixed-initiative interaction? In *Proceedings of the AAAI Spring Symposium on Computational Models for Mixed Initiative Interaction*, volume 2, page 12. doi:10.1016/s0167-6393(97)00049-6.
- Office of Management and Budget (2024a). Fact sheet: OMB issues guidance to advance the responsible acquisition of AI in government. <https://www.whitehouse.gov/omb/briefing-room/2024/10/03/fact-sheet-omb-issues-guidance-to-advance-the-responsible-acquisition-of-ai-in-go>
- Office of Management and Budget (2024b). Instructions for 2024 agency AI reporting per EO 14110. Technical report, Executive Office of the President. <https://www.whitehouse.gov/wp-content/uploads/2024/08/Instructions-for-2024-Agency-AI-Reporting-per-EO-14110.pdf>.
- Ogorelysheva, N., Vasileva, A., and Gramse, N. (2023). On troubleshooting in AGV-based autonomous systems. In *2023 International Conference on Control, Automation and Diagnosis (ICCAD)*, pages 1–6. doi:10.1109/ICCAD57653.2023.10152445.
- Oh, S., Seong, Y., Yi, S., and Park, S. (2020). Neurological measurement of human trust in automation using electroencephalogram. *International Journal of Fuzzy Logic and Intelligent Systems*, 20(4):261–271. doi:10.5391/IJFIS.2020.20.4.261.
- Orrange, S., Patel, A., Mack, W. J., and Cassetta, J. (2021). Patient satisfaction and trust in telemedicine during the covid-19 pandemic: retrospective observational study. *JMIR human factors*, 8(2):e28589. doi:10.2196/28589.
- Őrsi, B., Kovács, J., and Csukonyi, C. (2024). Accepting a robot request contradicting a human instruction in the function of robot attitudes and level of interdependency. *Computers in Human Behavior Reports*, page 100385. doi:10.1016/j.chbr.2024.100385.
- Ötting, S. K. and Maier, G. W. (2018). The importance of procedural justice in human-machine interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, 89:27–39. doi:10.1016/j.chb.2018.07.022.
- Paleja, R., Ghuy, M., Arachchige, N. R., Jensen, R., and Gombolay, M. (2021). The utility of explainable AI in ad hoc human-machine teaming. In *Advances in Neural Information Processing Systems*, volume 34, pages 610–623. <https://proceedings.neurips.cc/paper/2021/hash/05d74c48b5b30514d8e9bd60320fc8f6-Abstract.html>.
- Palmer, S., Richards, D., Shelton-Rayner, G., Izzetoglu, K., and Inch, D. (2020). Assessing variable levels of delegated control—a novel measure of trust. In *HCI International 2020—Late Breaking Papers: Cognition, Learning and Games*, pages 202–215. Springer. doi:10.1007/978-3-030-60128-7_16.
- Parakh, T., Bowers, R., and Feigh, K. (2025). Using mental models to understand the effect of AI transparency on trust. In *Workshop on Theory of Mind (ToM) at the 34th International Joint Conference on Artificial Intelligence (IJCAI 2025)*, Montreal, Canada. https://tomworkshop.github.io/assets/acceptedPapers/7_UUsing_Mental_Models_AI_Transparency.pdf.
- Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse.

- Human Factors*, 39(2):230–253. doi:10.1518/001872097778543886.
- Parasuraman, R., Sheridan, T. B., and 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. doi:10.1109/3468.844354.
- Pariser, E. (2011). *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Press, New York. doi:10.1353/pla.2011.0036.
- Park, H., Ahn, D., Hosanagar, K., and Lee, J. (2022). Designing fair AI in human resource management: Understanding tensions surrounding algorithmic evaluation and envisioning stakeholder-centered solutions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–22. doi:10.1145/3491102.3517672.
- Paul, E., Fancourt, D., and Razai, M. (2022). Racial discrimination, low trust in the health system and covid-19 vaccine uptake: a longitudinal observational study of 633 uk adults from ethnic minority groups. *Journal of the Royal Society of Medicine*, 115(11):439–447. doi:10.1177/01410768221095241.
- Perrig, S. A. C., Scharowski, N., and Brühlmann, F. (2023). Trust issues with trust scales: Examining the psychometric quality of trust measures in the context of AI. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–7. doi:10.1145/3544549.3585598.
- Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., and Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214. doi:10.1126/science.abe2629.
- Pinto, A., Sousa, S., Simões, A., and Santos, J. (2022). A trust scale for human-robot interaction: translation, adaptation, and validation of a human computer trust scale. *Human Behavior and Emerging Technologies*, 2022(1):6437441. doi:10.1155/2022/6437441.
- Plass, M., Kargl, M., Evans, T., Brcic, L., Regitnig, P., Geißler, C., others, and Müller, H. (2022). Human-AI interfaces are a central component of trustworthy AI. In *Explainable AI: Foundations, Methodologies and Applications*, pages 225–256. Springer, Cham. doi:10.1007/978-3-031-12807-3_11.
- Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S., and Carey, T. (1994). *Human–Computer Interaction*. Addison-Wesley, Wokingham, UK.
- Prompt Engineering Guide (2024). Prompt engineering guide. <https://www.promptingguide.ai/>.
- Pushkarna, M., Zaldivar, A., and Kjartansson, O. (2022). Data cards: Purposeful and transparent dataset documentation for responsible AI. In *2022 ACM Conference on Fairness Accountability and Transparency*, pages 1776–1826. ACM. doi:10.1145/3531146.3533231.
- Qinjun, L., Tianwei, C., Yan, Z., and Yuying, W. (2023). Facial recognition technology: A comprehensive overview. *Academic Journal of Computing & Information Science*, 6(7):15–26. doi:10.25236/ajcis.2023.060703.
- Ragone, G., Buono, P., Good, J., and Lanzilotti, R. (2026). Do children trust ai, and should they? designing and validating a child-centred k-ai trust scale for intelligent systems. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–18. doi:10.1145/3772318.3790765.
- Ramaraj, P., Sahay, S., Kumar, S. H., Lasecki, W. S., and Laird, J. E. (2019). Towards using transparency mechanisms to build better mental models. In *Advances in cognitive systems: 7th goal reasoning workshop*, volume 7, pages 1–6.
- Rampersad, G. (2020). Robot will take your job: Innovation for an era of artificial intelligence. *Journal of Business Research*, 116:68–74. doi:10.1016/j.jbusres.2020.05.019.
- Rasmussen, J. (1979). On the structure of knowledge—a morphology of mental models in a man-machine context. *Roskilde, Denmark: Riso National Laboratory*.
- Rasmussen, J., Pejtersen, A. M., and Goodstein, L. P. (1994). *Cognitive Systems Engineering*. Wiley, New York.
- Rasmussen, J., Pejtersen, A. M., and Schmidt, K. (1990). Taxonomy for cognitive work analysis. Technical report, Riso National Laboratory, Roskilde.
- Reed, M. S., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., others, and Stringer,

- L. C. (2009). Who's in and why? a typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management*, 90(5):1933–1949. doi:10.1016/j.jenvman.2009.01.001.
- Reeves, B. and Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press. doi:10.5860/choice.34-3702.
- Reiersen, J. (2017). Trust as belief or behavior? *Review of Social Economy*, 75(4):434–453. doi:10.1080/00346764.2016.1269939.
- Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, 49(1):95–112. doi:10.1037/0022-3514.49.1.95.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144. ACM. doi:10.1145/2939672.2939778.
- Rieder, B. and Hofmann, J. (2020). Towards platform observability. *Internet Policy Review*, 9(4). doi:10.14763/2020.4.1535.
- Riedl, D., Rothmund, M. S., Grote, V., Fischer, M. J., Kampling, H., Kruse, J., Nolte, T., Labek, K., and Lampe, A. (2023). Mentalizing and epistemic trust as critical success factors in psychosomatic rehabilitation: results of a single center longitudinal observational study. *Frontiers in psychiatry*, 14:1150422. doi:10.3389/fpsy.2023.1150422.
- Riedl, M. O. (2019). Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36. doi:10.1002/hbe2.117.
- Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. d. F., and Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, 14(1):e0210236. doi:10.1371/journal.pone.0210236.
- Roesler, E., Manzey, D., and Onnasch, L. (2021). A meta-analysis on the effectiveness of anthropomorphism in human-robot interaction. *Science Robotics*, 6(58):eabj5425. doi:10.1126/scirobotics.abj5425.
- Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408. doi:10.1037/h0042519.
- Rotter, J. B. (1971). Generalized expectancies for interpersonal trust. *American Psychologist*, 26(5):443–452. doi:10.1037/h0031464.
- Rouse, W. B. and Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, 100(3):349–363.
- Royce, W. W. (1970). Managing the development of large software systems. In *Proceedings of IEEE WESCON*, pages 328–338, Los Angeles, CA. doi:10.7551/mitpress/12274.003.0035.
- Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., Mutlu, B., and McDonnell, R. (2015). A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. *Computer Graphics Forum*, 34(6):299–326. doi:10.1111/cgf.12603.
- Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536. doi:10.1038/323533a0.
- Russell, S. J. and Norvig, P. (2021). *Artificial Intelligence: A Modern Approach*. Pearson, 4 edition.
- Ryan, M. (2020). In ai we trust: ethics, artificial intelligence, and reliability. *Science and engineering ethics*, 26(5):2749–2767. doi:10.1007/s11948-020-00228-y.
- SAE On-Road Automated Vehicle Standards Committee (2014). Taxonomy and definitions for terms related to On-Road motor vehicle automated driving systems. SAE Standard J3016, SAE International. doi:10.4271/j3016_201401.
- Sakai, A. (2010). Early caregiver relationships and dispositional trust.
- Sakamoto, T., Harada, Y., and Shimizu, T. (2024). Facilitating trust calibration in artificial intelligence-driven diagnostic decision support systems for determining physicians' diagnostic accuracy: Quasi-experimental study. *JMIR Formative Research*, 8(1):e58666. doi:10.2196/58666.

- Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 3(3):210–229. doi:10.1147/rd.33.0210.
- Sanchez, T., Vereschak, O., and Deroy, O. (2026). Mental models in human-AI interaction: Systematic review of empirical methodologies and guidelines. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pages 663–682.
- Saragih, F. N., Situmorang, J. T. L., Natsir, M., Barus, J. R. S., and Pangaribuan, V. K. N. (2026). Analysis a critical discourse analysis of the representation of artificial intelligence in online media and its impact on public perception. *Student Scientific Creativity Journal*, 4(3):426–439. doi:10.55606/sscj-amik.v4i3.6206.
- Saranya, A. and Subhashini, R. (2023). A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, 7:100230. doi:https://doi.org/10.1016/j.dajour.2023.100230.
- Sarter, N. B., Woods, D. D., and Billings, C. E. (1997). Automation surprises. In Salvendy, G., editor, *Handbook of Human Factors and Ergonomics*, pages 1926–1943. Wiley, New York, NY, 2 edition.
- Sasse, M. A. (1997). *Eliciting and describing users' models of computer systems*. PhD thesis, University of Birmingham.
- Saygin, A. P., Cicekli, I., and Akman, V. (2000). Turing test: 50 years later. *Minds and Machines*, 10:463–518. doi:10.1023/a:1011288000451.
- Scali, G. and Macredie, R. D. (2019). Shared mental models as a way of managing transparency in complex human-autonomy teaming. In *Proceedings of the 14th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2019)*, pages 153–159. doi:10.5220/0007484801530159.
- Schaefer, K. E. (2016). Measuring trust in human robot interactions: Development of the “trust perception scale-HRI”. In *Robust Intelligence and Trust in Autonomous Systems*, pages 191–218. Springer US. doi:10.1007/978-1-4899-7668-0_10.
- Schank, R. C. (1983). *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, Cambridge.
- Schank, R. C. and Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures*. Lawrence Erlbaum, Hillsdale, NJ. doi:10.2307/1421499.
- Scharowski, N., Perrig, S. A., von Felten, N., Aeschbach, L. F., Opwis, K., Wintersberger, P., and Brühlmann, F. (2025). To trust or distrust ai: A questionnaire validation study. In *proceedings of the 2025 ACM conference on fairness, accountability, and transparency*, pages 361–374. doi:10.1145/3715275.3732025.
- Schelble, B. G., Flathmann, C., McNeese, N. J., Freeman, G., and Mallick, R. (2022). Let's think together! assessing shared mental models, performance, and trust in human-agent teams. *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP):1–29.
- Schemmer, M., Kühl, N., Benz, C., Bartos, A., and Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 410–422. doi:10.1145/3581641.3584066.
- Schepman, A. and Rodway, P. (2023). The general attitudes towards artificial intelligence scale (gaais): Confirmatory validation and associations with personality, corporate distrust, and general trust. *International Journal of Human-Computer Interaction*, 39(13):2724–2741. doi:10.1080/10447318.2022.2085400.
- Schmeer, K. (1999). Stakeholder analysis guidelines. Technical report, Partnerships for Health Reform.
- Schmidt, A. (2020). Interactive human centered artificial intelligence: A definition and research challenges. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–4. doi:10.1145/3399715.3400873.
- Scholz, D. D., Kraus, J., and Miller, L. (2025). Measuring the propensity to trust in automated technology: Examining similarities to dispositional trust in other humans and validation of the PTT-A scale. *International Journal of Human-Computer Interaction*, 41(2):970–993. doi:10.1080/10447318.2024.2307691.

- Schomakers, E.-M., Biermann, H., and Ziefle, M. (2021). Users' preferences for smart home automation — investigating aspects of privacy and trust. *Telematics and Informatics*, 64:101689. doi:10.1016/j.tele.2021.101689.
- Schossau, J. and Hintze, A. (2023). Towards a theory of mind for artificial intelligence agents. In *Artificial Life Conference Proceedings 35*, volume 2023, page 21. MIT Press.
- Schrills, T., Franke, T., Hoesterey, S., and Roesler, E. (2026). Questioning trust in ai research: exploring the influence of trust assessment on dependence in ai-assisted decision-making. *Behaviour & Information Technology*, 45(8):1644–1660. doi:10.1080/0144929x.2025.2553153.
- Schumacher, R. M. and Czerwinski, M. P. (1992). Mental models and the acquisition of expert knowledge. In *The Psychology of Expertise: Cognitive Research and Empirical AI*, pages 61–79. Springer.
- Schvaneveldt, R. W., Dearholt, D. W., and Durso, F. T. (1988). Graph theoretic foundations of pathfinder networks. *Computers & Mathematics with Applications*, 15(4):337–345. doi:10.1016/0898-1221(88)90221-0.
- Schwaber, K. and Sutherland, J. (2020). The scrum guide. <https://scrumguides.org>.
- See, J. E., Craft, R. L., and Morris, J. D. (2019). Human readiness levels in the systems engineering process at Sandia national laboratories. Technical Report SAND2019-3123, Sandia National Laboratories, Albuquerque, NM. doi:10.2172/1761924.
- Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural Computation*, 35(3):309–342. doi:10.1162/neco_a_01563.
- Seong, Y. and Bisantz, A. M. (2000). Assessment of operator trust in and utilization of automated decision aids under different framing conditions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 44(1):5–8. doi:10.1177/154193120004400102.
- Setiono, R. and Leow, W. K. (2000). FERNN: An algorithm for fast extraction of rules from neural networks. *Applied Intelligence*, 12(1–2):15–25. doi:10.1023/A:1008307919726.
- Shang, R., Hsieh, G., and Shah, C. (2024). Trusting your ai agent emotionally and cognitively: Development and validation of a semantic differential scale for ai trust. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 1343–1356. doi:10.1609/aies.v7i1.31728.
- Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, 30(1):50–64. doi:10.1002/j.1538-7305.1951.tb01366.x.
- Sheridan, T. B. (1992). *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA. doi:10.1109/mtas.1993.232279.
- Sheridan, T. B. and Parasuraman, R. (2005). Human-automation interaction. *Reviews of Human Factors and Ergonomics*, 1(1):89–129. doi:10.1518/155723405783703082.
- Sheridan, T. B. and Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Technical report, MIT Man-Machine Systems Laboratory, Cambridge, MA. doi:10.21236/ada057655.
- Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 12(3):109–124. doi:10.17705/1thci.00131.
- Shneiderman, B. (2021). Human-centered AI. *Issues in Science and Technology*, 37(2):56–61. doi:10.1093/oso/9780192845290.001.0001.
- Shneiderman, B. (2022). *Human-Centered AI*. Oxford University Press, Oxford. doi:10.1093/oso/9780192845290.001.0001.
- Shneiderman, B., Plaisant, C., Cohen, M., Jacobs, S., Elmqvist, N., and Diakopoulos, N. (2016). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. Pearson. doi:10.1145/2977645.
- Shortliffe, E. H. (1976). *MYCIN: Computer-Based Medical Consultations*. Elsevier, New York.
- Simpson, T. W. (2012). What is trust? *Pacific Philosophical Quarterly*, 93(4):550–569. doi:10.1111/j.1468-0114.2012.01438.x.
- Singh, C., Inala, J. P., Galle, M., Caruana, R., and Gao, J. (2024). Rethinking interpretability in the era of large language models. *arXiv preprint*. arXiv:2402.01761.
- Singh, I. L., Molloy, R., and Parasuraman, R. (1993). Automation-induced “complacency”: De-

- velopment of the complacency-potential rating scale. *The International Journal of Aviation Psychology*, 3(2):111–122. doi:10.1207/s15327108ijap0302_2.
- Skitka, L. J., Mosier, K., and Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, 52(4):701–717. doi:10.1006/ijhc.1999.0349.
- Smith, A. et al. (2017). Default trust in human–machine interaction.
- Smuha, N. A. (2019). The EU approach to ethics guidelines for trustworthy artificial intelligence. *Computer Law Review International*, 20(4):97–106. doi:10.9785/crl-2019-200402.
- Spataro, J. (2023). Introducing microsoft 365 copilot—your copilot for work. The Official Microsoft Blog. <https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/>.
- Spatola, N., Marchesi, S., and Wykowska, A. (2022). Different models of anthropomorphism across cultures and ontological limits in current frameworks: The integrative framework of anthropomorphism. *Frontiers in Robotics and AI*, 9:863319. doi:10.3389/frobt.2022.863319.
- St-Cyr, O. and Burns, C. M. (2002). Mental models and ecological interface design: An experimental investigation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 46, pages 287–291. SAGE.
- Staggers, N. and Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies*, 38(4):587–605.
- Stanton, B. and Jensen, T. (2021). Trust and artificial intelligence. NIST IR 8332, National Institute of Standards and Technology. doi:10.6028/NIST.IR.8332-draft.
- Stegner, B. (2023). Your phone is secretly recording you: How to stop Google from listening. MSN. <https://www.msn.com/en-us/news/technology/your-phone-is-secretly-recording-you-how-to-stop-google-from-listening/ar-AAxayUM>.
- Stein, N. and Brooks, K. (2017). A fully automated conversational artificial intelligence for weight loss: longitudinal observational study among overweight and obese adults. *JMIR diabetes*, 2(2):e28. doi:10.2196/diabetes.8590.
- Sternberg, R. J. (2012). Intelligence. *Dialogues in Clinical Neuroscience*, 14(1):19–27. doi:10.31887/DCNS.2012.14.1/rsternberg.
- Stevens, A. F. and Stetson, P. (2023). Theory of trust and acceptance of artificial intelligence technology (traait): An instrument to assess clinician trust and acceptance of artificial intelligence. *Journal of biomedical informatics*, 148:104550. doi:10.2139/ssrn.4545976.
- Stockman, C. and Nottingham, E. (2024). Dark patterns of cuteness: Popular learning app design as a risk to children's autonomy. In *Children, Young People and Online Harms*, pages 113–137. Springer. doi:10.1007/978-3-031-46053-1_5.
- Strickland, E. (2023). OpenAI's moonshot: Solving the AI alignment problem. *IEEE Spectrum*. <https://spectrum.ieee.org/the-alignment-problem-openai>.
- Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:1–19. doi:10.1155/2009/421425.
- Subramanian, H. V., Canfield, C., and Shank, D. B. (2024). Designing explainable AI to improve human-AI team performance: A medical stakeholder-driven scoping review. *Artificial Intelligence in Medicine*, page 102780. doi:10.1016/j.artmed.2024.102780.
- Suchman, L. A. (1983). Office procedure as practical action: Models of work and system design. *ACM Transactions on Information Systems*, 1(4):320–328. doi:10.1145/357442.357445.
- Surianarayanan, C., Lawrence, J. J., Chelliah, P. R., Prakash, E., and Hewage, C. (2023). A survey on optimization techniques for edge artificial intelligence (AI). *Sensors*, 23(3):1279. doi:10.3390/s23031279.
- Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, 2 edition. doi:10.1109/tnn.1998.712192.
- Sweller, J. and Chandler, P. (1991). Evidence for cognitive load theory. *Cognition and Instruction*, 8(4):351–362. doi:10.1207/s1532690xci0804_5.
- Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1, National Institute of Standards and Technology. doi:10.6028/NIST.AI.100-1.

- Taddeo, M. (2009). Defining trust and e-trust. *International Journal of Technology and Human Interaction*, 5(2):23–35. doi:10.4018/jthi.2009040102.
- Tate, D., Grier, R., Martin, C. A., Moses, F. L., and Sparrow, D. (2016). A framework for evidence-based licensure of adaptive autonomous systems. IDA Paper P-5325, Institute for Defense Analyses. doi:10.13140/RG.2.2.11845.86247.
- Teo, S. (2024). How i won singapore's GPT-4 prompt engineering competition. Towards Data Science. <https://towardsdatascience.com/how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41>.
- Thelot, R. (2023). Searching for sentience. *AI & Society*, 39(6):3043–3045. doi:10.1007/s00146-023-01740-y.
- Tian, Y., Su, D., Lauria, S., and Liu, X. (2022). Recent advances on loss functions in deep learning for computer vision. *Neurocomputing*, 497:129–158. doi:10.1016/j.neucom.2022.04.127.
- Toulmin, S. (1958). *The Uses of Argument*. Cambridge University Press, Cambridge. doi:10.1017/cbo9780511840005.
- Traubinger, V., Heil, S., Grigera, J., Garrido, A., and Gaedke, M. (2024). In search of dark patterns in chatbots. In *Chatbot Research and Design*, pages 117–132. Springer. doi:10.1007/978-3-031-54975-5_7.
- Tsai, C.-H. and Brusilovsky, P. (2019). Designing explanation interfaces for transparency and beyond. In *IUI Workshops*, volume 19.
- Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236):433–460. doi:10.1093/mind/LIX.236.433.
- Üstün, A. G., Yavuz, M., Kayalı, B., Uçar, H., Erdoğan, E., and Bozkurt, A. (2026). Trust in ai: scale development and validation for online distance learners. *BMC psychology*. doi:10.1186/s40359-026-04679-z.
- van der Veer, G. C. and Melguizo, M. d. C. P. (2003). Mental models. In Jacko, J. A. and Sears, A., editors, *The Human-Computer Interaction Handbook*, pages 52–80. Lawrence Erlbaum Associates, Mahwah, NJ.
- Vargas, R., Mosavi, A., and Ruiz, R. (2017). Deep learning: A review. *Advances in Intelligent Systems and Computing*. doi:10.20944/preprints201810.0218.v1.
- Veinott, E., Klein, G. A., and Wiggins, S. (2010). Evaluating the effectiveness of the PreMortem technique on plan confidence. In *Proceedings of the 7th International ISCRAM Conference*, Seattle, WA.
- Venkatesh, V. and Davis, F. D. (2000). A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management Science*, 46(2):186–204. doi:10.1287/mnsc.46.2.186.11926.
- Vicente, K. J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. CRC Press.
- Vicente, L. and Matute, H. (2023). Humans inherit artificial intelligence biases. *Scientific Reports*, 13:15737. doi:10.1038/s41598-023-42384-8.
- Villareale, J., Hartevelt, C., and Zhu, J. (2022). “I want to see how smart this AI really is”: Player mental model development of an adversarial AI player. *Proceedings of the ACM on Human-Computer Interaction*, 6(CHI PLAY):1–26.
- Villareale, J. and Zhu, J. (2021). Understanding mental models of AI through player-AI interaction. *arXiv preprint. arXiv:2103.16168*.
- Vilone, G. and Longo, L. (2021). Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3), 615–661.
- Voiklis, J. et al. (2016). Cited in de graaf and malle (2017).
- Vultureanu-Albiși, A. and Bădică, C. (2021). Recommender systems: An explainable AI perspective. In *2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, pages 1–6.
- Vuong, Q.-H., La, V.-P., Nguyen, M.-H., Jin, R., La, M.-K., and Le, T.-T. (2023). How AI's self-prolongation influences people's perceptions of its autonomous mind: The case of U.S. residents. *Behavioral Sciences*, 13(6):470. doi:10.3390/bs13060470.

- Waern, Y. (1987). On the dynamics of mental models. In *Selected papers of the 6th Interdisciplinary Workshop on Informatics and Psychology: Mental Models and Human-Computer Interaction 1*, pages 73–93.
- Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY*, 36(2):585–595. doi:10.1007/s00146-020-01066-z.
- Walsh, S., Wang, Q., and Ying, L. (2026). Theory of mind in human-AI interaction and AI. In *Handbook of Human-Centered Artificial Intelligence*, pages 371–413. Springer.
- Walther, D. and Koch, C. (2006). Modeling attention to salient proto-objects. *Neural Networks*, 19(9):1395–1407. doi:10.1016/j.neunet.2006.10.001.
- Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., and Wang, Q. (2020). From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–6. doi:10.1145/3334480.3381069.
- Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence*, 10(2):1–37.
- Wang, P., Yin, K., Tian, M., Zheng, Y., Wu, H., Zhou, C., Zhang, M., Ma, J., and Yuan, X. (2026). A validation of the human-generative artificial intelligence trust scale. *International Journal of Human-Computer Interaction*, 42(7):4808–4821.
- Wang, Q., Walsh, S., Si, M., Kephart, J., Weisz, J. D., and Goel, A. K. (2024). Theory of mind in human-AI interaction. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–6.
- Wanner, J., Herm, L.-V., Heinrich, K., and Janiesch, C. (2022). The effect of transparency and trust on intelligent system acceptance: Evidence from a user-based study. *Electronic Markets*, 32(4):2079–2102. doi:10.1007/s12525-022-00593-5.
- Watson, I. and Marir, F. (1994). Case-based reasoning: A review. *The Knowledge Engineering Review*, 9(4):327–354. doi:10.1017/S0269888900007098.
- Watts, . S. (2023). What is observability? an introduction. *splunk*.
- Waytz, A., Cacioppo, J., and Epley, N. (2010). Who sees human?: The stability and importance of individual differences in anthropomorphism. *Perspectives on Psychological Science*, 5(3):219–232. doi:10.1177/1745691610369336.
- Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, 52:113–117. doi:10.1016/j.jesp.2014.01.005.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint. arXiv:2201.11903*.
- Weizenbaum, J. (1966). Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- Werbos, P. J. (1990). Backpropagation through time: what it does and how to do it. In *Proceedings of the IEEE*, 78(10), 1550–1560.
- Wickens, C. D., Clegg, B. A., Vieane, A. Z., and Sebok, A. L. (2015). Complacency and automation bias in the use of imperfect automation. *Human Factors*, 57(5):728–739. doi:10.1177/0018720815578777.
- Wiener, E. L. (1989). Human factors of advanced technology (“glass cockpit”) transport aircraft. NASA Contractor Report 177528, NASA Ames Research Center, Moffett Field, CA.
- Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. The MIT Press, Cambridge, MA.
- Wienrich, C. and Latoschik, M. E. (2021). eXtended artificial intelligence: New prospects of human-AI interaction research. *Frontiers in Virtual Reality*, 2. doi:10.3389/frvir.2021.686783.
- Wikipedia contributors (2024). PageRank. Wikipedia. <https://en.wikipedia.org/w/index.php?title=PageRank&oldid=1193600064>.
- Williams, J., Fiore, S. M., and Jentsch, F. (2022). Supporting artificial social intelligence with theory of mind. *Frontiers in Artificial Intelligence*, 5:750763.
- Williams, M. D., Hollan, J. D., and Stevens, A. L. (1983). Human reasoning about a simple physical

- system. In Gentner, D. and Stevens, A. L., editors, *Mental Models*, pages 131–153. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Wilson, J. R. and Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors*, 31(6):617–634.
- Wischnewski, M., Doebler, P., and Krämer, N. (2026). Development and validation of the trust in ai scale (tais). *Journal of Psychology and AI*, 2(1):2715475.
- Wixom, B., Someh, I., and Gregory, R. (2020). AI alignment: A new management paradigm. <https://espace.library.uq.edu.au/view/UQ:3698609>.
- Wojton, H. M., Porter, D., Lane, S. T., Bieber, C., and Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of Social Psychology*, 160(6):735–750. doi:10.1080/00224545.2020.1749020.
- Wood, J. A. (2006). Nlp revisited: Nonverbal communications and signals of trustworthiness. *Journal of Personal Selling & Sales Management*, 26(2):197–204.
- Woods, D. D. (1996). Decomposing automation: Apparent simplicity, real complexity. In Parasuraman, R. and Mouloua, M., editors, *Automation and Human Performance: Theory and Applications*, pages 3–17. Lawrence Erlbaum, Mahwah, NJ.
- Woods, D. D. and Roth, E. M. (1988). Cognitive systems engineering. In *Handbook of human-computer interaction*, pages 3–43. Elsevier.
- Wu, Y.-c. and Feng, J.-w. (2018). Development and application of artificial neural network. *Wireless Personal Communications*, 102(2):1645–1656. doi:10.1007/s11277-017-5224-x.
- Xiang, Y., Zhao, L., Liu, Z., Wu, X., Chen, J., Long, E., Lin, D., Zhu, Y., Chen, C., Lin, Z., and Lin, H. (2020). Implementation of artificial intelligence in medicine: Status analysis and development suggestions. *Artificial Intelligence in Medicine*, 102:101780. doi:10.1016/j.artmed.2019.101780.
- Xie, L., Guo, T., Yang, Y., Yu, L., Chen, Y., Peng, X., Zheng, Z., and Tu, H. (2025a). Trust in artificial intelligence-based follow-up in hospital information systems: development and validation of a new scale. *BMC psychology*, 14(1):95.
- Xie, Y., Zhou, R., Chan, A. H. S., and Xiong, D. (2025b). Do you trust me? measuring people's perception of being trusted by ai in a human-agent team. *International Journal of Human-Computer Interaction*, 41(20):13041–13058.
- Xu, J., Li, H., and Zhou, S. (2015). An overview of deep generative models. *IETE Technical Review*, 32(2), 131-139.
- Xu, W. (2019). Toward human-centered AI: A perspective from HCI. *Interactions*, 26(4):42–46. doi:10.1145/3328485.
- Xu, W., Dainoff, M. J., Ge, L., and Gao, Z. (2021). From human-computer interaction to human-AI interaction: New challenges and opportunities for enabling human-centered AI. [arXiv: 2105.05424](https://arxiv.org/abs/2105.05424).
- Yang, S., Huang, Y., Huang, X., Zhang, J., Meng, Z., and Yang, J. (2025). Impact of anthropomorphism in AI assistants' verbal feedback on task performance and emotional experience. *Ergonomics*, 69(6):980–993. doi:10.1080/00140139.2025.2497072.
- Yates, J. F., Veinott, E. S., and Patalano, A. L. (2003). Hard decisions, bad decisions: On decision quality and decision aiding. In *Emerging Perspectives on Judgment and Decision Research*, pages 13–63. Cambridge University Press.
- Yin, X., Doost, E. Z., Zhou, S., Yadav, G. A., and Gorman, J. C. (2025). When researchers say mental model/theory of mind of AI, what are they really talking about? [arXiv preprint arXiv:2510.02660](https://arxiv.org/abs/2510.02660).
- Young, R. M. (1981). The machine inside the machine: Users' models of pocket calculators. *International Journal of Man-Machine Studies*, 15(1):51–85. doi:10.1016/S0020-7373(81)80023-5.
- Young, R. M. (1983). Surrogates and mappings: Two kinds of conceptual models for interactive devices. In Gentner, D. and Stevens, A. L., editors, *Mental Models*, pages 35–52. Lawrence Erlbaum Associates, Hillsdale, NJ.
- Zanzotto, F. M. (2019). Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252.

- Zegwaard, K. E., Pretti, T. J., Rowe, A. D., and Ferns, S. J. (2023). *Defining work-integrated learning*. The Routledge International Handbook of Work-Integrated Learning (3rd ed. Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.
- Zewe, A. (2023). Explained: Generative AI. MIT News. <https://news.mit.edu/2023/explained-generative-ai-1109>.
- Zhang, G., Yan, C., Ji, X., Zhang, T., Zhang, T., and Xu, W. (2017). DolphinAttack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pages 103–117. doi:10.1145/3133956.3134052.
- Zhang, R., McNeese, N. J., Freeman, G., and Musick, G. (2021). “an ideal human”: Expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25.
- Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Xu, C., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., and Shi, S. (2023). Siren’s song in the AI ocean: A survey on hallucination in large language models. [arXiv:2309.01219](https://arxiv.org/abs/2309.01219).
- Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 295–305. doi:10.1145/3351095.3372852.
- Zhao, M., Lu, H., Yang, S., and Guo, F. (2020). The experience-memory Q-learning algorithm for robot path planning in unknown environment. *IEEE Access*, 8:47824–47844. doi:10.1109/ACCESS.2020.2978077.
- Zheng, J., Veinott, E., Bos, N., Olson, J. S., and Olson, G. M. (2002). Trust without touch: Jumpstarting long-distance trust with initial social activities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 141–146. doi:10.1145/503376.503402.
- Zhou, S., Yin, X., Scalia, M. J., Zhang, R., Gorman, J. C., and McNeese, N. J. (2023). Development of a real-time trust/distrust metric using interactive hybrid cognitive task analysis. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 67, pages 2128–2136. doi:10.1177/21695067231192549.
- Zirar, A., Ali, S. I., and Islam, N. (2023). Worker and workplace artificial intelligence (AI) coexistence: Emerging themes and research agenda. *Technovation*, 124:102747.
- Zondag, A. G., Rozestraten, R., Grimmelikhuijsen, S. G., Jongsma, K. R., van Solinge, W. W., Bots, M. L., Vernooij, R. W., and Haitjema, S. (2024). The effect of artificial intelligence on patient-physician trust: Cross-sectional vignette study. *Journal of medical Internet research*, 26:e50853.
- Zsombok, C. E., Kaempf, G. L., Crandall, B., and Kyne, M. (1997). A comprehensive program to deliver on-the-job training (OJT). DTIC Report ADA327576, U.S. Army Research Institute.
- Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. Basic Books, New York.

Index

- A* search, 19, **181**
- Absolute Trust, 120, **181**
- Absolute Trusting, 120, **181**
- Accountability, 68, 117, 126, 174, **181**
- ACT-R, 114, **181**
- Adaptive cruise control, 35, 48, **181**
- Agent, **181**
- AGI (Artificial General Intelligence), **181**
- Agile development, 56, **182**
- AI alignment (values), **182**
- AI Bias, **182**
- AI-assisted decision making, 38, **182**
- Algorithm, **182**
- Algorithm aversion, 106, **182**
- Algorithmic decision-making, **182**
- Alignment, 45
- Alignment problem (AI safety), **182**
- Altruistic punishment, 104, **182**
- Amershi HAI design guidelines, **182**
- ANI, **182**
- Anthropomorphic attributions, 94
- Anthropomorphic design, 86, 103, **182**
- Anthropomorphic expectations, 94
- Anthropomorphic expectations/attributions, **182**
- Anthropomorphism, 2, 65, 75, 105, 122, 166, **183**
- Anti-trusting, 120, **183**
- Applied Cognitive Task Analysis (ACTA), 138, **183**
- Appropriate trust, 119, **183**
- APT (Argument-based probabilistic trust), 113, **183**
- Artificial Intelligence, 1, 41, 42, 77, 83, 121, 126, **183**
- Artificial Neural Network, 18, 19, **183**
- ASI (Artificial Super Intelligence), **183**
- Auditability, **183**
- Authority Hypothesis, 70, 141, **183**
- Authority-based trusting, 120, **183**
- Automation, 1, 3, 38, 65, 70, 75, 97, **183**
- Automation Bias, 7, 70, 120, **183**
- Back Propagation, **183**
- Backtracking (search), **183**
- Backward chaining, 19, 23, **183**
- Bag-of-words model, 12, **183**

- Bayes Network, 26, **183**
- Bayesian inference, 5, 25, **183**
- Behavior Tree, **184**
- Beneficence, **184**
- Bias in computer systems, **184**
- Black box, 173, **184**
- Blackboard System, **184**
- Blind trust, 120, **184**
- Boundary conditions (automation), 49, 68, **184**
- Boundary conditions (theory), **184**
- CAPTCHA, 79, **184**
- CASA, 89, **184**
- Case-based reasoning (CBR), 18, **184**
- Chain-of-Thought Prompting, **184**
- Charrette, 62, **184**
- Chess engines, **184**
- Classifier or Classification System, **184**
- Clumsy automation, 50, 51, **185**
- Clustering, 13, **185**
- CO-STAR Method, **185**
- Coactive Design, 7, 44, **185**
- Coffee test, 85, **185**
- Cognition-based trust, 107, **185**
- Cognitive anthropomorphism, 86, **185**
- Cognitive architecture, 92, **185**
- Cognitive Decathlon, 80, **185**
- Cognitive digital twin, 92, 93, **185**
- Cognitive Load Theory, 73, **185**
- Cognitive Systems Engineering (CSE), 6, 41, 42, 122, **185**
- Cognitive Task Analysis (CTA), 135, **185**
- Cognitive Tutorial, **185**
- Cognitive Work Analysis (CWA), 72, **185**
- Cognitive Task Analysis, **185**
- Collaboration lifecycle, 58, **185**
- Collaborative Filtering, 12, **186**
- Competence (fidelity), **186**
- Competence envelope, 8, 9, 49, 83, 161, **186**
- Componential trust, 118, 122, **186**
- Computers in-the-loop, **186**
- Concept map, 172, **186**
- Constraint Satisfaction Problem (CSP), **186**
- Contestability, 175, **186**
- Context window, 125, **186**
- Contextual Design, 72, **186**
- Contextual trust, 125, **186**
- Contingent Trusting, 120, **186**
- Contractual Trust, 126, **186**
- Controller, **186**
- conversational agent, 166
- Cost Function, 5, 19, **186**
- Counter-trusting, 120, **186**
- Critical Decision Method, 44, 136, 170, **186**
- Critical periods (HCAI lifecycle), 54, **186**
- Cybernetics, 6, 37, 38, 156, **186**
- Dark patterns, 89, 92, 105, **186**
- Data Card, **187**
- Deep learning, 21, 31, 85, **187**
- Default trusting, 120, **187**
- Defensive monitoring, 140, **187**
- Dependability, 108, **187**
- Design thinking, 56, **187**
- Desirable difficulty, 73, **187**
- Desirements, 61, 62, **187**
- Digital Twin, 10, 93, **187**
- Digressive Trusting, 120, **187**
- Dijkstra's Algorithm, 19, **187**
- Dimensionality reduction, 13, **187**
- Directability, 44, **187**
- Distributed Cognition, 44, **187**
- Distributional verisimilitude, 80, **187**

- Distrust, 29, 49, 67, 87, 101, 105, **187**
- Disuse (automation), 119, **187**
- Domination (fidelity), **188**
- Ecological Interface Design, 157, **188**
- Effective anthropomorphism, 87, **188**
- ELIZA, 80, **188**
- Emotional Connection, 94, **188**
- Emulation (human-like AI), **188**
- Encapsulated trust, 49, **188**
- Ethopoeia, 90, **188**
- Evidence-based licensure, **188**
- Evolutionary algorithm, 28, **188**
- Experiential User Guide, 72, **188**
- Expert System, 5, 18, **188**
- Explainability, 2, 62, 87, 121, 126, 173, **188**
- Explainable AI, 40, 55, 121, 161, **188**
- Faith-based trusting, 120, **188**
- Feed-Forward Neural Network, **188**
- Feigenbaum Test, 79, **188**
- Few-shot prompting, **188**
- Filter bubble, 106, **189**
- Fine-tuning, 5, **189**
- formative assessment, 68, **189**
- Forward chaining, **189**
- Functional mental models, 160, **189**
- Functional understanding, 48, **189**
- Functionality (anthropomorphism space), **189**
- GAMs, **189**
- General Problem Solver (GPS), **189**
- Generate and test, 14, **189**
- Generative AI, 2, 5, 78, 88, 98, 121, 152, **189**
- generative model, 15, 17, **189**
- Genetic algorithm, 5, 19, **189**
- Glass Box Model, 173, **189**
- Global Explanation, 160
- Good regulator theorem, 48, 156, **189**
- Gradient descent, 19, 22, **189**
- Hallucination, 107, 125, 161, **189**
- heuristic, **190**
- Human Readiness Level (HRL), 57, **190**
- Human–Computer Trust (HCT), 149, **190**
- Human-agent teaming, 43
- Human-AI collaboration, 38, 40, 72, **190**
- Human-AI Interaction (HAI), 39, **190**
- Human-AI teaming, 38, 41, 162, **190**
- Human-automation interaction, 38, **190**
- Human-Autonomy Teaming, 43, **190**
- Human-autonomy teaming, 43
- Human-centered AI, 6, 37, 38, 53, 122, 124, **190**
- Human-centered machine learning, 45
- Human-in-the-loop, 38, **190**
- Human-in-the-loop ML, 45, **190**
- Human-machine teaming, 43
- human-machine teaming, 38, 98, **190**
- Human-Robot Interaction (HRI), 44, **191**
- Humanization, 91, **191**
- Hybrid Intelligence, 40, **191**
- ICAP framework, **191**
- IEEE P7007 Standard on Transparency, **191**

- Illusion of Understanding, 169
- Imitation Game, 77, **191**
- In-context learning in LLMS, **191**
- Information retrieval, 12, 31, **191**
- information transparency, **191**
- Intelligence, **191**
- Interactive Human-Centered AI (IH-CAI), 42, **191**
- Interactivity (anthropomorphism space), **191**
- Interpretability, **191**
- Ironies of automation, 7, **191**

- Jagged Technological Frontier, 83, **191**
- Joint activity, **191**
- Joint cognitive systems, 38, **192**
- Justifiability, **192**
- Justified mistrusting, 47, **192**
- Justified trusting, 47, 98, **192**

- Knowledge elicitation bottleneck, 21, **192**
- Knowledge transfer and exchange (KTE), **192**

- leakage, **192**
- Lens model, 107, **192**
- Levels of automation, 4, **192**
- LIME, **192**
- LLMs, 14, 16, 85, 125, 161, **192**
- Logistic regression, 9, **192**
- Loss Function, 5, **192**
- LSA, **192**

- Machine Learning, 1, 5, 61, 68, 93, 115, 116, **193**
- Macro cognition, 44, **193**
- Markov model, 15, 31, **193**
- material transparency, **193**
- Mayer-Hancock trust framework, 111, **193**
- Mechamorphism, 87, **193**
- Mechanistic mental models, 158, **193**
- Mental Model, 2, 47, 67, 69, 87, 112, 157, **193**
- Mental Model Matrix, 70, 71, 163, **193**
- Mental Model of a Dynamic System, **193**
- Mentalism, 91, **193**
- Metaphor Model, **193**
- Micro cognition, **193**
- Microworld, 132, **193**
- Minimum necessary rigor, 67, 173, **193**
- Mistrusting, 120, **193**
- Misuse (automation), 119, **193**
- Mixed Initiative Interaction/Planning, 45, **193**
- Model Card, **193**
- Monte Carlo Sampling, **194**
- Moravec's paradox, 77, 83, **194**
- Mutual Theory of Mind (MTOM), 168, **194**

- N-gram, 15, **194**
- natural language processing, 31, 33, 139, **194**
- Naturalistic Decision Making, 44, **194**

- Observability, 44, 173, **194**
- One-Shot Prompting, **194**
- Optimization, 1, 2, 47, **194**
- Oracle-based testing, **194**
- Organizational trust, 103, 104, **194**
- Overfitting, 6, **194**
- Overtrust, 119, 140, **194**

- PageRank, 12, 31, **194**
- Participatory design, 72, **194**
- Pathfinder networks, 172, **194**
- PCA, **194**
- Perceived trustworthiness, 103, **194**

- PID controller, 11, 35, **194**
- Pre-training, **195**
- Predictability, 44, 112, 150, **195**
- Premortem, 63, **195**
- Procedural Justice, **195**
- Production-Rule System, **195**
- Progressive Trust, 120, **195**
- Progressive Trusting, 120, **195**
- Prompt, **195**
- Prompt Engineering, **195**
- Propensity to trust, 106, 132, **195**

- Q-learning, 19, 26, 156, **195**

- RAG in LLMs, 161, **195**
- rational agent, 25, **195**
- Recognition-Primed Decision (RPD), **196**
- Recommender System, 11, 162, **196**
- Recurrent Neural Network, 16, 32, **196**
- reflective transparency, **196**
- Reinforcement learning, 5, 88, 166, **196**
- Reinforcement Learning from Human Feedback (RLHF), 88, **196**
- Reinforcement Learning with Human Feedback, 45
- Reliability, 42, 78, 101, 103, **196**
- Reliance, 92, 100, 101, 141, **196**
- Resemblance (fidelity), **196**
- Resilience Engineering, 43, **196**
- Reversal Curse, **196**
- Reverse Turing test, 79, **196**
- Rule-based system, 18, 29, **196**

- SAE J3016 (Levels of Driving Automation), **196**
- Safety, **196**
- Scalable Human Oversight, **196**
- Scrum, 56, **197**

- Self-Exfiltration, **197**
- Sentiment analysis, 31, 32, 139, **197**
- SHAP, 85, 101, 116, **197**
- Shapley Value, **197**
- Shared mental models, 162, **197**
- Shared responsibility, 48, **197**
- Shifting bottlenecks, 7, 50, **197**
- Simulacra, 75, 76, **197**
- Situated cognition, 71, 72, **197**
- Situation awareness, 38, 163, **197**
- Skeptical trusting, 120, **197**
- SLAM, 30, **197**
- Social Actor, 86, 89, **197**
- Social Robotics, 7, 44, 86, 90, 103, **197**
- Socio-technical system, 107, **197**
- Sociotechnical systems, 43, **197**
- Soft-max confidence, **197**
- Spiritualism, 91, **197**
- SQA (Social Q&A), **198**
- Stable trusting, 120, **198**
- Stakeholder alignment (adaptive management), **198**
- Stakeholder analysis, 61, 65, **198**
- Stakeholder Playbook, 62, **198**
- Stochastic parrot, **198**
- Strong AI, **198**
- Structural mental models, 158, 159, **198**
- Supervised learning, 5, 20, **198**
- Supervisory control, **198**
- Surrogate Model, 161, **198**
- SVD, **198**
- Swift Trust, 120, 136, **198**
- Symbolic AI, 23, **198**

- Technological Readiness Level (TRL), 57, **198**
- Technology Acceptance Model (TAM), 54, **199**
- Tentative trusting, 120, **199**
- Theory of Mind, 167, **199**

Thera-Turing test, 79, **199**
Think-aloud protocol, 138, 170, **199**
Three R's of trust, **199**
Toulmin model, 113, **199**
Training vs. deployment, **199**
Transparency, 173, **199**
transparency, transformational, **199**
transparency-in-use, **199**
Tree of knowledge, **199**
Trust, **199**
Trust calibration, 110, 142, **199**
Trust in Automation (TiA), 149, **199**
Trust in XAI Context (TXAI), 151, **199**
Trustworthiness, 59, 68, 91, 97, 98, **200**
Trustworthy AI, 98, 116, **200**
Turing Test, 75, **200**

Universal Intelligence Measure, 85, **200**
Unjustified mistrusting, **200**
Unjustified trusting, 98, 120, **200**
Unsupervised learning, 5, 20, **200**
Unwarranted trust, **200**
User expectations of AI, 66, **200**

Verisimilitude (fidelity), **200**
Vignette (research), 133, **200**

Warranted trust, 119, 163, **200**
Weaponized anthropomorphism, 92, **200**
Wizard of Oz (WOZ), 57, 134, **200**
Work domain analysis, **200**
Work-Integrated Learning (WIL), 73, **200**

Zero-Shot Prompting, **200**