ARTIFICIAL INTELLIGENCE GENERAL EDUCATION

Cognitive Autonomy in an Intelligent Society

Center for Artificial Intelligence General Education (AIGE)
Department of Computer Science and Technology
Tsinghua University

Copyright and Open License

Title:
Artificial Intelligence General Education: Cognitive Autonomy in an Intelligent Society

Authors:
Dong Wang and Shaoping Ma

Published by:
Center for Artificial Intelligence General Education (AIGE),
Department of Computer Science and Technology, Tsinghua University

© 2026 Dong Wang and Shaoping Ma. Some rights reserved.

Except where otherwise indicated, the text of this book and the figures originally created by the authors are licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

Under the terms of this license, users may copy and redistribute the unmodified work in any medium or format, provided that appropriate credit is given and the material is used only for non-commercial purposes.

Translations, abridgements, adaptations, rearrangements, remixes and other modified versions may not be distributed without separate written permission from the copyright holders.

Suggested attribution:

Dong Wang and Shaoping Ma, *Artificial Intelligence General Education: Cognitive Autonomy in an Intelligent Society*, Center for Artificial Intelligence General Education (AIGE), Tsinghua University, 2026.

Images, tables, quotations and other third-party materials for which a source, copyright holder or separate license is indicated are not covered by the license above. Their use is subject to the terms specified by the respective rights holders and may require separate permission.

The names and marks of Tsinghua University and the Center for Artificial Intelligence General Education (AIGE) are not included within the scope of this Creative Commons license.

For commercial use, translation, adaptation, publication or other permissions, please contact:

aigraph@cslt.org

Version: First Edition, 2026

Permanent URL or DOI:10.5281/zenodo.21959670

Redbook on Artificial Intelligence General Education

Translation of Chinese edition (v2.3.1)

Advisor Bo ZHANG

Editors-in-Chief Dong WANG Shaoping MA

AIGE Research Center, Department of Computer Science and Technology, Tsinghua University

2026

Preface

Part I

Artificial intelligence is profoundly changing how knowledge is produced, information is disseminated, judgments are formed, and social action is undertaken. AI is no longer merely a technology or a tool; it is becoming part of the underlying infrastructure of society. The information environment faced by individuals is likewise shifting from direct communication among people to a new environment extensively mediated by AI. In the future, AI will become more deeply embedded throughout the processes of knowledge discovery, information dissemination, wealth creation, and service provision. It will be closely coupled with foundational institutions, and together they will propel humanity toward a new social formation—an "intelligent society."

The emergence of the intelligent society places new demands on education. Once AI has become social infrastructure, acquiring knowledge of AI and learning to use it proficiently become essential tasks for everyone.

Yet this is far from sufficient. The most consequential change brought by the intelligent society is not the proliferation of highly intelligent devices, but the fundamental transformation of the human cognitive environment. The information one encounters may have passed through multiple layers of AI filtering or may have been generated by AI outright. A person's choices may contain subtle guidance from a system. A person's words and actions may, through AI systems, affect friends and family. More seriously, while using AI tools to improve efficiency, people may also be outsourcing their own capacity to think; the comforting reassurance they receive from an AI system may deepen their fixation on mistaken beliefs. Future AI may anticipate people's thoughts, sway their emotions, and shape their minds. This is not the predicament of a few individuals. It is a cognitive challenge confronting humanity as a whole.

Education must therefore respond to a crucial question: how can every citizen preserve the capacity for cognitive autonomy in the intelligent society of the future?

Cognitive autonomy is more than simple mental clarity. At its core lies the ability of each person to remain the author of their own thought and action, without being swept along by a complex cognitive environment—whether by intelligent systems or by other people.

The key to cognitive autonomy is self-direction: actively accepting, learning about, participating in, and making use of the tools, devices, and modes of operation of the intelligent age; embracing AI while retaining the ability to direct it. At the same time, individuals must be able to assess their own situation reasonably, understand the risks behind information, choices, and actions, and be prepared to bear responsibility for their consequences.

Viewed at the level of society, cognitive autonomy is a foundation of modern civil society. If individual citizens lose the capacity for cognitive autonomy, the foundation on which society operates will itself be hollowed out.

What kind of education, then, can assume the contemporary mission of cultivating this capacity?

In 1945, Harvard University published a report entitled *General Education in a Free Society*. Because of its plain red cover, it became widely known as the "Harvard Redbook." The report addressed a pressing educational problem of its time: as the division of labor became increasingly fine-grained and specialization intensified, how could a common body of knowledge, shared judgment, and common understanding of responsibility be built for a free society? The report set out Harvard's program of common education in detail, seeking to reconstruct a shared cognitive foundation for society through general courses in culture, the arts, and science. It became one of the important sources of modern general education.

Today, we confront a comparable situation. The operating foundations and cognitive environment of the intelligent society have changed dramatically. If human beings cannot maintain cognitive self-direction, individuals will lose their agency and the mechanisms of civil society will lose their foundation. We need another redbook: one that develops an educational program for citizens' cognitive autonomy in the intelligent society.

This is the historical context in which *Artificial Intelligence General Education: Cognitive Autonomy in the Intelligent Society* was conceived. Beginning with the mission of general education and the disciplinary characteristics of AI, this report develops a feasible path from AI education to the cultivation of cognitive autonomy, and on that basis establishes the theoretical foundations and implementation framework of Artificial Intelligence General Education (AIGE). It also compares the forms of intellectual training provided by AIGE with those provided by established foundational disciplines—mathematics, language and literature, physics, history, and computer science—and examines AIGE's distinctive advantages for developing cognitive autonomy. In doing so, it establishes the field's independent and distinctive educational character.

This report follows the Harvard Redbook in reflecting on the mission of education at a critical historical juncture, and it also adopts the tradition of a plain red cover. It is a redbook for the age of AI. The following table compares the social problems addressed by the two redbooks, with particular attention to their shared contribution to sustaining the foundations of civil society. The comparison further clarifies the historical mission of Artificial Intelligence General Education: it provides cognitive safeguards for individual development and builds the civic foundations required for the healthy operation of society as a whole.

Part II

This redbook grows out of several years of sustained exploration by the Artificial Intelligence General Education (AIGE) Research Center in the Department of Computer Science and Technology at Tsinghua University. The Center was established in January 2025, although its members had already completed a substantial body of related work. Early efforts focused on

Problem addressed by the Harvard Redbook	**Problem addressed by the Tsinghua Redbook**
Disciplinary fragmentation and social diversity weaken the common educational foundation.	Algorithmic mediation, generated content, and cognitive outsourcing weaken citizens' agency in judgment.
Modern society needs citizens capable of thinking, communicating, and judging.	The intelligent society needs citizens capable of regulating belief, evidence, trust, judgment, and responsibility.
General education sustains the basis for public dialogue in civil society.	Artificial Intelligence General Education sustains the foundation of cognitive agency in civil society.
It guards against the fragmentation of public judgment caused by the specialization of knowledge.	It guards against the hollowing out of public judgment caused by the erosion of cognitive agency.

university-level AI education and public science communication. They resulted in specialist books including *Introduction to Artificial Intelligence* (Lin Yaorui and Ma Shaoping, 1989), *Artificial Intelligence* (Ma Shaoping and Zhu Xiaoyan, 2004), *Introduction to Machine Learning* (Wang Dong, 2021), and *Dr. AI: Artificial Intelligence Made Accessible* (Ma Shaoping, 2023), as well as popular works such as *Artificial Intelligence Illustrated* (Wang Dong and Ma Shaoping, 2023). Work on AIGE for primary and secondary schools began to expand after 2024.

In January 2024, the AIGE Center began to formulate a systematic program for Artificial Intelligence General Education. At the time, many proposals were being advanced for how AI education should be conducted. Some scholars advocated hands-on practice through software and hardware programming, extending the tradition of project-based learning. Others emphasized experiences with large models and instruction in prompt writing, reflecting the contemporary surge of interest in large models. Still others favored robotics projects and learning through competitions, continuing the established logic of competitive programs. A further proposal was to simplify university courses in AI and machine learning for non-computer-science majors and for primary and secondary students, a familiar approach in cross-disciplinary university teaching. Each approach has some merit, yet none by itself fulfills the aims of general education. Some cannot reach all students; others privilege a narrow range of content and do not suit the general-education goal of developing thought and cognitive capacity.

We chose a distinctive path: the design presented in this redbook. The Tsinghua AIGE framework takes the cultivation of cognitive autonomy as its organizing goal and the disciplinary characteristics and modes of thought of AI as its foundation. It seeks to build a system of Artificial Intelligence General Education for all citizens and for the intelligent society of the future.

We emphasize foundational concepts in order to establish a coherent scientific conceptual system; attend to basic methods in order to construct a knowledge map of AI; analyze AI applications in order to reveal the principles behind them; and introduce interdisciplinary

advances in order to develop a view of the frontiers of AI. The scientific ethos of AI, training in AI-related modes of thought, and discussion of ethical responsibility run through every chapter, guiding the development of cognitive autonomy.

In April 2025, three volumes in the *Tsinghua Artificial Intelligence General Education Series for Primary, Secondary, and University Education*, covering primary, junior-secondary, and senior-secondary levels, were published on the basis of the Tsinghua AIGE framework. Lecture materials, lesson plans, student worksheets, and practical manuals were released openly at the same time. These resources have since been downloaded and used by thousands of schools and universities. We assembled more than 300 partner schools to conduct teaching research with these materials and improve instructional methods; worked with more than 200 certified instructors to co-develop course resources; and partnered with Tencent's volunteer-teaching platform to provide remote dual-teacher instruction in more than 2,000 rural primary schools across China.

In August 2026, building on this earlier exploration, the Tsinghua AIGE framework released grade-specific editions in the *Tsinghua Artificial Intelligence General Education Series for Primary, Secondary, and University Education*. These span Grades 3 through 6 in primary school, Grades 7 and 8 in junior secondary school, and Grades 10 and 11 in senior secondary school, with one volume per semester and sixteen volumes in total. This edition establishes three spiraling cycles across primary, junior-secondary, and senior-secondary education and includes substantially more learning activities and exercises designed to cultivate cognitive autonomy (early implementation began in Henan and Anhui provinces in 2025).

The design principles and basic criteria underlying these resources are explained in Parts III and IV of this book, and their overall structure is presented in the appendices. Artificial Intelligence General Education is an open field. Provided that the overarching goal of cultivating cognitive autonomy is maintained, it can accommodate many different designs. Specific learning content should remain flexible, diverse, and responsive to new developments. The Tsinghua AIGE framework is therefore one among many possible approaches, although its reliability has been strengthened through extensive testing in real classrooms.

We close by returning to the making of this redbook. We thank all the teachers who participated in its teaching and research activities. Their dedicated classroom practice, rigorous professional discussion, and generous sharing of materials have made the Tsinghua AIGE framework more systematic, feasible, and reliable. Their work has also given us the confidence to bring our practical exploration together in this redbook.

The Editors
Tsinghua University
August 2026

Contents

Preface **v**

Overview: From General Education in a Free Society to Cognitive Autonomy in an Intelligent Society **1**

0.1 Historical Point of Reference: From a Common Educational Foundation to Cognitive Autonomy . . . 2

0.2 Cognitive Autonomy . . . 2

0.2.1 Defining Cognitive Autonomy . . . 2

0.2.2 The Social Significance of Cognitive Autonomy . . . 4

0.3 Artificial Intelligence General Education . . . 4

0.3.1 Defining Artificial Intelligence General Education . . . 4

0.3.2 The Content Foundations and Cognitive Training Mechanisms of Artificial Intelligence General Education . . . 5

0.3.3 The Distinctive Contribution of Artificial Intelligence General Education to Cognitive Training . . . 6

0.4 From a Cognitive Goal to a Five-Part Content Framework . . . 7

0.5 Differentiated Implementation . . . 9

0.5.1 Primary and Secondary Education: Five Horizontal Content Areas and Three Vertical Stages . . . 9

0.5.2 Higher Education: Understanding Methods, Disciplinary Transfer, and Interdisciplinary Collaboration . . . 10

0.5.3 Lifelong Learning: A Common Core, Contextual Modules, and Frontier Updates . . . 10
0.6 Structure of the Book and Path of Argument . . . 11
0.7 Conclusion: Free Personhood in an Intelligent Society . . . 13

I Historical Context: Why Artificial Intelligence General Education Has Become Essential 14

Introduction to Part I 15

1 The Historical Mission of General Education: From the Crisis of Specialization to a Common Educational Foundation 17
1.1 Why General Education Became a Problem of Modern Education . . . 17
1.2 The Triumph of Specialization . . . 19
1.3 Intellectual Lineages of General Education . . . 21
1.4 Harvard Redbook: Common Education in a Free Society . . . 23
1.5 The Threefold Structure of a Common Educational Foundation . . . 24
1.6 The Internal Tensions of General Education . . . 25
1.7 The Historical Transition to the AI Age . . . 26
Chapter Summary . . . 27

2 The Intelligent Society and Its Cognitive Environment 28
2.1 AI and the Intelligent Society . . . 28
2.1.1 From the Information Society to the Platform and Algorithmic Societies . 29
2.1.2 From Technical Systems to Sociotechnical Systems . . . 29
2.2 A Working Definition of the Intelligent Society . . . 30
2.3 Basic Characteristics of the Intelligent Society . . . 31
2.4 The Cognitive Environment of the Intelligent Society . . . 32
2.4.1 Information Overload: Attention as a Scarce Resource . . . 33
2.4.2 Mixed Provenance: Opaque Knowledge-Production Chains . . . 33
2.4.3 Generated Content: The Separation of Fluency from Reliability . . . 34
2.4.4 Algorithmic Mediation: The Continuous Organization of Visibility and Attention . . . 34
2.4.5 Boundary Constraints: The Conditional Character of Model Capabilities 35
2.4.6 Governance by Metrics: Proxy Variables as Forces That Shape Reality . . 35
2.4.7 Automated Decision-Making: Judgment Separated from Responsibility . . 36
2.4.8 Relations among the Seven Aspects and Their Educational Implications . 36
2.5 Chapter Summary . . . 38

3 Cognitive Autonomy 40
3.1 Defining Cognitive Autonomy . . . 40
3.2 Six Dimensions of Cognitive Autonomy . . . 43

3.3 Individual Manifestations of Cognitive Autonomy 47
3.4 The Social Significance of Cognitive Autonomy 48
3.5 Relationship to Related Educational Concepts 50
3.6 Possible Objections and Responses 52
3.7 Chapter Summary 53

4 Artificial Intelligence General Education 55
4.1 Defining Artificial Intelligence General Education 55
4.2 AI Literacy and Cognitive Autonomy 57
4.3 Further Discussion of AIGE 59
4.3.1 The Dual Social Significance of Artificial Intelligence General Education: A Common Foundation and the Human as Cognitive Subject 59
4.3.2 Artificial Intelligence General Education as Distinct from a Simplified or Popularized Form of Professional Education 60
4.3.3 Artificial Intelligence General Education as Distinct from an Extension of Information Technology Education 62
4.3.4 The Independent Educational Standing of Artificial Intelligence General Education 64
4.4 Current Misalignments 64
4.4.1 The Tool-Centered Tendency: Reducing General Education to Operational Skill 66
4.4.2 The Technology-Centered Tendency: Reducing the Common Foundation to Algorithms and Programming 67
4.4.3 The Activity- and Competition-Centered Tendencies: Substituting Partial Achievements for Education of All 69
4.4.4 The Frontier-Chasing and Fragmentation Tendencies: Substituting Topical Content for a Stable Structure 70
4.4.5 The Ethics-as-Add-On Tendency: Substituting Risk Warnings for Judgments of Responsibility 71
4.4.6 The Superficial-Assessment Tendency: Substituting Visible Products for Judgments of Capacity 72
4.4.7 The Superficial-Teacher-Training Tendency: Substituting Tool Training for the Development of Teaching Capacity 74
4.4.8 The Common Sources of Misaligned Aims 74
4.5 The Cognitive Turn in Policy and Competency Frameworks 76
4.6 Chapter Summary 78

II Theoretical Foundations: How Artificial Intelligence General Education Cultivates Cognitive Autonomy 80

Introduction to Part II 81

5 Artificial Intelligence as a Science 82
5.1 How Artificial Intelligence Became a Science . . . 83
5.2 The Object of AI Research: Intelligent Behavior . . . 84
5.3 Two Basic Routes toward Machine Intelligence . . . 84
5.3.1 The Relationship between the Two Routes . . . 86
5.4 A Definition of Artificial Intelligence . . . 87
5.5 The Scientific System of Artificial Intelligence . . . 87
5.5.1 Real-World Tasks and System Implementation . . . 89
5.5.2 Mechanisms and Methods for Realizing Intelligent Behavior . . . 90
5.5.3 Mathematical and Computational Foundations . . . 91
5.5.4 Connections among the Three Levels . . . 92
5.6 Disciplinary Characteristics of Artificial Intelligence . . . 93
5.7 Modes of Thought in Artificial Intelligence . . . 94
5.7.1 Formalizing Real-World Problems as Computational Problems . . . 94
5.7.2 Learning Regularities from Data . . . 95
5.7.3 Using Domain Knowledge as a Constraint . . . 96
5.7.4 Reasoning under Uncertainty . . . 97
5.7.5 Balancing Performance and Efficiency . . . 97
5.7.6 Seeking Simple and Effective Methods First . . . 98
5.7.7 Testing Solutions through Practice . . . 99
5.7.8 Connections among the Modes of Thought . . . 99
5.8 A Developing Science . . . 100
5.9 Chapter Summary . . . 101

6 From Computational Intelligence to Social Force 103
6.1 How Intelligent Behavior Becomes a Social Force . . . 103
6.2 How Social Activity Becomes Computable . . . 105
6.2.1 Real-World Objects Become Data: What the System Can See . . . 105
6.2.2 Historical Experience Becomes a Model: How the System Makes Judgments 106
6.2.3 Social Objectives Become Metrics: What the System Pursues . . . 107
6.3 How Computational Results Enter Real-World Action . . . 108
6.3.1 Organizations and Institutions Give Model Outputs Social Efficacy . . . 108
6.3.2 The Redistribution of Power to Act across the System Chain . . . 109
6.3.3 The Reorganization of Responsibility along the System Chain . . . 109
6.4 From Local Effects to Structural Impact . . . 110
6.4.1 Replication at Scale Turns Local Rules into Common Conditions . . . 110
6.4.2 System Connections Allow Effects to Propagate across Contexts . . . 111
6.4.3 Social Feedback Enables Systems to Participate in Shaping Reality . . . 112
6.5 How an AI-Mediated Cognitive Environment Takes Shape . . . 113
6.5.1 Artificial Intelligence Enters between People and the Real World . . . 113
6.5.2 Artificial Intelligence Connects Cognition and Action . . . 113

6.5.3 Systems in Continuous Operation Become the Cognitive Environment of the Intelligent Society . . . 114
6.6 Chapter Summary . . . 115

7 Cultivating Cognitive Autonomy through AIGE 116
7.1 Framing the Problem . . . 117
7.2 How AIGE Cultivates Cognitive Autonomy . . . 119
7.2.1 Understanding Intelligent Mechanisms: Recognizing the Conditions under Which Conclusions Are Formed . . . 119
7.2.2 Participating in Computational Construction: Subjecting Understanding to Operational Tests . . . 120
7.2.3 Analyzing Social Embedding: Judging the Real-World Effects of Computational Results . . . 121
7.3 Induction . . . 122
7.3.1 The Basic Meaning of Induction . . . 122
7.3.2 The Cognitive Basis of Inductive Capacity . . . 122
7.3.3 Educational Objectives for Inductive Capacity . . . 123
7.3.4 A Classroom Case for Inductive Capacity . . . 123
7.4 Generalization . . . 125
7.4.1 Why Generalization Is a Core Capacity in the Age of AI . . . 125
7.4.2 Foundations of Generalization in the Learning Sciences . . . 125
7.4.3 Educational Objectives for the Capacity for Generalization . . . 126
7.4.4 A Classroom Case for Training Generalization . . . 126
7.5 Judgment . . . 127
7.5.1 Judgment Is More Than Reaching a Conclusion . . . 127
7.5.2 Psychological and Educational Foundations of Judgment . . . 128
7.5.3 Educational Objectives for the Capacity for Judgment . . . 128
7.5.4 A Classroom Case for Training Judgment . . . 129
7.6 Self-Awareness . . . 130
7.6.1 Why Self-Awareness Is Necessary . . . 130
7.6.2 Research Foundations of the Capacity for Self-Awareness . . . 131
7.6.3 Educational Objectives for the Capacity for Self-Awareness . . . 131
7.6.4 Classroom Tasks for Training Self-Awareness . . . 132
7.7 The Four-Capacity Cycle . . . 133
7.8 Evidence for Assessment: How to Observe the Four Capacities . . . 134
7.9 Chapter Summary . . . 136

8 AIGE and Other Disciplines 138
8.1 The Question: Why Compare Disciplines? . . . 138
8.2 Skills and Intellectual Training . . . 139
8.3 Mathematics: Necessary Reasoning under Formal Constraints . . . 140
8.4 Language and Literature . . . 141

8.5 Physics: Causal Explanation under Empirical Constraints . . . 141
8.6 History: Contextual Judgment from Limited Evidence . . . 142
8.7 Executable Construction in Computer Science . . . 143
8.8 The Distinctive Intellectual Training of AIGE . . . 143
8.9 Chapter Summary . . . 145

III Content Framework: What Artificial Intelligence General Education Should Teach 147

Introduction to Part III 148

9 Foundational Concepts of Artificial Intelligence 151
9.1 Why Foundational Concepts Are Indispensable . . . 152
9.2 What Is Artificial Intelligence? . . . 153
9.3 Origins of Artificial Intelligence . . . 153
9.4 The Formation of AI as a Discipline . . . 155
9.5 A History Organized Around Shifts in Method . . . 156
9.6 Distinguishing Easily Confused Concepts . . . 157
9.7 Principles for Designing Content on Foundational Concepts . . . 160
9.8 Recommended Content and Its Levels . . . 161
9.8.1 Common Core . . . 162
9.8.2 Extension Topics . . . 162
9.9 Foundational Concepts and Cognitive Autonomy . . . 163
9.10 Chapter Summary . . . 163

10 Foundational Methods of Artificial Intelligence 165
10.1 AI Methods as a System . . . 166
10.2 Two Basic Paths . . . 167
10.2.1 Knowledge-Based Methods: Converting Human Understanding into Machine-Usable Structures . . . 167
10.2.2 Learning-Based Methods: Forming Models from Experience . . . 168
10.2.3 The Two Paths as Complements . . . 168
10.3 Five Basic Elements of Machine Learning . . . 169
10.4 The Machine-Learning System . . . 172
10.5 Three Basic Modes of Learning . . . 173
10.6 Four Methodological Traditions . . . 175
10.7 Deep Learning . . . 178
10.8 Large Models . . . 180
10.9 Multimodal Models . . . 181
10.10 Agents: From Answering Questions to Taking Action . . . 181
10.11 Principles for Designing Content on AI Methods . . . 182
10.12 Recommended Content and Its Levels . . . 183

10.12.1 Common Core . . . 183
10.12.2 Extension Topics . . . 184
10.13 AI Methods and Cognitive Autonomy . . . 184
10.14 Chapter Summary . . . 185

11 AI Applications 186
11.1 Why AI Applications Deserve a Distinct Place in the Curriculum . . . 186
11.2 Boundaries of Application Education . . . 187
11.3 Principles for Selecting Representative Applications . . . 188
11.4 A Common Framework for Analyzing Applications . . . 189
11.5 Machine Vision . . . 191
11.6 Machine Audition . . . 193
11.7 Language Processing . . . 194
11.8 Game-Playing and Embodied Action . . . 195
11.9 Search and Recommendation . . . 197
11.10 Five Groups of Applications . . . 198
11.11 Principles for Organizing the Teaching of Applications . . . 200
11.12 Common Core and Extension Content . . . 201
11.12.1 Common Core . . . 201
11.12.2 Extension Content . . . 201
11.13 AI Applications and Cognitive Autonomy . . . 201
11.14 Chapter Summary . . . 202

12 Interdisciplinary Integration of AI 204
12.1 Why Interdisciplinary Integration Matters . . . 205
12.2 The Difference Between Interdisciplinary Integration and Application Analysis . 205
12.3 A Six-Stage Framework for Interdisciplinary Integration . . . 206
12.4 Six Domains . . . 208
12.5 Recommended Content Structure for the Six Domains . . . 211
12.6 Selecting and Organizing Interdisciplinary Content . . . 213
12.7 Common Core and Case-Based Extensions . . . 214
12.7.1 Common Core . . . 214
12.7.2 Case-Based Extensions . . . 214
12.8 Interdisciplinary Integration and Cognitive Autonomy . . . 215
12.9 Chapter Summary . . . 215

13 AI Risk, Ethics, and Responsibility 217
13.1 Why Risk, Ethics, and Responsibility Require a Chapter of Their Own . . . 217
13.2 Distinguishing Risk, Ethics, Responsibility, and Governance . . . 218
13.3 A Framework for Classifying AI Risks . . . 220
13.4 Knowledge and Information Risks . . . 221
13.5 Data, Privacy, and Intellectual-Property Risks . . . 223

13.6 Fairness and Inclusion Risks . . . 223
13.7 Reliability, Safety, and Resilience Risks . . . 224
13.8 Risks to Human Agency and Development . . . 225
13.9 Social, Institutional, and Sustainability Risks . . . 226
13.10 Highly Autonomous Systems and Long-Term Control Risks . . . 227
13.11 Core Principles of AI Ethics . . . 227
13.12 The Chain of Responsibility . . . 228
13.13 Independent Content and Dedicated Learning Activities . . . 230
13.13.1 A Seven-Step Deliberative Framework . . . 230
13.13.2 Recommended Dedicated Learning Tasks . . . 230
13.14 Integrating Ethics Throughout the Other Four Content Areas . . . 232
13.15 Design Principles for Risk and Ethics Content . . . 233
13.16 Common Core and Extension Content . . . 233
13.16.1 Common Core . . . 233
13.16.2 Extension Content . . . 234
13.17 Risk, Ethics, Responsibility, and Cognitive Autonomy . . . 234
13.18 Chapter Summary . . . 235

IV Differentiated Implementation: Artificial Intelligence General Education Across Educational Stages and Learner Groups 236

Introduction to Part IV 237

14 An Articulated K–12 AIGE Curriculum 240
14.1 Why K–12 Curricula Require Articulation . . . 240
14.2 Educational Foundations for Developmental Progression . . . 242
14.3 Primary School: Cultivating Interest and the Scientific Spirit . . . 244
14.4 Lower-Secondary School . . . 246
14.5 Upper-Secondary School . . . 247
14.6 Spiral Development Across Three Stages . . . 249
14.7 Principles of Curriculum Organization and Instruction . . . 250
14.8 Developmental Progression in Cognitive Autonomy . . . 251
14.9 Chapter Summary . . . 252

15 AIGE at University 253
15.1 Guiding Ideas . . . 253
15.1.1 Goals of University AIGE . . . 253
15.1.2 Defining the Boundary of General Education . . . 255
15.1.3 Cognitive Autonomy at University . . . 255
15.2 Content Organization . . . 256
15.2.1 Part One: Foundational AI Methods . . . 257
15.2.2 Part Two: Interdisciplinary Integration of AI . . . 260

15.2.3 The Place of Risk, Ethics, and Responsibility . . . 261
15.3 Modes of Teaching . . . 262
15.3.1 Basic Teaching Principles . . . 262
15.3.2 Optional Implementation Models . . . 263
15.3.3 Assessment . . . 264
15.4 A Transitional Curriculum . . . 264
15.4.1 Adding Interdisciplinary Practice . . . 265
15.4.2 Progressive Adjustment During the Transition . . . 265
15.5 Chapter Summary . . . 265

16 AIGE for Lifelong Learning 267
16.1 Guiding Ideas . . . 267
16.2 Content Organization . . . 269
16.2.1 Overall Structure . . . 269
16.2.2 Common Core . . . 269
16.2.3 Contextual Modules . . . 271
16.2.4 Frontier Updates . . . 272
16.3 Modes of Learning and Teaching . . . 273
16.3.1 Basic Principles . . . 273
16.3.2 Optional Implementation Models . . . 274
16.3.3 Assessment . . . 274
16.4 Science Communication and Public Learning . . . 275
16.5 Chapter Summary . . . 275

A Primary School 276
A.1 Integrated Stage Edition . . . 276
A.1.1 Textbook Contents . . . 276
A.1.2 Practice Manual Contents . . . 279
A.2 Grade-Level Volumes . . . 280
A.2.1 Grade 3, First Semester . . . 280
A.2.2 Grade 3, Second Semester . . . 281
A.2.3 Grade 4, First Semester . . . 281
A.2.4 Grade 4, Second Semester . . . 282
A.2.5 Grade 5, First Semester . . . 283
A.2.6 Grade 5, Second Semester . . . 283
A.2.7 Grade 6, First Semester . . . 284
A.2.8 Grade 6, Second Semester . . . 285

B Lower-Secondary School 287
B.1 Integrated Stage Edition . . . 287
B.1.1 Textbook Contents . . . 287
B.1.2 Practice Manual Contents . . . 289

B.2 Grade-Level Volumes 291
B.2.1 Grade 7, First Semester 291
B.2.2 Grade 7, Second Semester 292
B.2.3 Grade 8, First Semester 293
B.2.4 Grade 8, Second Semester 294

C Upper-Secondary School **296**
C.1 Integrated Stage Edition 296
C.1.1 Textbook Contents 296
C.1.2 Practice Manual Contents 298
C.2 Grade-Level Volumes 300
C.2.1 Grade 10, First Semester 300
C.2.2 Grade 10, Second Semester 301
C.2.3 Grade 11, First Semester 301
C.2.4 Grade 11, Second Semester 302

D Higher Education **304**
D.1 Textbook Contents 304
D.2 Practice Manual Contents 307

E Lifelong Learning **309**
E.1 Textbook Contents 309
E.2 Practice Manual Contents 311

Overview: From General Education in a Free Society to Cognitive Autonomy in an Intelligent Society

Artificial intelligence now permeates learning, everyday life, work, public governance, and many other domains. Search and recommendation systems shape what people see first; generative models participate in communication and creation; automated evaluation systems inform organizational decisions; and AI agents are beginning to call tools and carry out extended sequences of tasks. Artificial intelligence has consequently assumed two intertwined roles: it is both a basic instrument of everyday life and an increasingly active participant in human cognition.

These changes confront education with new demands. Members of society need to know what artificial intelligence is, understand how it develops its capabilities, use AI systems to address real-world problems, and grasp the risks, ethical questions, and responsibilities that arise as AI enters society. Yet the dissemination of knowledge, instruction in tool use, and rules for responsible use cannot fully answer the educational questions of the AI age. A deeper transformation is occurring within human cognition itself: answers are becoming easier to obtain while the grounds for them may remain obscure; tasks are becoming easier to complete while understanding may fail to develop alongside them; and systems can supply predictions and recommendations while responsibility continues to rest with human beings.

Education must therefore address a more fundamental set of questions. When artificial intelligence helps formulate questions, select materials, organize evidence, articulate positions, and inform decisions for action, how can people make active use of its contributions while retaining their own understanding and judgment? When a model produces polished and apparently complete content that may nevertheless contain bias, error, or fabricated information, how should people allocate their trust? When a method moves beyond a familiar setting to new objects, populations, or environments, how can people identify the conditions under which it remains valid and the boundaries at which it fails? When artificial intelligence participates in high-risk action, who is responsible for verification, explanation, and remediation?

Together, these questions lead to a still more basic one: what kind of person should an intelligent society seek to cultivate? Artificial Intelligence General Education has emerged as a historical response to this question.

0.1 Historical Point of Reference: From a Common Educational Foundation to Cognitive Autonomy

As modern society developed, the division of labor deepened and education became increasingly specialized. Professional education raised the level of vocational preparation, but it also contributed to the fragmentation of knowledge, the weakening of public judgment, and the narrowing of educational aims. General education arose in response. It sought to rebuild a common educational foundation so that educated people, beyond their professional identities, would continue to possess shared understanding, public judgment, and common norms.

Published in 1945, *General Education in a Free Society* gave concentrated expression to this historical concern(Harvard University Committee on the Objectives of a General Education in a Free Society, 1945). Commonly known as the "Harvard Red Book," the report asked what common education learners still required in a free society marked by an ever-deepening professional division of labor if they were to think effectively, communicate clearly, exercise sound judgment, and develop a shared understanding of values. Its curricular proposals plainly bore the cultural imprint of their time, but the proposition at their core remains fundamental: modern society needs each member to develop individual expertise, and it also needs all members to share essential knowledge, intellectual capacities, and a sense of responsibility.

The age of artificial intelligence carries this historical proposition into a new social setting. General education in the past primarily confronted disciplinary fragmentation and the need for a shared cultural foundation. Today, intelligent systems are profoundly changing how information is presented, how knowledge is formed, and how social action is organized. Individuals continue to depend on teachers, experts, and institutions, while relying increasingly on systems for search, recommendation, generation, and intelligent decision-making. The common educational foundation must therefore extend to an understanding of these mediating systems: how to explore appropriate ways of using them, calibrate trust in them, and preserve judgment and responsibility amid dependence on information, thereby sustaining human agency. Cognitive autonomy gives this problem its conceptual form.

0.2 Cognitive Autonomy: A Modern Proposition for General Education

0.2.1 Defining Cognitive Autonomy

Cognitive autonomy may be defined as **the process by which individuals actively regulate the formation of their beliefs, their use of evidence, their allocation of trust, the boundaries of their judgment, and their responsibility for action within a complex cognitive environment.** The capacity for cognitive autonomy is the ability to initiate, sustain, and adjust this process across different tasks and contexts.

Cognitive autonomy exists within open dependence. Knowledge in modern society is highly specialized, and everyone must rely on teachers, experts, institutions, media, databases, technological tools, and social collaboration(Hardwig, 1985; Hutchins, 1995). The autonomy at issue

consists in a person's continuing ability to question, explain, compare, verify, pause, and revise while drawing fully on external knowledge and intelligent systems. Individuals need to know what they are relying on, understand when that reliance is warranted, recognize its limits and risks, and adjust or withdraw trust when circumstances require.

To retain cognitive direction within open dependence is to retain command over the initiation, scrutiny, and coordination of cognitive activity. The questions a person poses determine the purpose and significance of that activity; the way people, tools, and materials are organized determines the role assigned to external capabilities; and the acceptability of the resulting conclusions must be examined by human beings in light of the evidence, relevant boundaries, and attendant risks. Artificial intelligence can help discover and extend questions, retrieve information, perform calculations, generate proposals, and test results. Human beings must continue to direct the selection and formulation of questions and make the final judgment about both results and the consequences of acting on them. Human direction lies in governing the orientation, regulation, and validity of cognitive activity; it does not depend on how many steps a person performs unaided.

Cognitive autonomy comprises six interrelated analytical dimensions. *Source vigilance* concerns who or what system produced a conclusion, as well as the source's expertise, incentive structure, and path of dissemination. *Evidence sensitivity* concerns the materials, data, experiments, or arguments that support a conclusion. *Boundary awareness* concerns whether samples, experience, and models apply to the object, population, time, and setting at hand. *Trust calibration* concerns the degree of trust warranted by the strength of the evidence, the risks inherent in the task, and the consequences of error. *Cognitive agency* concerns whether individuals can proceed from their own questions and purposes, actively exploring and engaging artificial intelligence to extend learning, creation, judgment, and action. *Responsibility judgment* concerns who should verify, decide, supervise, explain, and remedy once a conclusion enters into action.

To translate this framework into curriculum, teaching, and assessment, this book further identifies four capacities: induction, generalization, judgment, and self-awareness. Induction enables learners to form testable understandings from limited materials. Generalization enables them to transfer existing understanding to new contexts while recognizing its boundaries. Judgment enables them to make reasoned choices when evidence is insufficient, consequences are uncertain, and values conflict. Self-awareness enables them to remain attentive to their own cognitive state, the state of their trust, and their state of responsibility. The four capacities form a cycle: induction generates hypotheses, generalization tests their boundaries, judgment arrives at choices, and self-awareness monitors and corrects the entire process.

The six dimensions and the four capacities operate at different levels. The six dimensions provide a basic framework for analyzing the process of cognitive autonomy. The four capacities provide an educational framework for organizing curricular tasks, developmental progression, and evidence of learning. The two frameworks support one another and together constitute this book's full account of cognitive autonomy and the capacity for cognitive autonomy.

0.2.2 The Social Significance of Cognitive Autonomy

Cognitive autonomy bears both on whether individuals can direct their own understanding and action and on whether society can sustain citizens capable of genuine participation in public life. Modern societies entrust a growing range of public affairs to citizens for discussion, judgment, and collective decision. This institutional arrangement presupposes that citizens can form, express, test, and revise judgments of their own. They need to understand the sources of facts and claims, assess evidence and its limits, allocate trust appropriately among experts, institutions, and technological systems, and assume responsibility for their own expressions, choices, and actions. Cognitive autonomy therefore carries a public significance that reaches beyond individual development: it is the cognitive expression of civic agency.

A civil society requires citizens to exchange reasons amid disagreement, evaluate evidence, understand differing positions, and reach decisions for which they can share responsibility through public deliberation and legitimate procedures. Rights of expression and choice can embody citizens' status as agents only when they rest on a corresponding capacity for judgment. When citizens gradually lose the ability to form judgments, regulate trust, and reflect on decisions, the institutional foundations of civil society are weakened as well.

Artificial intelligence lends new urgency to this problem. Generative models, recommendation algorithms, and automated decision systems are entering the selection of issues, the acquisition of information, the organization of reasons, the expression of opinion, and the implementation of action. They can expand citizens' capacity to access information, analyze problems, and participate in public life. They can also generate, filter, and steer human judgment in ways that remain difficult to perceive. If citizens cannot trace information to its sources, verify evidence, recognize boundaries of applicability, calibrate their trust in systems, and assume final responsibility, the circulation of opinion may remain abundant even as the civic agency behind it is gradually hollowed out. Preserving cognitive autonomy is therefore both a necessary condition for individual agency in an intelligent society and a cognitive foundation for the continued functioning of public life and civil society.

0.3 Artificial Intelligence General Education

0.3.1 Defining Artificial Intelligence General Education

Cognitive autonomy establishes the contemporary mission of general education in an intelligent society and gives a concrete form of education its purpose. The ensuing question is what kind of education can assume this mission. Artificial intelligence is one of the forces that most directly shapes people in an intelligent society, making the study of AI a natural means of developing the capacity for cognitive autonomy. General education and AI education converge at this point as the two sources of Artificial Intelligence General Education. This book defines Artificial Intelligence General Education as follows:

Artificial Intelligence General Education is a form of education created through the integration of general education and AI education. It takes the knowledge,

methods, systems, and social operation of artificial intelligence as its content and assumes general education's mission of cultivating the capacity for cognitive autonomy in an intelligent society.

Viewed from the perspective of general education, Artificial Intelligence General Education is a distinctive form of general education that develops the capacity for cognitive autonomy through the teaching of artificial intelligence. Viewed from the perspective of AI education, it is a distinctive form of AI education directed by the aims of general education.

Cognitive autonomy supplies the integrating principle. Studying artificial intelligence enables learners to understand such components of a system as data, models, objectives, evaluation, generalization, error, and feedback, and thereby to recognize the sources and limits of AI capabilities. Examining the social operation of artificial intelligence allows learners to see how platforms, organizations, and institutions use it to filter information, form judgments, and implement action. Knowledge acquisition, capacity development, and the cultivation of responsibility can consequently be organized around the goal of developing the capacity for cognitive autonomy.

0.3.2 The Content Foundations and Cognitive Training Mechanisms of Artificial Intelligence General Education

Once cognitive autonomy has been established as the educational goal, a fundamental question remains: why is Artificial Intelligence General Education capable of developing the capacity for cognitive autonomy? The answer lies in the discipline of artificial intelligence itself and in the process by which AI enters social activity: the computationalization of intelligent activity and the computationalization of social activity.

Artificial intelligence first transforms abstract intelligent activities—learning, reasoning, judgment, creation, and action—into computational processes that can be studied and constructed. In doing so, experience from the world must be selected and represented, the objectives of a task must be made explicit, a system must develop capabilities from data, knowledge, or feedback, and the resulting performance must be tested on new objects and in new environments. Questions about where experience comes from, how objects are represented, how regularities are formed, whether conclusions transfer, how errors are detected, and under what conditions a method remains valid are thus transformed from implicit problems of cognition into observable, comparable, and discussable components of a system.

AIGE courses can organize samples, representations, objectives, models, evidence, feedback, and boundaries as system components open to analysis and testing. Learners can observe how an AI system derives conclusions from limited experience, applies existing capabilities in new situations, makes selections in accordance with its objectives, and reveals its errors and limits. Crucially, studying these processes does more than foster an understanding of AI systems; it also provides a way to reflect on human cognition. Computational implementations may differ markedly from the internal mechanisms of human cognition, yet both confront shared problem structures involving experience, evidence, transfer, and correction. The computationalization of

intelligent activity thus provides a constructive point of reference from which to examine human cognition, giving the fundamental problems of cognition an external and analyzable form.

Once artificial intelligence has developed computational capabilities, it proceeds to enter knowledge production, information dissemination, organizational judgment, and social action. The conversion of real-world objects into data determines what a system can see. The conversion of historical experience into models influences what a system draws upon in forming judgments. The conversion of social goals into metrics and constraints determines the direction in which a system tends. Model outputs are subsequently incorporated into workflows and acquire real-world force through platforms, organizations, and institutions, producing sustained effects through replication at scale, interconnection among systems, and social feedback. Understanding this process makes it possible to ask which parts of reality enter the data, which experiences become sedimented as model tendencies, who sets system objectives, how much authority to act is granted to model outputs, whether those affected can inspect, reject, correct, and appeal decisions, and how responsibility is distributed among developers, deployers, users, and institutions.

The computationalization of social activity renders the sources, objectives, divisions of labor, power relations, institutions, and responsibilities behind artificial intelligence as structures that can be traced. It supplies education for cognitive autonomy with concrete social materials and evidence, enabling learners to analyze how computational systems intervene in human understanding and action and how a computational result becomes a real-world consequence under particular organizational and institutional conditions.

Together, the two forms of computationalization constitute the fundamental objects of study for Artificial Intelligence General Education. The computationalization of intelligent activity principally reveals how conclusions are formed. The computationalization of social activity further reveals the social conditions under which conclusions are formed, circulated, and endowed with real-world force. The former makes visible the fundamental problems involved in cognition; the latter makes visible how collective cognitive activity is socially organized. Their connection enables AI education to engage both individual and social cognition.

These forms of content can become a basis for the capacity for cognitive autonomy only through deliberate curricular organization. Courses must help learners understand how artificial intelligence develops its capabilities. They must also lead learners to examine data sources, the strength of evidence, the setting of objectives, conditions of applicability, social force, and structures of responsibility, while requiring them to compare, verify, explain, revise, and reflect as they use AI. Artificial intelligence thereby becomes simultaneously an object of study, a practical instrument, and material for reflection. Cognitive autonomy can then be cultivated through a process that learners can understand and experience and that educators can test and assess.

0.3.3 The Distinctive Contribution of Artificial Intelligence General Education to Cognitive Training

Every discipline contributes to the development of students' cognitive capacities, with different emphases arising from its objects of inquiry and structure of knowledge. Mathematics foregrounds abstraction, deduction, and formalization. Physics emphasizes observation, experimentation,

induction, and causal explanation. Language and the humanities emphasize meaning, expression, evidence, and historical context. Computer science emphasizes problem decomposition, algorithmic formulation, executability, and systems thinking. Ethics education emphasizes the discernment of values, normative awareness, and responsibility judgment.

Artificial Intelligence General Education assumes general education's mission of cultivating the capacity for cognitive autonomy in an intelligent society. Its particular concern is whether individuals can understand and regulate their own cognitive activity in an information environment characterized by model mediation, mixed sources, and uncertain results; whether they can continue to explore actively within appropriate limits of safety; whether they can allocate trust in external systems judiciously; whether they can recognize the boundaries of conclusions; and whether they can assume responsibility for their final judgments and actions. It makes the individual's standing as a cognitive agent in an intelligent society a focused educational aim.

Cognitive training across disciplines is interconnected. Artificial Intelligence General Education likewise draws on mathematical reasoning, scientific evidence, linguistic expression, historical understanding, and value judgment, just as other disciplines foster independent thought and a sense of responsibility. The distinctive character of Artificial Intelligence General Education arises from an object of study that spans cognitive activity and social activity at once. Learners must understand how intelligent behavior is transformed into computational processes that can be constructed and tested, and how those processes enter society and affect knowledge, judgment, and action. This dual focus provides systematic subject matter and targeted intellectual training for the development of the capacity for cognitive autonomy.

Artificial Intelligence General Education thereby carries forward, under the conditions of an intelligent society, general education's historical mission of sustaining a common cognitive foundation. Together with mathematics, language and literature, physics, history, computer science, and ethics education, it contributes to the development of human cognitive capacities while concentrating on the cognitive problems created by AI's participation in knowledge formation, information dissemination, social judgment, and real-world action.

0.4 From a Cognitive Goal to a Five-Part Content Framework

The computationalization of intelligent activity and the computationalization of social activity define the fundamental objects that Artificial Intelligence General Education must address. Learners need to understand what artificial intelligence is, how it develops capabilities, and how it forms real-world systems. They also need to understand how artificial intelligence enters different disciplines and social activities and how people should use, evaluate, oversee, and govern these systems. The two forms of computationalization provide the sources of curricular content; cognitive autonomy determines how that content should be selected and organized.

To transform these fundamental objects into a stable, coherent, and teachable curriculum, Artificial Intelligence General Education must answer five interrelated questions.

1. What is artificial intelligence, where did it come from, and where do its boundaries lie?
2. How does artificial intelligence represent, learn, reason, generate, and act?

3. How does artificial intelligence form real-world systems, and under what conditions do those systems succeed or fail?
4. How does artificial intelligence enter different disciplines and participate in the production of knowledge?
5. How should people use, oversee, and govern artificial intelligence, and what responsibilities should they assume?

These five questions give rise respectively to five areas of content: foundational concepts of artificial intelligence, AI methods, AI applications, the integration of AI with other disciplines, and risk, ethics, and responsibility. Governance—the means by which values and responsibilities are translated into institutions, processes, and technical mechanisms—is developed within the fifth area.

Foundational concepts establish the historical, disciplinary, and conceptual coordinates needed to understand artificial intelligence. They help learners distinguish artificial intelligence from automation, robotics, algorithms, the Internet, big data, machine learning, deep learning, and generative AI, while locating contemporary technologies within the field's longer development.

AI methods provide the conceptual architecture for explaining system capabilities. Learners need to understand the two fundamental approaches based on knowledge and on learning; the relationships among objectives, models, algorithms, data, and knowledge; and the complete process of training, testing, model selection, deployment monitoring, and feedback-driven updating. Deep learning, large models, multimodal models, and agents must be situated within this larger framework.

AI applications return abstract principles to the workings of real-world systems. Representative applications—including machine vision, machine audition, language processing, game-playing and embodied action, and search and recommendation—can be analyzed through a common framework: "task and setting–data and representation–model and method–system integration–output and evaluation–failure conditions–risk and responsibility." Learners thereby acquire a shared method for explaining unfamiliar applications.

The interdisciplinary integration of AI examines how artificial intelligence enters mathematics, engineering, the physical sciences, biology and medicine, Earth and space sciences, and the social sciences, culture, and the arts. Courses should present the full chain of "domain problem–formalization and representation–data and knowledge–method transfer–domain validation–interpretation and responsibility." Learners can then understand that general methods, when introduced into a discipline, remain answerable to domain knowledge, disciplinary standards of evidence, and professional responsibility.

Risk, ethics, and responsibility constitute an independent area of content and also run through the other four. Risk concerns harms that may occur under conditions of uncertainty. Ethics concerns the values that should be upheld. Responsibility identifies who must prevent, supervise, explain, and remedy harm. Governance translates these requirements into institutions, processes, and technical mechanisms. Privacy, fairness, authenticity, safety, human agency, and social responsibility must be incorporated throughout data selection, model evaluation, application analysis, and domain validation.

0.5 Differentiated Implementation: From School Education to Lifelong Learning

Artificial Intelligence General Education is intended for all learners. Its shared goal must be realized through curricular forms suited to learners' cognitive development, disciplinary backgrounds, and practical needs. Every stage shares the five-part content framework and the goal of cognitive autonomy, while progressively increasing the depth of content, complexity of tasks, openness of practice, and demands of responsibility.

0.5.1 Primary and Secondary Education: Five Horizontal Content Areas and Three Vertical Stages

Primary and secondary curricula adopt a structure of "five horizontal content areas and three stages of vertical progression." Foundational concepts, AI methods, AI applications, interdisciplinary integration, and risk, ethics, and responsibility extend across primary, lower-secondary, and upper-secondary school, ensuring the completeness of curricular content. Fostering interest, building a systematic understanding, and extending knowledge form the leading tasks of the three stages and ensure continuity in learning.

At the primary level, the emphasis falls on fostering interest and a scientific spirit. Through stories, everyday situations, observation, experience, and discussion, courses encourage children to engage with artificial intelligence while developing an initial critical distance from it. Students need to know that AI is designed by people, that it can help people and can also make mistakes, that important information must be checked, that personal information must be protected, and that the use of AI should observe principles of safety, privacy, and honesty. The use of complex algorithms and open-ended tools must remain subordinate to children's developmental and safety needs.

At the middle-school level, the emphasis falls on building a systematic understanding and a broad perspective. Courses organize fragmented experiences into the historical development, conceptual relationships, and methodological structure of artificial intelligence, using "data–model–output–evaluation" as the principal technical thread. Students gradually learn to explain simple systems, analyze the effects of data, identify errors in outputs, evaluate results, and undertake project-based problem solving. Programming may serve as an extension pathway, while systematic understanding, awareness of evidence, and a sense of responsibility form the common foundation. At this stage, cognitive autonomy takes shape as a structured framework for judgment.

At the high-school level, the emphasis falls on extending knowledge and undertaking innovative practice. Students can develop a deeper understanding of key technologies such as deep learning, generative AI, multimodal models, agents, and human–AI collaboration; analyze the mechanisms and development behind representative applications; engage with frontier questions in different disciplines; and cultivate cross-domain transfer, evidence analysis, disciplinary interests, and public responsibility through research projects and complex cases. At this stage, cognitive autonomy takes the form of responsible cognitive action.

Across the three stages, learners progress from engaging with and questioning AI, through understanding and evaluating it, to using and creating it responsibly. The same concepts and cases may be revisited in a spiral curriculum, with increasing levels of abstraction, system complexity, and demands on judgment.

0.5.2 Higher Education: Understanding Methods, Disciplinary Transfer, and Interdisciplinary Collaboration

Building on the foundational understanding developed in primary and secondary education, higher education further advances students' understanding of AI methods, disciplinary transfer, and interdisciplinary collaboration. Students need a systematic grasp of the fundamental modes of thought and research paradigms of artificial intelligence. They must understand the roles played by computation, data-driven induction, the integration of knowledge and data, and experimental validation, and they must situate artificial intelligence within the systems of problems, standards of evidence, and structures of responsibility specific to their own disciplines.

University curricula can comprise two main components: foundational AI methods and the integration of AI with other disciplines. The former deepens theoretical understanding. The latter organizes representative cases from mathematics, engineering, the physical sciences, biology and medicine, Earth and space sciences, and the social sciences, culture, and the arts. Each case should analyze where AI enters the field, the conditions under which it works, the standards by which it is validated, and the boundaries of responsibility.

At the university level, cognitive autonomy continues to be developed through the four capacities of induction, generalization, judgment, and self-awareness, while the objects of training expand from everyday information and general problems to the formation, testing, and application of professional knowledge. Students currently differ greatly in their prior exposure to AI. During a transitional period, university courses may therefore retain the high-school structure of "concepts–methods–applications–integration," drawing on the high-school knowledge framework while increasing analytical depth. Learning time can then shift progressively toward deeper study of methods, the reading of research papers, professional cases, interdisciplinary practice, and innovative research.

0.5.3 Lifelong Learning: A Common Core, Contextual Modules, and Frontier Updates

Lifelong learning extends across professional development, family life, social participation, and retirement. It encompasses employed adults, people undergoing career transitions, adults learning in family contexts, older adults, the general public, and learners with limited digital skills or special needs. Multiple points of entry, flexible pathways, and continuous updating are therefore essential.

Lifelong-learning curricula adopt a structure of "common core–contextual modules–frontier updates." The common core establishes a stable understanding of artificial intelligence and the intelligent society, AI systems encountered in everyday life, foundational methods, large models and agents, and risk, ethics, and responsibility. Contextual modules connect learning with

information access, work and career development, family life, health, aging and care, cultural creation, and integration with professional fields. Frontier updates incorporate new methods, products, applications, and social issues, helping learners revise their existing understanding.

Adult learning can begin with real-world situations, draw on foundational principles to support operation and judgment, and establish a continuing process of learning through "real-world problem–AI task–foundational principle–practical operation–result verification–transfer and reflection." Working professionals need greater emphasis on workflows, human–AI collaboration, organizational data protection, and professional responsibility. Older adults need support that balances everyday convenience, health and care, fraud prevention, social connection, and personal agency. Learners with limited baseline digital skills or special needs require accessible introductory materials, in-person support, accessible design, local-language support, and the continued availability of human services. Public science communication can serve as an entry point and a mechanism for dynamic updating, linked to sustained courses, community learning, and practical activities.

0.6 Structure of the Book and Path of Argument

The book proceeds through the following sequence: how the historical mission of general education becomes a new modern task in an intelligent society; how cognitive autonomy captures that task; how Artificial Intelligence General Education responds to it; why artificial intelligence can cultivate the relevant capacities; what the curriculum should teach; and how it should be implemented for different stages and groups.

Part I begins with the modern task of general education in an intelligent society. It explains step by step why cognitive autonomy should be treated as the primary educational goal and how Artificial Intelligence General Education takes shape on that basis. Chapter 1 returns to the historical mission of modern general education and explains how a common educational foundation responds to specialization and growing social complexity. Chapter 2 draws on relevant research to define the intelligent society and its fundamental characteristics and then analyzes its cognitive environment from seven perspectives. Chapter 3 defines cognitive autonomy and the capacity for cognitive autonomy, proposes six analytical dimensions—source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment—and explains why cognitive autonomy has become the defining mission of general education in an intelligent society. Chapter 4 defines Artificial Intelligence General Education and its relationships with AI education, AI literacy, cognitive autonomy, and the capacity for cognitive autonomy. It also examines major deviations in current practice and the cognitive turn evident in policy and competency frameworks. Together, the four chapters provide a complete argument for the historical positioning of Artificial Intelligence General Education.

Part II explains why Artificial Intelligence General Education can and should assume the task of developing the capacity for cognitive autonomy. Chapter 5 examines the object of study, scientific system, and fundamental modes of thought of artificial intelligence, revealing how AI transforms abstract and implicit intelligent activity into observable tasks, computable represen-

tations, constructible mechanisms, and testable systems. Chapter 6 analyzes how real-world objects, historical experience, and social goals enter data, models, and metrics respectively, and how model outputs become social forces through workflows, organizational institutions, replication at scale, system interconnection, and social feedback. These two chapters examine the computationalization of intelligent activity and the computationalization of social activity respectively; together they establish the scientific foundation of Artificial Intelligence General Education. Chapter 7 translates this foundation into a four-capacity model of induction, generalization, judgment, and self-awareness, together with corresponding approaches to curriculum and assessment, through three pathways: understanding intelligent mechanisms, engaging in computational construction, and analyzing social embedding. Chapter 8 compares the cognitive training provided by mathematics, language and literature, physics, history, and computer science. It explains how Artificial Intelligence General Education organizes these cognitive resources into the complete chain of "data–model–output–action–feedback," thereby developing a distinctive form of cognitive training.

Part III addresses what Artificial Intelligence General Education should teach. Chapters 9 through 13 examine foundational concepts of artificial intelligence, AI methods, AI applications, the integration of AI with other disciplines, and risk, ethics, and responsibility. Each area is organized around its educational function, principles of content selection, common core, extended content, and contribution to cognitive autonomy, allowing rapidly changing technologies to be situated within a relatively stable structure of knowledge.

Part IV considers implementation for different educational stages and social groups. Primary and secondary curricula adopt five horizontal content areas and three stages of vertical progression. University curricula emphasize understanding methods, disciplinary transfer, and interdisciplinary collaboration. Lifelong learning combines a common core, contextual modules, and frontier updates, with differentiated support based on stage of life, practical tasks, baseline digital skills, and special needs.

The appendices present textbook structures and practical projects designed by the Tsinghua University Research Group on Artificial Intelligence General Education for primary and secondary education, higher education, and lifelong learning. These materials bring together existing curricular resources and forms of implementation, providing a reference for the subsequent reorganization of knowledge, tasks, capacities, and assessment evidence around the goal of cognitive autonomy.

The book thus develops a complete line of argument:

General education's historical mission of sustaining a common cognitive foundation → the new cognitive conditions of an intelligent society → cognitive autonomy as a modern task → the response and positioning of Artificial Intelligence General Education → the computationalization of intelligent and social activity → understanding mechanisms, computational construction, and social embedding → the four capacities of induction, generalization, judgment, and self-awareness → the five-part curriculum → differentiated curricular implementation

0.7 Conclusion: Free Personhood in an Intelligent Society

Artificial intelligence is expanding human capabilities while reorganizing the human cognitive environment. Education must help learners understand and use these new capabilities while safeguarding their standing as cognitive agents in an intelligent society. A person who can use AI proficiently but cannot explain its outputs may remain deeply dependent at the cognitive level. A person who can rapidly generate a work but cannot reconstruct its evidence and reasoning has yet to complete the corresponding learning. A person who delegates an important choice to a system but cannot explain the reasons for that choice or its consequences will likewise struggle to assume genuine responsibility.

Artificial Intelligence General Education ultimately seeks to cultivate people who possess the capacity for cognitive autonomy. Such people can form testable understandings from limited materials, transfer what they know to new contexts while recognizing its boundaries, make reasoned choices when evidence is insufficient and values are complex, and remain alert to their own trust, dependence, biases, and responsibilities. They are willing to enlist artificial intelligence in extending their capabilities, and they can also verify, revise, suspend, or reject its use when necessary. They can participate actively in human–AI collaboration while explaining the grounds for their final judgments and assuming responsibility for the consequences of their actions.

From common education in a free society to cognitive autonomy in an intelligent society, the historical mission of general education is extended under new technological conditions. The more deeply artificial intelligence participates in knowledge and action, the more firmly education must place human understanding, judgment, reflection, and responsibility at its center. Artificial Intelligence General Education thus becomes a foundational education for every member of society. It helps people know and understand artificial intelligence; it also helps them preserve independent thought, calibrated trust, and responsible public participation in an intelligent society, thereby cultivating a form of free personhood that is neither directed nor swept along by the external environment.

Part I

Historical Context: Why Artificial Intelligence General Education Has Become Essential

Introduction to Part I

General education has always carried a foundational mission: amid the differentiation of knowledge and social change, it establishes a shared foundation of knowledge, intellectual capacities, value awareness, and public responsibility for members of society. The 1945 Harvard Redbook raised anew the question of what common education a free society requires, against a background in which war, scientific progress, increasing specialization, and the reconstruction of democratic order were intertwined. Today, the development of artificial intelligence has given this historical question a new contemporary significance.

AI technologies and systems are steadily entering knowledge production, information dissemination, education and learning, the organization of work, public governance, and everyday life. Data, models, platforms, and institutions jointly participate in representing reality, filtering information, generating content, forming judgments, and implementing action. Fundamental social activities—including knowledge production, information dissemination, judgment formation, and social action—are increasingly mediated jointly by AI systems and institutions. An intelligent society is thus gradually taking shape.

An intelligent society expands human cognitive capacities while also changing the cognitive environment in which people live. Information and knowledge have become easier to obtain, yet problems involving mixed provenance, unclear evidence, distorted generated content, algorithmic mediation, automated evaluation, and blurred lines of responsibility have also become more pronounced. While making open use of AI and other external resources, individuals need to retain active regulation of their own understanding and action. This book summarizes that requirement as cognitive autonomy: the process by which individuals actively regulate the formation of their beliefs, their use of evidence, allocation of trust, boundaries of judgment, and responsibility for action in a complex cognitive environment.

Cultivating citizens' capacity for cognitive autonomy has therefore become an important goal of general education in an intelligent society, and AI education provides a natural route toward that goal. General education and AI education consequently converge to form the distinctive educational field of "Artificial Intelligence General Education": it takes the knowledge, methods, systems, and social operation of AI as its subject matter and assumes general education's mission of cultivating cognitive autonomy in an intelligent society.

This part develops that argument. Chapter 1 returns to the history of general education and explains its historical mission of rebuilding a common educational foundation. Chapter 2 defines the intelligent society and analyzes the new cognitive environment produced by AI. Chapter 3 defines cognitive autonomy and explains its basic meaning as the contemporary mission of general

education in an intelligent society. Chapter 4 defines Artificial Intelligence General Education, clarifies its relationship to AI education, AI literacy, cognitive autonomy, and the capacity for cognitive autonomy, and analyzes major distortions and emerging shifts in current practice.

Chapter 1

The Historical Mission of General Education: From the Crisis of Specialization to a Common Educational Foundation

Abstract. This chapter examines why general education has repeatedly become a central concern in the history of modern education. Its emergence and reconstruction have often coincided with profound changes in social structures, institutions of knowledge, and the functions of universities. Modern specialized education has increased the efficiency of knowledge production and strengthened professional preparation, while also contributing to the fragmentation of knowledge, the erosion of public judgment, and the narrowing of educational aims. General education has consequently assumed the task of rebuilding a common educational foundation: helping learners develop a shared language, shared forms of reasoning, and shared frameworks for value judgment beyond their professional identities. Tracing the traditions of liberal education, the modern university system, Dewey's thought on democratic education, Whitehead's account of the rhythm of education, the core curricula at Chicago and Columbia, and the 1945 Harvard Redbook, this chapter analyzes the historical mission of general education in modern society. It concludes that general education in the AI age must carry this tradition forward while extending the common educational foundation beyond the cultural community of industrial society to the problem of cognitive autonomy in an intelligent society.

Keywords: general education; liberal education; specialization; common educational foundation; Harvard Redbook; cognitive autonomy

1.1 Why General Education Became a Problem of Modern Education

General education has remained a recurring concern throughout the history of modern education because societies continually ask what kind of person education should cultivate. Once educational

systems expand in scale and become increasingly specialized and vocationally oriented, they must confront a basic tension. Society needs large numbers of people with specialized expertise, yet it also needs them to possess a shared language, mutual understanding, and common frameworks for value judgment that extend beyond their particular occupations. The first demand drives schools to subdivide disciplines, establish specialized programs, and intensify skills training. The second requires education to retain its concern for the development of the whole person and for public reason. General education derives its distinctive mission from this tension.

Historically, general education has taken many forms. Classical liberal education emphasized training in grammar, rhetoric, logic, arithmetic, music, geometry, and astronomy, with the aim of preparing free persons to participate in public life. With the rise of the modern university, scientific research, professional education, and nation-building moved progressively into the institutional core of higher education, reshaping the traditional model of liberal education. Since the nineteenth century, research universities, professional schools, and vocational education have continually expanded. Education has assumed ever greater responsibility for preparing engineers, physicians, lawyers, teachers, administrators, and researchers. Modern education thereby acquired immense social capacity, but at the cost of greater fragmentation of knowledge and a narrowing of educational purpose.

General education emerged to address this internal tension within modern education. Its concern is whether learners can understand the world beyond their professional identities; whether they can place specialized knowledge within wider historical, scientific, social, and ethical contexts; whether they can reason with people from different intellectual backgrounds; and whether they can make judgments on public questions in which facts, values, and action are intertwined. In his nineteenth-century reflections on the idea of a university, Newman (1996) understood university education as the cultivation of the intellect and the formation of habits of mind, rather than as preparation for some immediately useful occupation. Dewey (1916) connected education with civic life and emphasized the school's responsibility to cultivate the capacity to participate in public life. Whitehead (1929) criticized the educational harm caused by "inert ideas" and insisted that knowledge must be integrated with understanding, imagination, and application. Although developed in different historical contexts, these arguments converge on a central claim: short-term skills and isolated knowledge cannot be allowed to absorb the entire purpose of education.

Modern general education also occupies a distinctive position because it must continually balance two pressures. On the one hand, social production and the growth of knowledge require students to enter specialized fields early and acquire capacities that can be verified, put into practice, and transferred to professional settings. On the other hand, the greater the complexity of society, the more individuals need forms of judgment that cross the boundaries of any single profession. Professional education enables people to address local problems effectively; general education helps them understand the place of those problems within society as a whole. Specialized knowledge provides instruments, while general education supplies orientation, an awareness of boundaries, and a sense of responsibility. The two should be understood as mutually sustaining.

Within the argument of this book, Chapter 1 establishes a foundation. It shows that general

education extends far beyond a supplementary place in the curriculum. It is modern society's institutional response to the shared human capacities required under conditions of specialization, expanding civic participation, and technological development. Only after this point is clear can we understand why the AI age requires a new form of general education and why Artificial Intelligence General Education should take cognitive autonomy in an intelligent society as its central aim.

1.2 The Triumph of Specialization and Its Educational Consequences

The rise of the modern university was inseparable from specialization. Humboldt (1810) envisioned a university founded on the unity of research and teaching, academic freedom, and the intrinsic value of inquiry. His conception exerted a far-reaching influence on the research university from the nineteenth century onward. Modern disciplines subsequently acquired stable institutional forms. Academic journals, laboratories, graduate education, professional associations, and systems of occupational qualification jointly shaped modern knowledge production and the preparation of skilled personnel. Specialization accelerated the growth of knowledge and strengthened higher education's support for industrialization, state governance, and the modern system of professions.

This specialization had a clear historical rationale. Modern medicine requires systematic training in anatomy, pathology, clinical practice, and standards of evidence. Modern engineering draws on mathematics, physics, materials science, mechanics, control, and design. Law, finance, education, public administration, and other modern professions have likewise developed complex bodies of professional standards and institutional practice. Without specialization, modern society could not sustain advanced technological and organizational capacities. Flexner (1930) argued in his analysis of universities that the modern university must serve the highest forms of intellectual inquiry while resisting complete subordination to short-term utility. His argument reminds us that specialization is not itself the source of the educational problem. The decisive question is whether specialization constricts people's understanding of the whole, of values, and of public responsibility.

In his critique of American higher education, Veblen (1918) had already warned that commercial pressures, professional interests, and institutional expansion could divert universities from their academic mission. Kerr (2001), in his account of the "multiversity," further showed that by the twentieth century the university had become a complex organization bringing together research, teaching, government programs, industrial collaboration, and public service. This multifunctional structure enlarged the social reach of universities while also dispersing their educational purposes. Students acquired an ever wider range of choices, and universities offered a growing number of specialized pathways. At the same time, shared learning experiences and a common basis for judgment became more difficult to sustain.

The first educational consequence of specialization is the fragmentation of intellectual experience. Students may acquire advanced knowledge in one field while lacking the basic language needed to understand problems in others. Engineering students may be well versed in technical

design yet receive little training in social consequences and ethical responsibility. Students in the humanities may develop strong capacities for textual interpretation yet lack basic scientific and data literacy. Social science students may be able to analyze institutions and behavior while having little understanding of the internal mechanisms of technical systems. Such fragmentation impedes communication across fields and weakens the integrated judgment that complex problems require.

The second consequence is the instrumentalization of educational purpose. When specialization is driven exclusively by the logic of employment, schools can readily reduce their aims to career preparation and the certification of competencies. Students, too, may come to regard learning chiefly as capital that secures entry into a particular occupational channel, rather than as a process of developing enduring intellectual capacities and a stable structure of personhood. In her discussion of the humanities and citizenship in a democratic society, Nussbaum (2010) warns that an excessively utilitarian education erodes the sympathy, imagination, critical reflection, and public responsibility required for democratic life. The warning applies equally to the AI age: the more powerful our technical capacities become, the more education must safeguard human judgment and responsibility.

The third consequence is the weakening of public judgment. Many problems in modern society—climate change, public health, AI governance, energy transitions, urban planning, educational equity, and bioethics among them—cannot be resolved within a single field of expertise. They simultaneously involve scientific facts, statistical evidence, institutional arrangements, cultural meanings, ethical values, and political choices. A person trained only in a highly specialized field may operate with precision and effectiveness inside that field while struggling to understand competing evidence and conflicting values in complex public issues. General education therefore has a vital role in cultivating public reason.

Research on education lends support to these concerns. In their study of learning outcomes among American university students, Arum and Roksa (2011) reported that a substantial proportion of students showed limited improvement in critical thinking, complex reasoning, and writing. The large-scale synthesis by Pascarella and Terenzini (2005) likewise showed that the effects of university education on student development depend on multiple factors, including curricular structure, faculty–student interaction, intellectually demanding tasks, and integrative learning experiences. Such findings do not diminish the value of professional education. They do show that general capacities do not arise automatically from specialized coursework. Shared learning, reflection, expression, the evaluation of evidence, and transfer across contexts require deliberate curricular design.

The modern mission of general education can therefore be stated in preliminary form: **to preserve human wholeness in a specialized society, rebuild common understanding within institutions of differentiated knowledge, sustain public judgment alongside individual development, and uphold value and responsibility alongside technical efficiency**. This mission provides the shared background to many subsequent reforms of education.

1.3 Intellectual Lineages of Liberal Education, Democratic Education, and Common Learning

Modern general education is not the product of a single theory. It draws on several intellectual lineages, including liberal education, democratic education, science education, and the core curriculum. These traditions often stand in tension with one another, yet together they constitute the intellectual resources of general education.

The tradition of liberal education emphasizes the free development of the mind. Newman (1996) held that the value of university education lies in a general cultivation of the intellect through which people acquire habits of clear, precise, careful, and orderly thought. This tradition gives particular weight to language, logic, philosophy, and history and emphasizes the internal relations among branches of knowledge. Its strength lies in preserving the non-instrumental dimension of education and reminding universities that immediate occupational needs cannot dictate their entire mission. Its limitations are equally apparent. When liberal education relies too heavily on an elite canon and humanistic study, it can neglect modern science, social diversity, and the educational needs of a mass public.

The tradition of democratic education emphasizes the relationship between education and associated life. Dewey (1916) understood democracy as a mode through which shared experience is communicated and regarded the school as a crucial setting for organizing experience and cultivating reflection and participation. From this perspective, education contributes to personal development and to the formation of capacities for public life. Students must learn to formulate questions, negotiate possible courses of action, understand others, examine evidence, and assume responsibility for consequences. General education thereby acquires a strong social dimension: it develops people's capacity to participate in public life, which requires forms of understanding and judgment broader than specialized knowledge alone can provide.

The tradition of science education emphasizes that modern citizens must understand scientific methods and the scientific ethos. Whitehead (1929) argued that education cannot remain an accumulation of inert knowledge; knowledge must circulate through understanding, imagination, and application. Bruner (1960) emphasized that every discipline has a fundamental structure and that education should help students understand its central ideas and modes of inquiry. These arguments are especially important for general education. Its task is deeper than offering superficial introductions to multiple disciplines: it must help students understand how different disciplines formulate questions, produce evidence, construct explanations, and test conclusions.

The core-curriculum tradition sought an institutional response to the problem of common learning. Columbia University's course in "Contemporary Civilization" originated in 1919, in the aftermath of the First World War. It was designed to bring students into sustained engagement with the urgent problems of modern society and to develop their powers of expression through reading, discussion, and writing (Columbia College, 2026). The University of Chicago introduced its core curriculum in the autumn of 1931. Its "New Plan" sought to counteract an excessively dispersed elective system through common coursework and to preserve a shared intellectual experience in undergraduate education (The College, University of Chicago, 2026).

These initiatives show that general education is at once an educational idea and a curricular institution.

Table 1.1 summarizes these intellectual lineages. The summary does not exhaust the origins of general education, but it reveals the multiple traditions from which it has developed.

Table 1.1: Selected intellectual lineages of general education

Intellectual lineage	Central concern	Representative formulation	Implications for AIGE
Tradition of liberal education	Intellectual cultivation, habits of mind, and the unity of knowledge	Newman's account of the idea of a university, emphasizing intellectual cultivation and relations among branches of knowledge	AI education must extend beyond the operation of tools to human understanding and habits of judgment
Tradition of democratic education	Associated life, reflection on experience, and public participation	Dewey's account linking education with democratic experience	AI education should serve public reason and responsible participation in an intelligent society
Tradition of science education	Disciplinary structure, standards of evidence, and methods of inquiry	Whitehead's critique of inert ideas and Bruner's emphasis on the fundamental structure of disciplines	AI education must explain data, models, evidence, error, and methodological boundaries
Tradition of the core curriculum	Shared learning experiences, common texts, and common questions	The core curricula at Columbia and Chicago	AI education needs a common curriculum for all students and a shared awareness of the questions at stake

These lineages show that general education has no single fixed form. It can be organized through the reading of canonical works, scientific literacy, social problems, interdisciplinary projects, writing and discussion, or a core curriculum. Its enduring core lies in a common purpose: education must help individuals move beyond partial forms of training and develop the capacities required to engage with the world as a whole, participate in public life, and direct their own development.

1.4 Harvard Redbook: Common Education in a Free Society

Published in 1945, *General Education in a Free Society*—widely known as the "Harvard Redbook"—is one of the most important documents in the history of modern general education (Harvard University Committee on the Objectives of a General Education in a Free Society, 1945). Produced by a Harvard committee, the report opened with a foreword in which James Bryant Conant set out its educational concerns. Its historical setting was exceptional. As the Second World War approached its end, American society faced the problems of postwar democratic order, scientific and technological development, mass education, and university expansion. Education had to address a fundamental question for a free society: as specialization and the social division of labor continued to deepen, what common education would enable learners to preserve their distinctively human qualities and a shared sense of civic responsibility?

The Redbook's first major contribution was to frame general education as a problem of common education within a free society, rather than as a matter of internal curricular adjustment within universities. The report distinguished specialized education from general education. Specialized education serves particular professions, disciplines, and technical capacities; general education serves everyone by cultivating the civic responsibilities of the individual as a member of society. The distinction has deep significance for the philosophy of education. It indicates that modern education must fulfill two kinds of task at once: enabling individuals to perform particular roles within the social division of labor and enabling them to understand and participate in a shared social life.

The Redbook's second major contribution was to express the aims of general education in terms of intellectual capacities and value judgment. It emphasized the development of effective thinking, the communication of thought, relevant judgments, and the discrimination of values. In his review of the report, Wilson (1946) likewise summarized its concerns in terms of the capacities to think, communicate, make relevant judgments, and discriminate among values. These aims closely anticipate contemporary discussions of critical thinking, communication, ethical judgment, and civic literacy.

Its third major contribution was to place secondary and higher education within a single educational continuum. The report considered both American secondary education and undergraduate education at Harvard College, seeking to establish a coherent framework of common education across the two. This conception of continuity remains instructive for Artificial Intelligence General Education across primary, secondary, and higher education today. Common education must begin early in learners' experience and develop into a more systematic understanding as they advance through successive stages of education.

The Redbook's integrated framework of common education in the humanities, social sciences, and natural sciences is equally instructive. It sought to connect cultural traditions, social institutions, scientific methods, and public responsibility, enabling students to understand both human experience and modern science and society. Although its curriculum bears the unmistakable marks of its time and a Western-centered outlook, the way the Redbook framed the problem remains valuable. A free society cannot rest solely on dispersed forms of professional competence. It also requires shared learning experiences that give people from different backgrounds a

foundation for communication, common judgment, and responsibility.

The Redbook has, of course, also attracted substantial criticism. The cultural tradition, canon, and social vision on which it relied could not adequately address the later concerns of multiculturalism, gender equality, racial justice, globalization, and postcolonial critique. Kliebard (2004), in his history of the American curriculum, shows that twentieth-century curricular reform unfolded through persistent conflict among educational orientations that included humanism, social efficiency, developmentalism, and social reconstruction. The Redbook represents a powerful humanistic and common-culture position within this debate. Its limitations remind us that every common educational foundation must be reinterpreted as historical conditions change; it cannot be fixed once and for all as a list of knowledge drawn from a particular civilization or social class.

This book therefore invokes the Redbook to inherit its mode of posing the question, rather than to reproduce its particular curricular program. The Redbook of 1945 addressed the relationship among a free society, specialization, and a common educational foundation. A Redbook for the AI age must address the relationship among intelligent society, cognitive mediation, and shared cognitive capacities. The historical conditions differ, but the educational problem displays a structural continuity: the more complex society becomes, the more it requires a common foundation; the more differentiated knowledge becomes, the more it requires integrated understanding; and the more powerful technology becomes, the more it requires human judgment and responsibility.

1.5 The Threefold Structure of a Common Educational Foundation

The central task of general education can be described as rebuilding a common educational foundation. This foundation does not require everyone to study an identical list of facts, nor can it be created simply by adding a handful of common courses around the margins of professional education. A more adequate conception gives the common educational foundation a threefold structure comprising knowledge, capacity, and responsibility.

First, a common educational foundation includes essential shared knowledge. Every society needs a degree of common background knowledge to make basic communication possible. In modern society, this shared knowledge encompasses historical understanding, scientific literacy, mathematical and data awareness, linguistic expression, social institutions, technological foundations, the ecological environment, cultural diversity, and ethical questions. Its purpose is to provide the minimum common background required for public discussion, so that members of society can develop mutually intelligible understandings of the same problem.

Second, a common educational foundation includes cross-domain capacities. Educational research has long emphasized that the transfer of knowledge requires deep understanding and reflective regulation. In *How People Learn*, Bransford, A. L. Brown, and Cocking (2000) argues that transfer depends on whether learners understand the organizing structure beneath what they have learned and can recognize relevant principles in a new context. Perkins and Salomon (1988) further distinguish between relatively automatic low-road transfer and high-road transfer,

which requires deliberate abstraction. General education must therefore do more than expose students to multiple fields. It must train them to build connections across domains, understand the structure of problems, and transfer concepts and methods.

Third, a common educational foundation includes a sense of responsibility and the capacity for value judgment. Education cultivates capacities while also shaping the purposes toward which they are directed. In his analysis of educational purpose, Biesta (2010) argues that education cannot be centered solely on the measurement of outcomes; it must also negotiate the relationship among qualification, subjectification, and socialization. The point is crucial for general education. A person may possess formidable analytical abilities while lacking the basic judgment needed to decide when those abilities should be used, what ends they should serve, and who must answer for the consequences of action. Such a person remains impaired in both cognition and action. A common educational foundation must therefore encompass ethics, public responsibility, and self-reflection.

These three dimensions are internally connected. Shared knowledge supplies the material through which people understand the world. Cross-domain capacities provide the means to organize and transfer that knowledge. Responsibility guides the use of both knowledge and capacity. Specialized education often separates the three: knowledge is assigned to disciplinary courses, capacities to training programs, and responsibility to ethics courses. General education must bring them back into relation so that learners develop judgment as they acquire knowledge and understand responsibility as they develop their capacities.

At the institutional level, a common educational foundation can be realized in different ways. Core curricula emphasize common reading and discussion; distribution requirements promote engagement across fields; project-based curricula integrate learning through real-world problems; writing courses cultivate expression and argument; scientific literacy courses emphasize evidence and method; and civic education courses focus on public questions and value judgment. Each institutional form has strengths and limitations. What matters is whether the curriculum establishes a genuine awareness of common problems, presents learners with intellectual challenges, and advances their development in the use of evidence, expression, judgment, and reflection.

This point is crucial for Artificial Intelligence General Education. Its curriculum must integrate shared knowledge about artificial intelligence, cognitive capacities that operate across contexts, and ethical responsibility, connecting technical understanding, the cultivation of judgment, and concrete application. Only in this way can it carry forward the threefold structure of general education.

1.6 The Internal Tensions of General Education

General education has always been contested because it must negotiate several tensions that cannot be eliminated once and for all. Understanding these tensions helps prevent the reduction of general education to a curricular slogan.

The first tension lies between commonality and diversity. General education requires a common foundation, yet modern society encompasses diversity in culture, language, class, gender,

region, and intellectual background. A common education that lacks openness can become a vehicle for reproducing a single cultural tradition. Diversity without common questions, however, leaves students without a foundation for public communication. General education should therefore locate commonality in shared questions and shared capacities, rather than in a closed and uniform list of knowledge.

The second tension lies between breadth and depth. General education seeks a broad horizon, but breadth without depth deteriorates into a superficial survey. Professional education seeks depth, but depth without lateral connections narrows the field of vision. Strong general education pursues structured breadth. In several key domains, it enables students to understand the context of problems, standards of evidence, and the boundaries of methods, thereby preparing them for continued learning and communication across fields.

The third tension lies between freedom of choice and institutional requirements. Modern universities value student autonomy, and elective systems allow students to develop their interests and individuality. Yet when everything is left to individual choice, professional pressures, strategic concerns about grades, and immediate interests often discourage students from choosing general education courses. Core curricula and similar programs restrict choice to some degree, but they also secure an institutional basis for common learning. The central questions concern the educational rationale for such requirements, the openness of the curriculum, and the quality of what students are required to study.

The fourth tension lies between professional preparation and human development. Occupational competence is vitally important to students' future lives, and education cannot disregard employment or the social division of labor. Professional preparation, however, must be understood within the longer course of human development. As working environments continue to change, excessively narrow occupational training may rapidly become obsolete. People who can continue learning, judge complex problems, collaborate with others, and understand the technological and social consequences of their work are more likely to sustain their development amid change. Here general education performs a long-term and foundational role.

The fifth tension lies between the measurability of assessment and education's deeper aims. Recall of knowledge and the performance of technical operations are relatively easy to assess; judgment, responsibility, and self-reflection are much harder to evaluate. Educational systems tend to favor what can be readily measured, and curriculum design follows accordingly. Biesta (2010) warns that when educational purpose is governed solely by the logic of measurement, its value dimension is easily compressed. General education therefore needs richer forms of assessment attentive to evidence of learning processes, performance-based tasks, quality of argument, depth of reflection, and transfer across contexts.

1.7 The Historical Transition to the AI Age

The historical mission of general education assumes a new form in the AI age. The rapid advance of artificial intelligence has carried it beyond the domain of tools and techniques and into the basic operating mechanisms of social production and everyday life, where it is becoming

part of the infrastructure of an intelligent society. Within such a society, the fragmentation of knowledge produced by specialization persists, while a deeper problem of cognitive mediation has emerged. In the past, the principal challenge for students was to acquire broader knowledge beyond their specialized fields. Today, they must also understand how data, models, platforms, and automated systems reorganize information, knowledge, and judgment. General education once responded primarily to disciplinary differentiation. It must now also respond to cognitive outsourcing, algorithmic recommendation, generated content, automated decision-making, and model mediation.

This chapter therefore reaches a basic conclusion. Modern general education seeks to establish a common foundation, public judgment, intellectual cultivation, and a sense of responsibility for a free society. In the AI age, both the objects and the environment with which it engages have changed, calling for new conceptual structures and curricular designs. The Harvard Redbook of 1945 sought to determine what common education a free society required. Today, we must determine what cognitive capacities an intelligent society requires. This is the central task of the present book.

Chapter Summary

Beginning with the historical mission of modern general education, this chapter has examined the achievements and consequences of specialized education; traced the traditions of liberal education, democratic education, science education, and the core curriculum; analyzed the representative significance of the Harvard Redbook in the history of modern general education; and proposed that a common educational foundation has a threefold structure comprising knowledge, capacity, and responsibility. General education's central task is to help learners develop common understanding, public judgment, and a sense of responsibility amid the differentiation of knowledge and the growing complexity of society.

The AI age has not diminished the traditional problems of general education. It has made them more urgent. The next chapter examines the formation of intelligent society and the new cognitive predicaments individuals face within it, leading to the central concept of cognitive autonomy.

Chapter 2

The Intelligent Society and Its Cognitive Environment

Abstract. Artificial intelligence is becoming a vital infrastructure of modern society, giving rise to the intelligent society. This chapter addresses two fundamental questions. The first concerns what an intelligent society is and which basic characteristics define it. Scholars have developed several lines of inquiry around the information society, the platform society, the algorithmic society, general-purpose technologies, distributed cognition, and sociotechnical systems. Building on this research, the chapter defines the intelligent society as a social formation in which human knowledge production, information dissemination, judgment formation, and social action are continuously mediated by artificial intelligence systems in conjunction with institutions. The second question concerns the cognitive environment created by this social formation. The chapter examines seven parallel aspects of that environment: information overload, mixed provenance, generated content, algorithmic mediation, boundary constraints, governance by metrics, and automated decision-making. It explains how an abundance of information can coexist with uncertain reliability, easy access to knowledge with difficulty in discerning its grounds, and systemic efficiency with dispersed responsibility. These conditions, in turn, give rise to the intelligent society's demand for cognitive autonomy.

Keywords: intelligent society; sociotechnical systems; algorithmic mediation; cognitive environment; uncertainty; trust calibration; cognitive autonomy

2.1 Artificial Intelligence and the Emergence of the Intelligent Society

Artificial intelligence has moved beyond the bounds of a single technology or a specific tool. It now reaches into nearly every domain of social production and everyday life, where it combines with existing mechanisms of social operation to form a new social order. We call this the "intelligent society."

No single definition of the "intelligent society" has yet gained acceptance across disciplines. Sociology, communication studies, economics, law, cognitive science, and science and technology studies each describe a different aspect of the structural changes that occur as artificial intelligence enters society. A sound account of the concept therefore begins by examining how existing

scholarship understands the relations among technology, information, knowledge, organizations, and institutions, and by identifying the structures these accounts hold in common.

2.1.1 From the Information Society to the Platform and Algorithmic Societies

Research on the information society first revealed the foundational role of information production, information processing, and network connectivity in modern society. Digitization allows a growing range of social activities to leave data traces that can be recorded, transmitted, and computed, while networks reshape the connections among information, capital, organizations, and people. Research on big data further shows that data analytics expands the scale of information processing and reorganizes the epistemological and methodological structures of knowledge production(Kitchin, 2014). This research tradition establishes a premise for understanding the intelligent society: the objects processed by artificial intelligence arise from ever-expanding processes of social datafication.

Research on the platform society shifts attention from the growing volume of information to the mechanisms that organize information and social activity. Dijck, Poell, and Waal (2018) characterize the platform society as one in which social, economic, and interpersonal activities increasingly unfold through a global ecosystem of platforms. They analyze how datafication, commodification, and selection enter such social domains as news, transportation, health care, and education. Platforms thereby become organizational structures linking users, institutions, markets, and the public sector. They also help determine which content becomes visible, how activities can be accessed, and under what conditions public values can be realized.

Research on the algorithmic society focuses more closely on how algorithms and artificial intelligence participate in social governance. Balkin (2018) observes that large digital platforms stand between states and individuals and govern their user populations through data, algorithms, and artificial intelligence. Here, "governance" extends beyond formal decisions by public authorities. It also includes the continuing effects of ranking, recommendation, classification, prediction, moderation, and rule enforcement on people's opportunities, expression, and behavior. Research on platform and algorithmic societies thus shows that artificial intelligence has entered the processes through which social order is formed.

2.1.2 From Technical Systems to Sociotechnical Systems

Research on general-purpose technologies in economics helps explain the extensive diffusion of artificial intelligence. Bresnahan and Trajtenberg (1995) identify broad applicability, continuous improvement, and the capacity to generate complementary innovations as defining features of a general-purpose technology. Artificial intelligence is one such technology. When combined with particular industries, organizational processes, and bodies of professional knowledge, it gives rise to numerous complementary applications and drives continuing changes in existing tasks, institutions, and divisions of labor. The intelligent society is therefore grounded in the ongoing integration of artificial intelligence with every part of society.

Research on sociotechnical systems reminds us that an algorithmic model is only one component of a real-world system. Whether a technology achieves its intended purpose depends

on the interplay among data, operators, affected parties, organizational rules, legal institutions, and contexts of use. Selbst et al. (2019) shows that separating a technical model from its social context can lead to a mistaken abstraction of the real-world problem. Human choices shape a model's inputs, objectives, and outputs, and the model in turn alters human behavior and institutional processes once it enters an organization. The appropriate basic unit for analyzing an intelligent society is consequently a system jointly constituted by people, technology, and institutions.

Research on distributed cognition in cognitive science supplies another important perspective. Hutchins (1995) demonstrates that cognitive activity may be distributed across people, tools, symbols, and collaborative procedures. As artificial intelligence enters knowledge production and social decision-making, such cognitive functions as memory, retrieval, induction, expression, prediction, and information filtering become further distributed across people and intelligent systems. Individuals remain the agents who understand, judge, and bear responsibility, yet the external structures on which their cognitive activity depends have changed substantially. The intelligent society is therefore also a new environment of distributed cognition.

2.2 A Working Definition of the Intelligent Society

Drawing on the research reviewed above, this book defines the intelligent society as follows:

> **An intelligent society is a social formation in which human knowledge production, information dissemination, judgment formation, and social action are continuously mediated by artificial intelligence systems in conjunction with institutions.**

This definition has two essential components. First, artificial intelligence participates throughout the processes of knowledge production, information dissemination, judgment formation, and social action, thereby becoming part of the underlying conditions on which social activity depends. Second, artificial intelligence systems become deeply coupled with institutions, creating social and organizational structures that mutually sustain one another.

Under this definition, whether a society has become an intelligent society depends principally on whether artificial intelligence has become a routine mediator of knowledge and action and whether AI technologies have become deeply coupled with social institutions. The number of devices of a particular kind or the sophistication of a particular technology is not decisive.

The intelligent society is a descriptive concept. It does not presume that technological progress necessarily produces a better society. Artificial intelligence can improve information processing and social coordination, while also generating new power relations, epistemic risks, and problems of responsibility.

The concept of the "intelligent society" is continuous with those of the digital society, information society, platform society, and algorithmic society. The digital society emphasizes the digitization of social activity; the information society, information production and network structures; the platform society, the organization of social activity through platform ecosystems;

and the algorithmic society, the governance of populations through data and algorithms. The intelligent society incorporates the structures revealed by this scholarship while drawing further attention to the coupled changes in individuals and social organizations that occur as artificial intelligence penetrates social processes and systems of governance.

2.3 Basic Characteristics of the Intelligent Society

The following dimensions summarize several basic characteristics of the intelligent society. They do not provide an exhaustive portrait, but they do reveal the fundamental mechanisms through which artificial intelligence systems and social institutions jointly participate in the operation of society.

- **Artificial intelligence as infrastructure:** Artificial intelligence is expanding from a specialized tool into a general capability available across many domains. It enters everyday operations through cloud services, platform interfaces, embedded systems, and organizational processes. As AI becomes infrastructural, many activities come to rely by default on algorithmic search, model generation, intelligent recommendation, and automated analysis, even though users may not directly see the technical processes involved. The social significance of artificial intelligence consequently extends from its performance on individual tasks to the foundational conditions it provides for knowledge, communication, and action.
- **The datafication and modeling of social processes:** The intelligent society continually converts human behavior, relationships, and environments into data, then uses models to classify, predict, generate, and evaluate. Data are never a complete copy of reality, nor are models reality itself. Through the selection of variables, the scope of samples, labeling rules, and optimization objectives, they construct particular representations of reality. Datafication and modeling make social processes more computable while turning the question of how reality is represented into a fundamental issue for knowledge and decision-making(Kitchin, 2014; Gillespie, 2014).
- **The human–AI hybridization of cognition and action:** Knowledge formation and social action are increasingly accomplished jointly by humans and artificial intelligence. Within this arrangement, humans set objectives, interpret contexts, verify results, and bear responsibility; artificial intelligence performs some of the cognitive or practical work involved in retrieval, matching, generation, prediction, and execution. Human–AI hybridity involves more than combining capabilities. It redistributes who frames the question, who supplies the evidence, who formulates the recommendation, who reviews the result, and who bears the consequences. Cognitive agents in an intelligent society consequently operate within more complex distributed systems(Hutchins, 1995).
- **The platform mediation of social connections:** Platforms connect data, algorithms, service providers, users, organizations, and public agencies, organizing their interactions through rules, interfaces, and ranking mechanisms. What content people encounter, which services they receive, and with whom they connect are increasingly shaped by platform

selection. A platform is more than a channel of communication: it can design rules, record behavior, allocate attention, and create new mechanisms of social coordination(Dijck, Poell, and Waal, 2018; Balkin, 2018).

- **The evolution of systems through feedback:** Intelligent systems adjust continually in response to usage data, while people adapt their behavior to the recommendations, scores, and opportunities that systems present. New behavioral data then reenter models and platforms, creating a feedback loop in which systems shape behavior and behavior updates systems. This loop gives the intelligent society its dynamic character: model performance, usage patterns, organizational rules, and social norms interact and continue to evolve. The consequences of a system must therefore be assessed through its long-term effects after deployment as well as through the functions envisioned at the design stage.
- **The deep coupling of technology, institutions, and values:** Once artificial intelligence enters society, it always operates within specific institutional arrangements. What objectives a model optimizes, which metrics it uses, what forms of appeal it permits, and who bears responsibility are all related to organizational interests, legal rules, and public values. Technical choices have institutional consequences, while institutional arrangements shape the design and use of technology. Fairness, privacy, safety, explainability, the public interest, and responsibility consequently become properties of the system as a whole and must be analyzed within the integrated arrangement of people, models, and institutions(Selbst et al., 2019; Dijck, Poell, and Waal, 2018).

Together, these six characteristics describe the foundations of the intelligent society. The infrastructuralization of artificial intelligence identifies its social position; datafication and modeling, its way of knowing; human–AI hybridization, the agents of cognition and action; platform mediation, its organizational form; evolution through feedback, its operating mechanism; and institutional coupling, its structure of values and responsibility. Collectively, they also transform the conditions under which individuals obtain information, form understanding, and make judgments.

2.4 The Cognitive Environment of the Intelligent Society

The structural changes of the intelligent society ultimately take concrete form in the cognitive environment that each person inhabits. Modern education has long confronted a scarcity of knowledge and information. Digital networks and artificial intelligence have greatly reduced the cost of obtaining information and generating content. Abundant information consumes finite attention and makes attention scarce(H. A. Simon, 1971); research on information overload likewise shows that the quantity of information, task complexity, time pressure, organizational processes, and individual capabilities jointly affect the quality of information processing(Eppler and Mengis, 2004).

Artificial intelligence further extends the chain of information production and social judgment. Content may be generated by a model, ranked by a platform, amplified by a recommendation system, and then converted by organizational algorithms into scores and recommendations for

action. Conclusions become easier to obtain, while assessing their sources, evidence, scope of applicability, degree of credibility, and implications for responsibility becomes more difficult. The challenge extends well beyond distinguishing truth from falsehood. Many conclusions hold only under specified conditions; evidence varies in strength; models may perform differently across groups and contexts; and different levels of risk call for different standards of verification.

This book analyzes the cognitive environment of the intelligent society through seven parallel aspects. Each reveals a form of uncertainty involving attention, provenance, expression, the organization of information, model boundaries, institutional metrics, or responsibility for action. Together, they constitute the fundamental conditions to which Artificial Intelligence General Education must respond.

2.4.1 Information Overload: Attention as a Scarce Resource

Information overload arises when the volume, speed, and complexity of the information received by an individual or organization exceed its processing capacity, impairing understanding, judgment, and action. Even when every item is of some value, finite attention cannot adequately process a large body of competing content. Generative artificial intelligence has reduced the cost of producing and rewriting content, while search, summarization, and recommendation deliver information to individuals continuously. The gap between the supply of information and human processing capacity is consequently growing.

Information overload can encourage people to make rapid judgments based on search rankings, headlines, numbers of likes, familiar sources, or model-generated summaries, substituting ease of access for the quality of evidence. Repeated exposure to the same content may also create the illusion of corroboration by multiple sources, even when those repetitions originate from the same primary source or the same chain of generation.

This environment requires individuals to filter information, establish a hierarchy of sources, compare the strength of evidence, and allocate their verification effort. Education must help learners do more than locate additional material. It must also train them to define the boundaries of a question when materials are abundant, set priorities among different forms of evidence, and decide how much attention verification warrants in light of the risks associated with the task.

2.4.2 Mixed Provenance: Opaque Knowledge-Production Chains

Content in an intelligent society is often produced jointly by humans and models and passes through multiple stages of collection, generation, editing, selection, publication, recommendation, and retelling. A person may supply the ideas for a text, a model may generate its first draft, an editor may revise it, a platform may recommend it, and a search engine may then summarize it. By the time readers encounter the final content, the named author reveals only part of the structure of responsibility and cannot fully represent the chain through which the knowledge was produced.

Sperber et al. (2010) uses the term “epistemic vigilance” for the human capacity to evaluate the reliability of sources and the information they provide. Research on epistemic cognition likewise stresses that learners need to understand where knowledge comes from, how it is

supported by evidence, and under what conditions it can be revised(Hofer and Paul R. Pintrich, 1997; Sandoval, Greene, and Bråten, 2016). The intelligent society broadens the objects of epistemic vigilance: beyond authors and institutions, one must also consider how models, data, platforms, and disseminators contribute to the formation of content.

Judging a source must therefore progress from identifying the named author to tracing the production chain. Institutional reports may contain model-generated material; expert documents may be coedited by teams and tools; and everyday online comments may originate from automated marketing systems. Learners need to distinguish among original research, authoritative reports, journalistic accounts, commercial promotion, platform comments, and model-generated content, and to judge how the different stages affect reliability and responsibility.

2.4.3 Generated Content: The Separation of Fluency from Reliability

Generative artificial intelligence can produce grammatically correct, well-structured, and confidently worded prose, as well as visually coherent images, audio, and video. The quality of expression is thereby further separated from factual reliability. Research on natural language generation shows that models may produce fluent content that is inconsistent with facts, inputs, or external evidence(Ji et al., 2023; L. Huang et al., 2025).

Research on processing fluency indicates that people tend to experience familiarity with, or develop a preference for, information that is easy to process(Reber, Schwarz, and Winkielman, 2004). Generated content can readily lead learners to equate clear expression with correct explanation, structural completeness with sufficient argument, and numerical precision with reliable evidence. Linguistic form can heighten persuasiveness, yet cannot by itself provide epistemic warrant.

Education must evaluate expressive quality separately from epistemic evidence. Scientific conclusions require support from data, experimental design, model assumptions, reproducibility, and peer scrutiny. Historical interpretations require support from primary sources, context, chronology, and explanatory coherence. Public judgments must consider sources, interests, values, and the consequences of action. Fluency aids understanding and communication; truth claims still require external evidence.

2.4.4 Algorithmic Mediation: The Continuous Organization of Visibility and Attention

The information encountered by an individual is always organized in some way. A distinctive feature of the intelligent society is the ability of search, ranking, recommendation, and personalization mechanisms to adjust visible content continuously in response to user data. Gillespie (2014) argues that algorithms possess public relevance because they help determine what counts as important, relevant, and visible. The information stream presented by a platform is jointly produced by algorithmic selection, individual behavior, social connections, the available supply of content, and commercial objectives. It cannot be treated as a direct representation of reality as a whole.

Algorithmic mediation also amplifies feedback. Users' clicks, viewing time, reposts, and purchases become new data with which systems adjust their recommendations; the altered information environment then continues to shape users' interests, beliefs, and actions. Research on platforms shows that algorithmic ranking, individual choice, and social relations all affect exposure to information(Bakshy, Messing, and L. A. Adamic, 2015). The spread of misinformation is likewise shaped by the characteristics of content, social motivations, and platform structures(Vosoughi, Roy, and Aral, 2018; D. M. J. Lazer et al., 2018).

Understanding algorithmic mediation requires attention to both technical mechanisms and human participation. Learners should ask why particular information appears before them, what signals inform its ranking, which objectives the platform seeks to optimize, and how their own behavior reshapes the information stream. Such analysis helps distinguish "what I see" from "what exists in the world."

2.4.5 Boundary Constraints: The Conditional Character of Model Capabilities

Artificial intelligence systems are generally trained and evaluated on specified data before being deployed in new settings. Differences among training data, test data, and real-world environments may result in declining performance, amplified bias, or shifts in risk. Guo et al. (2017) finds that accuracy and confidence in modern neural networks do not align automatically; Recht et al. (2019) observes a decline in model accuracy after recreating a classic image test set; and Hendrycks and Dietterich (2019) uses a benchmark of common corruptions to reveal problems in model robustness under changing environmental conditions.

Boundary constraints also have a social dimension. In their evaluation of commercial facial-analysis systems, Buolamwini and Gebru (2018) find marked disparities in error rates across gender and skin-type groups. Overall accuracy may conceal differences among groups, just as average performance may conceal localized harms. Strong performance on one sample, population, or demonstration setting provides no guarantee of equal reliability with new subjects and in new contexts.

These conditions require learners to develop stable boundary awareness. They should ask what samples were used to build a model, what its evaluation metrics establish, to which populations or cases its conclusions can be generalized, what changes may cause it to fail, and whether its errors have the same consequences across groups and settings. Understanding the conditional character of model capabilities is central to managing uncertainty in an intelligent society.

2.4.6 Governance by Metrics: Proxy Variables as Forces That Shape Reality

Modern organizations make extensive use of quantitative indicators, scoring systems, rankings, and risk models. Once artificial intelligence enters these systems, metrics serve both to describe reality and to guide model optimization and organizational action. Goals that are difficult to measure directly are often converted into computable proxy variables. The choice of proxies affects what a system sees, what it overlooks, and what it rewards.

Espeland and Sauder (2007), in their study of law-school rankings, show that rankings can alter how organizations allocate resources and understand themselves. Strathern (1997), in

an analysis of audit culture, likewise reveals that when an indicator becomes a target, the activity under measurement changes in response. Algorithmic models optimize metrics, while organizations adjust their behavior in response to model outputs. Descriptive numbers can thereby become institutional forces that shape reality.

Learning data in education can help teachers understand parts of the learning process, but cannot capture the full extent of understanding, intellectual development, or creativity. Risk scores in health care, finance, and public administration likewise represent reality only in relation to particular objectives and proxy variables. Learners need to understand how metrics are constructed, what proxy relationships they assume, which dimensions they omit, and how they affect behavior, so that they do not mistake the measurable part for the whole of reality.

2.4.7 Automated Decision-Making: Judgment Separated from Responsibility

When model outputs enter such processes as employment screening, educational early-warning systems, credit approval, clinical assessment, and the allocation of public services, they affect people's opportunities and rights. Automated systems can increase the efficiency and consistency of processing. At the same time, recommendations, decisions, and responsibility may become distributed among model developers, data providers, system deployers, professionals, and organizational managers. As the decision chain grows longer, it becomes harder for individuals to identify who can explain a decision, correct it, and bear responsibility for its consequences.

Lee and See (2004) argues that appropriate trust in automated systems depends on system performance, an understanding of process, and alignment with purpose; Parasuraman and Riley (1997) analyzes the misuse, disuse, and abuse of automation. When organizations treat model outputs as objective evidence, human review can become a formal exercise in confirmation, while responsibility is displaced onto "the result produced by the system." Automated recommendations therefore create a risk that the authority to judge will become separated from the obligation to answer for the judgment.

Different settings require different standards of evidence, levels of human oversight, and mechanisms of accountability. Low-risk content recommendation and high-risk decisions in health care, education, law, and public governance cannot be subject to the same intensity of review. Learners need to ask who sets the objectives, who interprets the outputs, who has the authority to change a decision, who handles appeals, and who bears the consequences. Models can support decisions, while the chain of responsibility must still be explicitly established by people and institutions(National Institute of Standards and Technology, 2023).

2.4.8 Relations among the Seven Aspects and Their Educational Implications

The seven cognitive conditions described above are interconnected. Information overload increases the pressure of selection; mixed provenance lengthens the chain of knowledge production; generated content separates expressive form from reliability; algorithmic mediation organizes visibility and attention; boundary constraints delimit model capabilities; governance by metrics alters organizational objectives and behavior; and automated decision-making carries model

outputs into action and responsibility. These conditions interact, yet each corresponds to a distinct and indispensable task of judgment.

Table 2.1: Seven Aspects of the Cognitive Environment of the Intelligent Society and Their Educational Implications

Cognitive environment	**Structural source**	**Principal cognitive risk**	**AIGE task**
Information overload	Declining costs of content generation; continuous information supply through search, recommendation, and summarization	Fragmented attention; substituting accessibility and repetition for assessment of evidence	Define the boundaries of questions; train selection, source classification, and the prioritization of evidence
Mixed provenance	Production and dissemination chains jointly formed by people, models, institutions, and platforms	Attending only to attribution or form; failing to identify how content was produced	Trace production chains; identify levels of provenance and stages of responsibility
Generated content	Models rapidly produce formally complete, confidently expressed content	Mistaking fluency, completeness, and precision for truth and reliability	Distinguish expressive quality from epistemic grounds; verify facts and evidence
Algorithmic mediation	Ranking, recommendation, and feedback mechanisms continually organize the visibility of information	Mistaking a selected information stream for a comprehensive view of reality	Understand platform objectives, ranking signals, and behavioral feedback
Boundary constraints	Differences among training, testing, and deployment environments	Generalizing local performance to new settings; overlooking group differences	Assess sample representativeness, scope of applicability, and failure conditions

Cognitive environment	**Structural source**	**Principal cognitive risk**	**AIGE task**
Governance by metrics	Organizations represent and manage reality through proxies, scores, and rankings	Equating a metric with the objective; overlooking omissions and behavioral effects	Analyze how metrics are constructed, their proxy relationships, and their institutional consequences
Automated decision-making	Model outputs enter organizational action; judgment and responsibility are distributed among multiple actors	Excessive trust in systems; insufficient review, appeal, and accountability	Distinguish levels of risk; clarify the division of labor between humans and AI and the chain of responsibility

Within this cognitive environment, indiscriminate doubt would make epistemic trust and social cooperation difficult to sustain, while excessive trust would weaken judgment and responsibility. The critical capacity is trust calibration: assigning an appropriate degree of trust to a conclusion or system in light of the source of information, strength of evidence, scope of applicability, situational risk, and consequences of error. Sources, evidence, boundaries, risk, and responsibility jointly determine the appropriate intensity of verification.

2.5 Chapter Summary

This chapter has addressed two questions: what an intelligent society is and what kind of cognitive environment it creates. Drawing together research on the information society, platform society, algorithmic society, general-purpose technologies, distributed cognition, and sociotechnical systems, it defines the intelligent society as "a social formation in which human knowledge production, information dissemination, judgment formation, and social action are continuously mediated by artificial intelligence systems in conjunction with institutions." Six basic characteristics constitute this social formation: artificial intelligence as infrastructure; the datafication and modeling of social processes; the human–AI hybridization of cognition and action; the platform mediation of social connections; the evolution of systems through feedback; and the deep coupling of technology, institutions, and values.

These structural changes, in turn, shape seven interrelated cognitive conditions, each with its own emphasis. Information overload makes attention scarce; mixed provenance makes knowledge-production chains harder to trace; generated content further separates expressive fluency from factual reliability; algorithmic mediation continually organizes the visibility of information; boundary constraints render model capabilities markedly conditional; governance by metrics enables proxy variables to shape organizational behavior; and automated decision-making can

separate the authority to judge from the obligation to bear responsibility. The primary difficulty confronting individuals consequently shifts from obtaining information toward assessing its grounds, recognizing boundaries, calibrating trust, and assuming responsibility.

The intelligent society thereby places new demands on general education. Learners need to understand how intelligent systems participate in knowledge and action, regulate their judgments in light of sources, evidence, scope of applicability, risks, and consequences, and maintain a sense of responsibility in human–AI collaboration. These demands give rise to "cognitive autonomy," the subject of the next chapter.

Chapter 3

Cognitive Autonomy: The Contemporary Mission of General Education in an Intelligent Society

Abstract. In an intelligent society, human cognition is increasingly mediated by data, models, platforms, and institutions. Sources of information are becoming more mixed; content can be generated at scale; algorithms participate in filtering and amplifying information; and model outputs are beginning to enter both judgment and real-world action. General education must therefore help individuals retain direction over their cognition amid open dependence. This chapter defines cognitive autonomy as the process by which individuals actively regulate the formation of their beliefs, their use of evidence, their allocation of trust, the boundaries of their judgment, and their responsibility for action within a complex cognitive environment. The capacity for cognitive autonomy is the ability to initiate, sustain, and adjust this process across different tasks and contexts. The chapter develops the basic structure of cognitive autonomy through six dimensions—source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment. It explains their individual manifestations and social significance, distinguishes cognitive autonomy from critical thinking, metacognition, self-regulated learning, epistemic cognition, information literacy, and media literacy, and responds to major objections to cognitive autonomy as an educational aim. Cognitive autonomy thereby emerges as the contemporary mission of general education in an intelligent society.

Keywords: intelligent society; general education; cognitive autonomy; capacity for cognitive autonomy; source vigilance; evidence sensitivity; trust calibration; cognitive agency

3.1 Defining Cognitive Autonomy

The previous chapter examined the changes that intelligent society has brought to the cognitive environment. Changes in how information is generated, filtered, disseminated, and used place new demands on human cognition. This book brings these demands together under the concept of cognitive autonomy.

Cognitive autonomy may be defined as **the process by which individuals actively**

regulate the formation of their beliefs, their use of evidence, their allocation of trust, the boundaries of their judgment, and their responsibility for action within a complex cognitive environment. The capacity for cognitive autonomy is the ability to initiate, sustain, and adjust this process across different tasks and contexts. This definition can be understood through the following considerations.

- **Cognitive autonomy is a process of actively regulating cognitive activity.**
  Cognitive autonomy extends throughout the processes of acquiring information, understanding knowledge, solving problems, forming judgments, and choosing actions. Individuals must continually identify sources, evaluate the quality of evidence, adjust degrees of trust, recognize the boundaries of judgment, and assume responsibility for actions taken on the basis of those judgments. Cognitive autonomy is continuous and dynamic: it unfolds and changes as task objectives, available evidence, external conditions, and potential consequences change.

  Cognitive autonomy operates at a different level from disciplinary modes of thought such as data-driven reasoning and practical testing. Nor can it be equated with the design, operation, or use of AI tools. Disciplinary modes of thought provide concrete ways of approaching a problem, while tool use consists of concrete actions undertaken to complete a task. Cognitive autonomy runs through the process by which individuals understand and select those approaches, evaluate the results produced by tools, and revise their judgments in light of new evidence. The capacity for cognitive autonomy is manifested in a person's ability to initiate this process actively, sustain it effectively, and adjust it in a timely manner. Because the capacity can operate across different disciplines, tasks, and situations in life, it is both general and transferable.
- **Cognitive autonomy addresses the complex cognitive environment pervasive in modern society.**
  A complex cognitive environment is one in which sources of information are diverse, the quality of evidence varies, knowledge is continually updated, competing judgments coexist, and the generation, filtering, dissemination, and amplification of information are increasingly mediated by algorithms, platforms, organizations, and institutions. The content encountered by individuals in such an environment has often already passed through multiple layers of selection and processing. A greater volume of information does not automatically produce more reliable understanding.

  Artificial intelligence further increases the complexity of this environment. It can rapidly generate content, organize knowledge, offer recommendations, and participate in decision-making; it can also produce conclusions that contain errors or omissions or lack adequate grounds. Individuals must now ask more than whether a particular item of information is true. They must also determine how the information was produced, which conclusions the evidence can support, where the system applies, and how much trust it deserves.

  The specific manifestations, distribution of risks, and forms of governance differ across countries and institutions. Yet structural changes such as information overload, mixed provenance, algorithmic mediation, generated content, and rapidly changing knowledge are

widespread. The active exercise of cognitive autonomy therefore transcends any particular technology, platform, or social system and constitutes a general requirement for entering the modern cognitive environment. The capacity that sustains this process has likewise become a foundational capacity in modern society.

- **Active regulation lies at the heart of cognitive autonomy.**
  The agency emphasized by cognitive autonomy lies in a person's ability to recognize which information, tools, and social mechanisms are shaping their understanding and then actively adjust their cognitive activity in light of the task, the evidence, and the possible consequences. Individuals can accept reliable information, draw on expert knowledge, and use artificial intelligence. They can also recheck the evidence, revise their judgments, and change their actions when conditions shift, evidence proves insufficient, or risks increase.
  Such agency presupposes a basic understanding of AI technology and the ways it operates socially. Understanding how artificial intelligence develops capabilities from data, how objectives and modes of evaluation shape it, and how it enters real life through platforms and organizations enables individuals to form well-grounded trust and confidence in action. On this basis, people can approach artificial intelligence without apprehension, choose and use tools with confidence, and actively explore new ways to learn, create, and solve problems, while continuing to examine whether their understanding, judgment, and responsibility remain under their own active regulation.
  Cognitive autonomy therefore includes an awareness of risk, but it does not end with warnings about risk. It seeks a positive process in which individuals continue to understand, choose, act, and revise within a complex environment. The ability to carry out this process reliably manifests the capacity for cognitive autonomy.
- **Cognitive autonomy preserves human cognitive direction amid open dependence.**
  Human cognition has always been grounded in open dependence. Language, knowledge, education, experts, institutions, and technological tools all participate in the process by which people come to know. Artificial intelligence further enlarges the scope of this dependence by providing external support for information retrieval, the organization of materials, pattern recognition, computational inference, proposal generation, and the expression of content. Cognitive autonomy permits individuals to invoke and organize external cognitive resources selectively in response to the needs of a task; it does not require them to complete every cognitive operation unaided.
  This cognitive dependence is also a fundamental condition of modern knowledge societies. In his discussion of epistemic dependence, Hardwig (1985) observes that individuals in modern knowledge communities must often form beliefs on the basis of others' testimony and expert labor. Hutchins (1995) further demonstrates through his account of distributed cognition that cognition in complex tasks is often distributed among individuals, tools, symbolic systems, and organizational environments. Artificial intelligence makes these relations of dependence denser, faster, and less visible, thereby making it still more important for people to organize them actively.

Cognitive autonomy emphasizes human direction within this open dependence. Such direction lies primarily in retaining command of the initiation, scrutiny, and coordination of the cognitive process. Two roles are especially important to keep on the human side: posing questions and verifying results. The questions posed determine the ends toward which cognition is directed and embody a person's understanding of real needs, values, and the significance of the problem. Verifying results requires people to determine whether a conclusion is adequately grounded, applies to the present context, omits important factors, and can properly serve as a basis for action. Together, these two roles establish the direction, boundaries, and ultimate validity of cognitive activity.

Artificial intelligence can help people discover, elaborate, and generate candidate questions. It can also participate in retrieving evidence, making cross-comparisons, and validating results. Yet the selection and formulation of questions must remain under human direction, and the acceptability of results must remain subject to human verification. The cognitive operations between these two points—searching for information, invoking knowledge, performing calculations, recognizing patterns, generating proposals, and simulating consequences—may be delegated selectively to artificial intelligence according to the nature of the task, the capabilities of the tool, and the degree of risk. The division of labor between people and AI can vary flexibly from task to task. Human direction does not depend on how many steps a person performs personally; it depends on whether the person still governs the posing of the problem, the regulation of the process, and the verification of the result.

Using artificial intelligence to extend human capabilities is thus compatible with maintaining cognitive autonomy. Individuals can draw fully on AI's powers of memory, computation, generation, and analysis while continuing to understand how the task is unfolding, which parts AI is performing, what grounds the results, and which conclusions require further verification. Cognitive autonomy preserves human direction over the overall orientation and validity of cognitive activity, so that the capabilities of artificial intelligence remain embedded in a cognitive process initiated, verified, and ultimately answered for by human beings.

3.2 Six Dimensions of Cognitive Autonomy

The process of cognitive autonomy can be elaborated through six dimensions: source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment. Source vigilance, evidence sensitivity, boundary awareness, trust calibration, and responsibility judgment correspond respectively to five objects of regulation: belief formation, the use of evidence, the boundaries of judgment, the allocation of trust, and responsibility for action. Cognitive agency expresses the inner disposition involved in "active regulation." It inclines individuals to engage with changes in the cognitive environment, explore the new possibilities created by artificial intelligence, and actively seek ways to extend their own capabilities. Its realization in practice depends on the combined support of multiple cognitive capacities. Together,

the six dimensions constitute the basic structure of cognitive autonomy and provide an analytical framework for curriculum design, classroom observation, and process-oriented diagnosis.

- **Source vigilance.** Individuals need to ask who, or what system, produced a piece of information. Its source may be a teacher, expert, research paper, media outlet, institution, platform, model, or commercial system; it may also be a hybrid source jointly produced by people and machines. Source vigilance requires individuals to avoid accepting or rejecting information on the strength of identity alone and to examine the mechanisms of production, relevant expertise, incentive structures, accountability, and pathways of dissemination. The concept of epistemic vigilance proposed by Sperber et al. (2010) emphasizes that human beings depend on communication with others while also having to assess the reliability of both communicators and communicated content. AI environments complicate this problem because model outputs take a linguistic form resembling testimony but lack the relations of responsibility found in ordinary human communication.
- **Evidence sensitivity.** Individuals need to distinguish the form in which a conclusion is expressed from the evidential basis supporting it. A conclusion may be based on personal experience, statistical data, experimental research, expert judgment, historical material, model generation, benchmark tests, or commercial promotion. These kinds of evidence differ in strength and scope of application. Research on epistemic cognition by Chinn, Buckland, and Samarapungavan (2011) shows that learners need to understand how evidence supports knowledge claims and how different disciplines establish justification. In an AI environment, evidence sensitivity means that students do not equate linguistic fluency, visual realism, or numerical precision directly with reliability.
- **Boundary awareness.** AI systems learn from data, and data come from particular populations, settings, languages, cultures, and periods. Performance on one distribution does not guarantee equal reliability in a new setting. Boundary awareness translates the machine-learning problem of generalization into a habit of everyday judgment: does the present case lie beyond the boundaries of the evidence? Does the sample represent the population? Can experience be transferred? Questions of this kind are pervasive in artificial intelligence. For example, the reevaluation of the ImageNet and CIFAR-10 test sets by Recht et al. (2019) found that models could suffer substantial declines in performance on new test sets. Such studies remind educators that claims about performance must be understood in relation to data sources and contexts of use.
- **Trust calibration.** Individuals need to assign an appropriate degree of confidence to their beliefs and choices in light of the strength of the evidence, the quality of the source, the risks of the task, and the consequences of error. Research in psychology and neuroscience shows that confidence in one's judgment does not always correspond to accuracy. In a review of the neural basis of metacognition, Fleming and Dolan (2012) notes that metacognitive accuracy can be distinguished from task performance. In AI research, Guo et al. (2017) finds that the prediction probabilities of modern neural networks may also be poorly calibrated: accuracy and confidence do not coincide automatically. Both people and models can be overconfident, and education must train learners to regulate the strength of their

confidence.

- **Cognitive agency.** On the basis of an understanding of AI's fundamental principles, the boundaries of its capabilities, and its social effects, individuals can approach artificial intelligence with openness and confidence, regard it as a cognitive resource that can be understood, explored, and employed, and actively consider how it might participate in learning, creation, judgment, and problem solving. Cognitive agency is primarily a cognitive disposition: a willingness to understand and experiment with new technologies and an openness to new ways of knowing. In concrete tasks, it is also manifested in a person's ability to begin with their own problems and aims, identify where AI might play a role, and envision possible ways of extending their cognitive reach through external capabilities. Cognitive agency supplies the internal motivation for selecting tools, organizing tasks, and collaborating with AI. It also ensures that regulating sources, evidence, boundaries, trust, and responsibility serves positive learning, creation, and action.
- **Responsibility judgment.** Cognitive autonomy ultimately enters action. A conclusion used for brainstorming requires a relatively low standard of evidence; one used in medicine, law, finance, educational assessment, or public governance demands a much higher standard. Research on trust in automation by Lee and See (2004) shows that trust between people and automated systems must correspond to system capability, process transparency, and the purpose of use. The NIST Artificial Intelligence Risk Management Framework likewise emphasizes that AI risks must be managed in relation to specific contexts and affected parties (National Institute of Standards and Technology, 2023). Responsibility judgment requires learners to understand that different settings call for different standards of verification, requirements for human review, and arrangements for accountability.

Table 3.1: Six Dimensions of Cognitive Autonomy

Dimension	**Core Question**	**Manifestation in AI Contexts**	**Manifestation in Social Contexts**
Source vigilance	Which information sources and generative mechanisms are influencing my understanding?	Attending to how models, training data, prompts, retrieval systems, platform interfaces, and deploying institutions shape outputs	Attending to how authors, institutions, media, expert identities, interests, and pathways of dissemination shape information

Dimension	**Core Question**	**Manifestation in AI Contexts**	**Manifestation in Social Contexts**
Evidence sensitivity	What evidence supports the present conclusion, and how strong is that support?	Distinguishing model generation, retrieved evidence, benchmark testing, human review, and fact-checking, and determining whether an output is adequately grounded	Distinguishing personal experience, statistical data, experimental research, historical material, expert judgment, and commercial promotion, and evaluating their evidential force
Boundary awareness	Under which conditions and within what scope does the present conclusion hold?	Examining training distributions, test settings, group differences, linguistic and cultural differences, and conditions of model applicability	Analyzing the representativeness of samples, the transfer of experience, the extension of policies, historical analogies, and regional differences, and avoiding judgments that exceed the evidence
Trust calibration	How much trust should I place in this information, actor, or system?	Adjusting trust in AI answers, probability scores, risk predictions, and generated content in light of the task, evidence quality, model performance, and potential risk	Adjusting trust in media, authorities, peers, institutional rankings, and personal intuition in light of source reliability, evidential strength, and context
Cognitive agency	Can I engage actively with changes in the cognitive environment and explore ways to extend my own cognitive capabilities?	On the basis of an understanding of AI's capabilities and boundaries, approaching it without apprehension and actively exploring its possible roles in learning, creation, judgment, and problem solving	Remaining open, confident, and willing to explore in the face of new knowledge, technologies, and environments, and actively discovering and using external resources that can support understanding and action

Dimension	Core Question	Manifestation in AI Contexts	Manifestation in Social Contexts
Responsibility judgment	What action will follow from the conclusion, what consequences may result, and who should bear responsibility?	Deciding whether to adopt, review, or reject AI results according to task risk, and clarifying decision authority and responsibility within the human–AI division of labor	Distinguishing among discussion, learning, publication, medical, legal, financial, and public-policy contexts, and judging the consequences of action and the corresponding responsibilities

3.3 Individual Manifestations of Cognitive Autonomy

Cognitive autonomy is manifested first in the active regulation that individuals continually undertake within concrete cognitive tasks. The six dimensions are not six separate skills. Together, they run through the entire process of posing questions, using external resources, and verifying results. The development of an individual's capacity for cognitive autonomy is manifested in whether that person can initiate the process actively, sustain it effectively, and adjust it in a timely manner.

Posing questions expresses the individual's initiation and direction of cognitive activity. Learners need to begin with real needs and their own aims, decide what is worth understanding, solving, or creating, and transform an ill-defined situation into a question that can be explored. Artificial intelligence can help identify clues, elaborate questions, and generate candidate directions; cognitive agency makes people willing to enlist these capabilities. The choice and formulation of the question, together with its value and significance, must remain under human direction.

Using external resources expresses the individual's active organization of cognitive dependence. In writing, learning, programming, translation, design, and research, artificial intelligence can organize materials, recognize patterns, perform computational inference, generate proposals, simulate comparisons, and check errors. Learners need to decide whether to use AI, which parts of a task to assign to the system, and how to do so in light of the nature of the problem, the capabilities of the tool, and the risks of the task. Throughout this process, they must maintain source vigilance, evidence sensitivity, and boundary awareness. Human direction does not depend on personally completing every step; it depends on whether one's own questions and aims continue to coordinate the human–AI division of labor.

Verifying results expresses the individual's scrutiny of the cognitive process and exercise of final judgment. Learners need to distinguish facts, explanations, evaluations, and recommendations; ask about a conclusion's evidential basis, conditions of validity, and omitted factors; and calibrate trust according to the context of use. Low-risk, exploratory, and creative tasks can accommodate

greater uncertainty. High-risk, professional, and public tasks require stricter evidence, human review, and arrangements for responsibility. Lee and See (2004) emphasizes that appropriate trust is more important than either blind trust or complete distrust. This principle should occupy an important place in general education for an intelligent society.

Cognitive autonomy also requires individuals to remain aware of how external tools affect their own cognitive states. Artificial intelligence may increase efficiency, yet diminish patience; stimulate ideas, yet make expression more uniform; aid understanding, yet conceal its absence; or provide multiple perspectives, yet reinforce existing preferences. Learners should be able to explain how they arrived at a judgment, where AI assisted them, which content they verified, and what remains uncertain, and then adjust their subsequent use of the tool accordingly.

These demands can be expressed as a stable set of questions: What problem am I actually trying to solve? Which of my capabilities can AI extend? Which parts of the task should I assign to it? Who, or what system, produced the information? What is the evidence? Under what conditions does the conclusion hold? How much trust should I place in it? What consequences might follow if I act on it? Sustained practice in asking these questions is essential if Artificial Intelligence General Education is to move from the acquisition of knowledge to the cultivation of mind.

3.4 The Social Significance of Cognitive Autonomy

Cognitive autonomy also has a clear social significance. Judgments in an intelligent society rarely affect only the individual who makes them. Algorithmic recommendation shapes public discussion, AI-generated content influences the circulation of knowledge, automated assessment affects educational opportunity, risk models influence the allocation of medical resources, and credit scores affect access to financial services. At the same time, artificial intelligence is lowering barriers to knowledge acquisition, content creation, technological development, and professional collaboration. Whether individuals can use artificial intelligence actively and judiciously is therefore closely connected to public institutions, organizational responsibility, social trust, and opportunities to participate in new forms of learning and production.

First, cognitive autonomy bears on the quality of public discussion. Research on false news by D. M. J. Lazer et al. (2018) emphasizes that misinformation results from the interaction of technology, cognition, social motives, and platform structures. The study of diffusion on Twitter by Vosoughi, Roy, and Aral (2018) finds that false news can travel faster and farther than true news. As generative AI lowers the cost of producing content, participation in public discussion demands greater source vigilance and evidence sensitivity. If citizens cannot judge how content was produced or what evidence supports it, public reason will be weakened by fluent narratives, emotional appeals, and platform amplification.

Second, cognitive autonomy bears on the quality of organizational decision-making. As automated systems enter recruitment, credit, medicine, educational assessment, and public services, members of organizations need to understand the boundaries within which model outputs apply. Research on the use of automation by Parasuraman and Riley (1997) shows that

automation can be misused, abused, or disused. Excessive organizational reliance on a system can lead to the outsourcing of responsibility, while wholesale rejection can forfeit improvements in efficiency and quality. Cognitive autonomy requires professionals to retain the ability to interpret, review, and assume responsibility within organizational processes.

Third, cognitive autonomy bears on the individual's capacity to participate in learning, production, and creation. Artificial intelligence can help people enter unfamiliar fields rapidly, compare multiple proposals, cross some of the expressive barriers to specialized work, and turn preliminary ideas into objects that can be tested and improved. Cognitive agency inclines individuals to discover and explore these possibilities. Source vigilance, evidence sensitivity, boundary awareness, trust calibration, and responsibility judgment ensure that this expansion of capability rests on understanding and judicious regulation. Cognitive autonomy thus both preserves the individual's position as an agent and expands the range of practices in which that individual can participate.

Cognitive autonomy also bears on educational equity. If education in an intelligent society attends only to the operation of advanced tools, well-resourced schools may gain platforms and opportunities for practice more quickly while less-resourced schools are excluded. When cognitive autonomy becomes a central aim of general education, low-cost tools, open resources, pencil-and-paper simulations, analyses of everyday situations, and human–AI collaborative tasks can all play important roles. Students need not have the most advanced equipment to learn how to pose questions, select tools, and organize tasks while understanding data bias, model boundaries, the verification of generated content, recommendation mechanisms, and responsibility judgment. This orientation helps shift education away from disparities in equipment and toward the development of foundational capacities in which all learners can participate.

Finally, cognitive autonomy bears on human dignity and agency. The AI4People principles proposed by Floridi et al. (2018) emphasize beneficence, non-maleficence, autonomy, justice, and explicability, situating AI ethics within the broader context of human well-being and social governance. From the perspective of sociotechnical systems, Selbst et al. (2019) criticizes abstract conceptions of fairness and shows that algorithmic systems are always embedded in institutional and social settings. Taken together, these studies suggest that general education in an intelligent society must enable students to understand the relations among technology, human beings, institutions, and society. Cognitive autonomy enables individuals to retain their standing as agents within a technological environment while recognizing that individual judgments must enter a structure of public responsibility.

Taken as a whole, preserving citizens' capacity for cognitive autonomy bears directly on the foundations of democratic society. Democratic society depends on citizens capable of independent judgment, participation in public discussion, and responsibility for action. Public deliberation, collective decision-making, and institutional life possess a genuine foundation in human agency only when citizens can form beliefs autonomously, scrutinize evidence, calibrate trust, and recognize the boundaries of their judgments. If citizens' understanding and judgment become increasingly governed by algorithmic filtering, generated content, and automated systems, and if their capacity for cognitive autonomy is continually weakened, the civic foundation sustaining

social life will gradually erode. Preserving cognitive autonomy therefore preserves both the individual's direction over understanding and action and the basic order on which public life and the normal functioning of society depend.

3.5 Relationship to Related Educational Concepts

Cognitive autonomy is not a new slogan created from nothing. It is closely related to established concepts such as critical thinking, metacognition, self-regulated learning, epistemic cognition, information literacy, and media literacy. This book uses the concept of cognitive autonomy to reorganize these established capacities in relation to the shared conditions of intelligent society.

Critical thinking emphasizes reasoned judgment. The Delphi Report led by Facione (1990) identifies interpretation, analysis, evaluation, inference, explanation, and self-regulation as core skills of critical thinking and emphasizes dispositions toward rationality, open-mindedness, and prudence. Kuhn (1999) likewise regards argument, evidence, and reflective judgment as central to the development of critical thinking. Cognitive autonomy inherits critical thinking's orientation toward reasons while giving greater prominence to the new conditions of judgment created by model mediation, platform dissemination, and automated decision-making.

Metacognition concerns an individual's knowledge and monitoring of their own cognitive processes. In an early formulation, Flavell (1979) describes metacognition as encompassing knowledge about cognitive tasks, strategies, and one's own capabilities, together with the monitoring of cognitive processes. The framework developed by Nelson and Narens (1990) distinguishes two directions in metacognition: monitoring and control. Cognitive autonomy extends metacognition into human–AI environments. Learners need to monitor whether they understand an AI answer, whether fluent expression has swayed them, whether they have become overdependent on a tool, and whether they have assigned excessive confidence under conditions of uncertainty.

Self-regulated learning emphasizes the learner's active planning, monitoring, and adjustment of learning in pursuit of goals. Zimmerman (2002) describes it as a cyclical process comprising forethought, performance, and self-reflection. Cognitive autonomy retains this goal-directed, cyclical regulation while extending its scope from the learning process to information acquisition, belief formation, tool use, judgment, choice, and responsibility for action.

Epistemic cognition concerns how learners understand the nature, sources, certainty, and justification of knowledge. Hofer and Paul R. Pintrich (1997) argues that students' beliefs about knowledge affect learning. Sandoval, Greene, and Bråten (2016) further maintains that research on epistemic cognition needs to examine how learners reason about sources of knowledge and evidential justification in different contexts. Cognitive autonomy carries epistemic cognition into the settings of intelligent society: sources of knowledge may include people, institutions, models, platforms, or hybrid systems, while the justification of knowledge may involve data, algorithms, retrieved evidence, model evaluation, and human judgment.

Information literacy and media literacy offer another perspective. Information literacy emphasizes identifying information needs, gaining access to information, evaluating it, and using

it ethically. Media literacy emphasizes understanding media production, symbolic representation, structures of dissemination, and relations of power. Intelligent society further incorporates data, algorithms, generative models, and automated assessment into the information environment. Students need to understand both the source and quality of information and the ways in which these technologies and systems alter its production, presentation, and dissemination.

Table 3.2: Cognitive Autonomy and Related Educational Concepts

Related Concept	**Central Concern**	**Contribution to Cognitive Autonomy**	**Further Extension Through Cognitive Autonomy**
Critical thinking	Reasons, evidence, argument, inference, and evaluation	Helps students analyze the quality of evidence in AI outputs and social information	Situates reasoned judgment within environments of model mediation and platform dissemination
Metacognition	Monitoring and regulating one's own cognitive processes	Helps students notice their degree of understanding, strength of confidence, and dependence on tools	Extends self-monitoring to processes of human–AI collaborative learning
Self-regulated learning	Planning, monitoring, adjusting, and reflecting on learning in pursuit of goals	Helps students organize learning strategies actively and improve continually in light of feedback	Extends cyclical regulation to beliefs, tools, judgment, and action
Epistemic cognition	The sources, certainty, and justification of knowledge	Helps students understand how different knowledge claims acquire justification	Incorporates people, institutions, models, platforms, and hybrid sources into judgments about knowledge

Related Concept	Central Concern	Contribution to Cognitive Autonomy	Further Extension Through Cognitive Autonomy
Information and media literacy	Information access, media production, structures of dissemination, and ethical use	Helps students judge information sources, forms of expression, and mechanisms of dissemination	Incorporates data, algorithms, generative models, and automated assessment into analysis of the information environment

This comparison shows that cognitive autonomy reorganizes established educational concepts within the shared conditions of intelligent society. Students need critical thinking to analyze reasons; metacognition and self-regulated learning to monitor and adjust their own activity; epistemic cognition to understand the sources and justification of knowledge claims; and information and media literacy to evaluate information and its mechanisms of dissemination. Cognitive autonomy asks, further, whether these capacities can work together in tool selection, human–AI collaboration, everyday judgment, and real-world action, forming a continuous process of active regulation.

3.6 Possible Objections and Responses

Establishing cognitive autonomy as the contemporary mission of general education in an intelligent society may invite three kinds of objection. They concern the operationalization of the aim, the need for the concept, and its underlying attitude toward technology.

The first objection is that the aim is too abstract to be implemented. Responding to this objection requires observing the process of cognitive autonomy through the six dimensions of source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment, and cultivating it through AI cases, social cases, authentic tasks, human–AI collaboration, cross-context tasks, and reflective records. Subsequent parts of this book establish this operational basis more fully through accounts of capacity structure, curricular content, pedagogical implementation, and learning assessment.

The second objection is that established concepts such as critical thinking, metacognition, and information literacy already cover the relevant educational aims, leaving no need to introduce cognitive autonomy. Cognitive autonomy absorbs the central achievements of these concepts while responding to structural changes in the cognitive environment of intelligent society. Sources of information have expanded from people and institutions to include models, platforms, and hybrid human–machine systems. Cognitive activity now extends beyond the individual's mind into distributed processes constituted jointly by data, algorithms, tools, and organizations. Model outputs may also enter real-world action directly. By taking active human regulation as its

main thread, cognitive autonomy organizes source, evidence, boundaries, trust, agency, and responsibility into a continuous process. It thereby provides a shared problem context and an integrating aim for capacities that originated in different traditions of research.

The third objection is that an emphasis on sources, evidence, boundaries, trust, and responsibility may turn cognitive autonomy into risk-prevention education and leave learners defensive toward artificial intelligence. Cognitive autonomy, however, extends beyond the communication of risk. It seeks a mature understanding of technology and the cognitive agency to engage actively with it, explore its possibilities, and extend one's own capabilities. On the basis of an understanding of how AI capabilities arise and where they apply, learners can determine which tasks suit the technology and use it actively for learning, creation, and problem solving while maintaining evidence sensitivity, boundary awareness, and responsibility judgment. Understanding mechanisms and boundaries creates well-grounded confidence. It enables people to avoid both indiscriminate rejection and uncritical dependence and to develop mature ways of acting that combine active use, judicious trust, and human intervention when necessary.

3.7 Chapter Summary

Intelligent society has transformed the fundamental conditions under which information is generated, knowledge is organized, judgments are formed, and actions are carried out. It has thereby given general education a new contemporary task. Cognitive autonomy is the process by which individuals actively regulate the formation of their beliefs, their use of evidence, their allocation of trust, the boundaries of their judgment, and their responsibility for action within a complex cognitive environment. The capacity for cognitive autonomy is the ability to initiate, sustain, and adjust this process across different tasks and contexts. Distinguishing the two enables general education to attend both to cognitive activity as it actually occurs and to the individual capacity that allows this activity to be sustained reliably and transferred across contexts.

Cognitive autonomy is grounded in open dependence. Individuals may selectively draw on expert knowledge, institutional resources, and AI capabilities in retrieval, computation, generation, and analysis; they need not perform all cognitive activity independently. At the same time, they must retain command of the initiation, scrutiny, and coordination of the cognitive process, especially the roles of posing questions and verifying results. Source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment form the six-dimensional structure of cognitive autonomy. Source vigilance, evidence sensitivity, boundary awareness, trust calibration, and responsibility judgment correspond respectively to belief formation, the use of evidence, the boundaries of judgment, the allocation of trust, and responsibility for action. Cognitive agency supplies the inner motivation to engage actively with technological change, explore cognitive possibilities, and extend one's own capabilities. Together, the six dimensions express what it means to retain human direction while using external capabilities and provide an analytical framework for curriculum design, classroom observation, and process-oriented diagnosis.

Cognitive autonomy has both individual and public significance. It helps individuals organize

external cognitive resources actively, verify results, and recognize how tools affect their own cognitive states. It also bears on public discussion, organizational decision-making, participation in learning and production, educational equity, and human dignity and agency. Critical thinking, metacognition, self-regulated learning, epistemic cognition, information literacy, and media literacy provide important foundations for this aim. Cognitive autonomy situates these traditions within the shared environment of model mediation, platform dissemination, human–AI collaboration, and automated decision-making, bringing them together in a continuous process of active regulation.

Cognitive autonomy has now been established as the contemporary mission of general education in an intelligent society. The next chapter turns to Artificial Intelligence General Education itself. It examines how this form of education has taken shape, which distortions of purpose have appeared in current practice, and how its educational orientation can be recalibrated around cognitive autonomy. Explaining why Artificial Intelligence General Education can and should assume the task of developing the capacity for cognitive autonomy requires a further examination of the scientific character and modes of thought of artificial intelligence and of the fundamental mechanisms by which AI enters social life. Part II develops that argument.

Chapter 4

Artificial Intelligence General Education: Positioning, Significance, and Deviations in Practice

Abstract. The preceding chapter established cognitive autonomy as the defining mission of general education in an intelligent society. This chapter further defines the form of education capable of assuming that mission, which is the main topic of this book: Artificial Intelligence General Education. Artificial Intelligence General Education is an educational form created through the integration of general education and artificial intelligence education. It takes the knowledge, methods, systems, and social operation of artificial intelligence as its subject matter and assumes general education's mission of developing the capacity for cognitive autonomy in an intelligent society. In current practice, the conceptual boundaries among artificial intelligence education, Artificial Intelligence General Education, and AI literacy remain unclear. Partial forms of education—including tool use, technical training, project competitions, frontier topics, ethical reminders, and the assessment of finished products—can consequently come to stand for Artificial Intelligence General Education as a whole. This chapter analyzes the resulting tool-centered, technology-centered, activity- and competition-centered, frontier-chasing and fragmented, ethics-as-add-on, assessment-superficial, and teacher-training-superficial tendencies. It also examines the cognitive turn emerging in domestic and international policies and competency frameworks. Clarifying the relationships among educational positioning, learning outcomes, and the overarching goal is a prerequisite for Artificial Intelligence General Education to advance from the expansion of content to a conscious understanding of its aims.

Keywords: Artificial Intelligence General Education; artificial intelligence education; AI literacy; cognitive autonomy; educational positioning; deviations in practice; cognitive turn

4.1 Defining Artificial Intelligence General Education

Beginning with the cognitive environment of an intelligent society, the preceding chapter established cognitive autonomy as the defining mission of general education. Once that goal has been defined, a further question must be answered: what form of education can assume this mission?

Artificial Intelligence General Education acquires its educational position in response to this question.

At an intuitive level, many of the cognitive predicaments of an intelligent society arise from the involvement of artificial intelligence. Studying artificial intelligence can help learners understand the basic principles and modes of operation of AI systems and thereby develop a foundational understanding of intelligent society, which is a prerequisite for cognitive autonomy.[1] General education and artificial intelligence education converge at this point, forming the two sources of Artificial Intelligence General Education.

From the perspective of general education, developing the capacity for cognitive autonomy is a historical mission it must assume, while the study of artificial intelligence provides a viable path for accomplishing that mission. Artificial Intelligence General Education is therefore a distinctive form of general education: it develops the capacity for cognitive autonomy through the teaching and study of artificial intelligence.

From the perspective of artificial intelligence education, Artificial Intelligence General Education is a distinctive form within the broader field. Artificial intelligence education is the superordinate concept encompassing all forms of education organized around artificial intelligence. It includes the knowledge, methods, systems, tools, applications, and social impact of AI, as well as educational forms directed toward professional preparation, career development, and general learners. In most contexts, it places greater emphasis on technical learning and skill development and does not take cognitive autonomy as its primary aim. Artificial Intelligence General Education is AI education directed by the aims of general education. It is a distinctive form of AI education constructed to fulfill general education's goal of developing the capacity for cognitive autonomy.

Drawing these perspectives together, this book defines Artificial Intelligence General Education as follows:

Artificial Intelligence General Education is an educational form created through the integration of general education and artificial intelligence education. It takes the knowledge, methods, systems, and social operation of artificial intelligence as its subject matter and assumes general education's mission of developing the capacity for cognitive autonomy in an intelligent society.

This integration proceeds in two interrelated directions. General education supplies Artificial Intelligence General Education with its educational mission and value aims, directing the study of AI toward all members of society and toward human intellectual development, value judgment, and public life. Artificial intelligence education supplies direct objects of study, bodies of knowledge, and forms of practice through which learners can understand how AI capabilities are formed, how AI systems operate, and how artificial intelligence enters information dissemination,

[1]The statement that "many of the cognitive predicaments of an intelligent society arise from the involvement of artificial intelligence" is only a broad intuition. Demonstrating that the study of artificial intelligence can develop the capacity for cognitive autonomy requires a deeper analysis of the characteristics and modes of thought of the discipline, as well as of the ways in which artificial intelligence enters social activity. These questions are addressed in Part II.

knowledge production, social organization, and real-world action. Without the mission of general education, AI education can contract into professional knowledge, tool operation, or vocational skill. Without systematic study of artificial intelligence, the cognitive problems of intelligent society can remain at the level of general admonitions. Their integration produces Artificial Intelligence General Education as a complete educational form. This integration does not permit Artificial Intelligence General Education to be subsumed under either source alone. It cannot be reduced to a simplified version of professional AI education or an enlarged version of traditional general education with several AI topics added. It is an independent educational form created by integrating the two.

Cognitive autonomy provides the integrating principle. Studying artificial intelligence allows learners to understand such components of a system as data, models, objectives, evaluation, generalization, error, and feedback and thus to recognize both the sources and the limits of AI capabilities. Analyzing the social operation of artificial intelligence allows them to see how platforms, organizations, and institutions use it to participate in information filtering, judgment formation, and the implementation of action. Such learning supports learners in actively using artificial intelligence to extend their own capabilities while training them to question sources, analyze evidence, recognize boundaries, calibrate trust, and assume responsibility. Cognitive autonomy therefore constitutes the fundamental aim of Artificial Intelligence General Education and organizes knowledge acquisition, capacity development, and responsibility into an integrated whole.

4.2 AI Literacy, Cognitive Autonomy, and the Capacity for Cognitive Autonomy

Artificial Intelligence General Education must also be distinguished from AI literacy, cognitive autonomy, and the capacity for cognitive autonomy. Existing policies and competency frameworks commonly describe AI literacy in terms of understanding artificial intelligence, using it, evaluating its outputs, applying it creatively, and addressing ethics and responsibility(Miao, Shiohira, and Lao, 2024; Teaching Steering Committee for Basic Education, Ministry of Education, 2025a; OECD and European Commission, 2026). On this basis, this book understands AI literacy as the integrated learning outcomes—including knowledge, skills, attitudes, and responsibility—that learners develop in AI-related contexts. It answers what learners have acquired, through AI education, in the domains of understanding, use, evaluation, and responsible participation.

Cognitive autonomy addresses a different level of the problem: whether individuals can actively regulate their own belief formation, use of evidence, allocation of trust, boundaries of judgment, and responsibility for action in a complex cognitive environment. Cognitive autonomy is an ongoing cognitive process and the overarching goal of Artificial Intelligence General Education. The capacity for cognitive autonomy is the capacity to initiate, sustain, and adjust this process so that it can continue across different tasks and contexts.

The respective positions of the four concepts in this book are summarized in Table 4.1.

Table 4.1: Levels and relationships among concepts related to Artificial Intelligence General Education

Concept	**Conceptual nature**	**Core content**	**Position in this book**
Artificial Intelligence General Education	Educational form and educational process	Organizes the study of AI knowledge, methods, applications, and social impact for all members of society	Integrates general education and AI education and assumes the mission of general education in an intelligent society
AI literacy	Direct learning outcomes in AI-related contexts	Knowledge, skills, attitudes, and responsibility required to understand, use, evaluate, and participate responsibly in artificial intelligence	Represents the integrated literacy learners have developed in the domain of artificial intelligence
Cognitive autonomy	Overarching educational goal and process of active regulation	Active regulation of belief formation, use of evidence, allocation of trust, boundaries of judgment, and responsibility for action	Provides a common direction for curricular content, learning activities, and assessment evidence
Capacity for cognitive autonomy	Stable capacity that can be developed and transferred	Initiates, sustains, and adjusts the process of cognitive autonomy so that it can continue across tasks and contexts	Translates cognitive autonomy into a capacity that can be cultivated, observed, and assessed

AI literacy and cognitive autonomy are closely connected, although their scope and function differ. AI literacy primarily describes learners' direct learning outcomes in AI-related contexts. Cognitive autonomy extends across AI settings as well as more general forms of information judgment, knowledge acquisition, social participation, and choice of action. Artificial Intelligence General Education must cultivate substantive AI literacy and further enable learners to transfer the awareness of sources, evidence, and boundaries, the calibration of trust, and judgments of responsibility developed through the study of AI to broader social contexts. Part II explains how the scientific character of artificial intelligence and the ways in which it enters social activity

provide the foundation for this development of capacity and transfer across contexts.

4.3 Further Discussion of Artificial Intelligence General Education

Having defined the central concepts, we must further clarify the relationships between Artificial Intelligence General Education and general education, professional AI education, and information technology education. These relationships directly shape the curricular positioning, selection of content, and forms of implementation of Artificial Intelligence General Education. It inherits the historical mission of general education while responding to new problems posed by intelligent society. It draws its foundation of knowledge from the discipline of artificial intelligence while pursuing educational aims that differ from those of professional and information technology education. Only by clarifying these relationships can Artificial Intelligence General Education develop into a stable and complete educational form.

4.3.1 The Dual Social Significance of Artificial Intelligence General Education: A Common Foundation and the Human as Cognitive Subject

General education has long assumed the task of building a common social foundation. The differentiation of knowledge leads members of society to master different fields; the division of labor places them within different institutions and practices; and the diversity of experience and value commitments makes public problems difficult to understand through any single tradition of knowledge. General education must therefore help members of society develop basic common knowledge, modes of thought, capacities for expression, value awareness, and public responsibility. These resources allow people from different backgrounds to understand one another, exchange reasons, participate in public discussion, and address social affairs together. In *General Education in a Free Society*, the Harvard Committee connected a common culture and common citizenship to the preservation of a free society, emphasizing the foundational role of general education in public life(Harvard University Committee on the Objectives of a General Education in a Free Society, 1945).

Artificial Intelligence General Education first inherits this traditional task. Artificial intelligence is entering knowledge production, information dissemination, public discussion, organizational management, and social decision-making, gradually becoming a cognitive infrastructure shared by people across society. Whether or not individuals work professionally in artificial intelligence, they need a basic understanding of data, models, algorithmic recommendation, generated content, and automated decision-making. Without this common knowledge, people cannot readily discuss the capabilities and risks of AI systems or engage effectively in public dialogue about the boundaries of their use, the allocation of responsibility, and appropriate institutional arrangements. Basic knowledge of artificial intelligence and the modes of judgment associated with it are therefore becoming important components of the common structure of knowledge in an intelligent society.

Artificial Intelligence General Education also assumes a distinctive task in an intelligent society: sustaining the human being as a cognitive subject. The influence of artificial intelligence has extended beyond the ways in which people obtain information and use tools. It has begun to enter such cognitive activities as posing questions, selecting materials, organizing reasons, forming judgments, and implementing action. Individuals can use artificial intelligence to extend their capacities to learn, create, and act. Through sustained dependence, however, they may also gradually lose awareness and regulation of their own cognitive processes. Artificial Intelligence General Education must enable learners to understand how AI produces results, how those results are shaped by data and objectives, and where the boundaries of system capabilities lie. On this basis, learners must learn to calibrate trust, verify evidence, organize the division of labor between human beings and AI, and assume responsibility for action. The UNESCO AI Competency Framework for Students likewise places human agency, responsibility, citizenship, AI ethics, technical understanding, and system design within a single framework, indicating that the public significance of AI education now extends beyond technical operation(Miao, Shiohira, and Lao, 2024).

Both forms of significance bear on the stable operation of civil society. Common knowledge and a public language enable members of society to enter into dialogue, while cognitive autonomy ensures that those who participate in dialogue and make decisions retain the capacity for independent judgment. If citizens cannot understand the information environment mediated by artificial intelligence, public discussion loses an essential common foundation. If citizens gradually hand over information selection, the formation of reasons, and final judgment to AI systems, public life will also lose its foundation in citizens as independent subjects. Democratic institutions may continue to operate formally while the citizens whose agency sustains them are progressively hollowed out. Building a common foundation and sustaining citizens as cognitive subjects are therefore interdependent, and together they constitute the public mission of general education in an intelligent society.

In this sense, Artificial Intelligence General Education is a new form of general education for an intelligent society. It continues general education's historical task of developing common knowledge, public judgment, and social responsibility, while extending its field of concern to a cognitive world jointly mediated by data, models, platforms, and AI systems. It also advances the cultivation of an independent mind toward cognitive direction under conditions of human–AI collaboration, enabling learners to maintain active regulation over beliefs, evidence, trust, boundaries, and responsibility amid open dependence. Artificial Intelligence General Education thereby extends the boundaries of the content, capacities, and modes of realization of traditional general education.

4.3.2 Artificial Intelligence General Education as Distinct from a Simplified or Popularized Form of Professional Education

Artificial Intelligence General Education cannot be understood as a simplified or popularized form of professional AI education. Although both take the concepts, knowledge, methods, and systems of artificial intelligence as objects of study, they differ in their intended learners, educational

aims, organization of content, and expected learning outcomes.

Professional AI education primarily prepares specialists who will undertake research, development, and engineering practice in artificial intelligence. It must establish systematic and advanced foundations in mathematics, computation, and engineering, enabling learners to understand algorithmic principles and to carry out model design, system implementation, experimental evaluation, and technological innovation. Depth of learning, technical completeness, and professional practical competence are central requirements of professional education.

Artificial Intelligence General Education addresses every member of society. Here, knowledge of artificial intelligence is both content to be understood and a major vehicle for developing the capacity for cognitive autonomy. Through data labeling, learners examine how facts are transformed into computable objects. Through machine learning, they examine how conclusions are induced from limited experience. Through model testing, they examine how established regularities transfer to new contexts. Through generative models, they examine the difference between fluent output and reliable knowledge. Through human–AI collaboration, they examine the relationships among tool use, the allocation of trust, and responsibility for action. Professional content from artificial intelligence thus enters a broader process of cognitive training that supports the development of the capacities for induction, generalization, judgment, and self-awareness.

The same AI concept can perform different functions in the two forms of education. In studying classification models, professional education may require learners to derive an objective function, implement a training algorithm, and compare model performance. General education can approach classification models through sample selection, the formation of labels, the distribution of errors, and the consequences of application, allowing learners to understand how data shape conclusions and why models have boundaries of applicability. Professional study of generative models emphasizes model architectures, training methods, inference efficiency, and system optimization. General education focuses on generative mechanisms, uncertainty in outputs, verification of information, calibration of trust, and responsibility in use. Differences in difficulty and depth are only the most visible distinction. The fundamental difference lies in the purpose of learning and the organization of knowledge.

General education and professional education pursue their own educational aims. They develop in parallel and support one another, without forming a hierarchy or a fixed sequence. Artificial Intelligence General Education is more than preparation for professional study, while professional education does not automatically encompass every task of general education. AI professionals also need to develop cognitive autonomy, public responsibility, and social judgment; learners in other fields and members of the public likewise need to understand artificial intelligence and retain cognitive direction in an intelligent society.

The two forms of education also perform different social functions. Professional education prepares specialists capable of advancing AI research and technological development. Artificial Intelligence General Education addresses a much broader public and sustains the common knowledge, basic capacity for judgment, and citizen agency required by an intelligent society. Professional education is irreplaceable in the depth of specialized preparation it provides. Artificial Intelligence General Education has a broader public and foundational role because it reaches

across society and sustains common cognitive conditions. Its social value does not derive from a higher level of knowledge; it derives from addressing every member of society and from its direct bearing on whether an intelligent society will possess citizens capable of understanding, judging, and taking responsibility.

4.3.3 Artificial Intelligence General Education as Distinct from an Extension of Information Technology Education

Information technology education provides an important foundation for Artificial Intelligence General Education. Data processing, algorithms, programming, networks, digital tools, and responsibility in the information society are all closely connected with the study of artificial intelligence. China's compulsory-education information technology curriculum has established a curricular system centered on information awareness, computational thinking, digital learning and innovation, and responsibility in the information society, and it has gradually incorporated content related to algorithms and artificial intelligence(Ministry of Education of the People's Republic of China, 2022). The inclusion of AI knowledge can update the content of information technology courses and help students understand a new stage in the development of digital technology.

This connection does not place Artificial Intelligence General Education within information technology education. Through its long development, artificial intelligence has established a relatively independent object of study, set of fundamental problems, conceptual system, genealogy of methods, and modes of evaluation. It asks how intelligent activity can be represented, computed, and realized, and it investigates how machines perceive their environment, represent knowledge, learn from data, reason and search, make decisions, and interact with people and the environment. Representation, learning, generalization, reasoning, search, optimization, decision-making, and feedback together constitute the disciplinary structure of artificial intelligence. Chapter 5 explains further how artificial intelligence has developed a complete scientific system extending from foundational theories and core mechanisms to system applications.

Information technology education is concerned with the acquisition, representation, transmission, processing, communication, and application of information. It emphasizes individuals' capacity to use digital technologies to solve problems and participate in the information society. Artificial intelligence and information technology share such foundations as data, algorithms, and computational systems, but they organize these foundations in different directions. Artificial intelligence asks further how intelligent activity is transformed into a computational object, how models form regularities from experience, why the resulting regularities can or cannot apply to new objects, and how systems produce outputs and actions in uncertain environments. These questions express the distinctive object and modes of thought of artificial intelligence. Treating AI entirely as a local application of information technology can compress its disciplinary structure into a collection of tools, algorithms, or operational activities, weakening students' systematic understanding of its fundamental ideas and mechanisms.

In terms of educational aims, Artificial Intelligence General Education assumes the distinctive mission of developing the capacity for cognitive autonomy. Artificial intelligence possesses the

dual character of a cognitive mirror and a social mediator. On one side, it transforms intelligent activities such as representation, learning, induction, generalization, judgment, and decision-making into computational processes that can be observed and analyzed. On the other, AI systems have entered information dissemination, knowledge production, organizational evaluation, and decisions about action in society. Studying artificial intelligence therefore enables learners to observe both machine cognition and their own cognitive processes and to understand the effects produced when those computational processes enter society. This educational mission requires teaching to be organized around sources, evidence, boundaries, trust, cognitive agency, and responsibility. It also calls for learning tasks such as comparing models, testing boundaries, verifying outputs, organizing the human–AI division of labor, and making judgments in social contexts. The aims and content of conventional information technology education cannot fully encompass these requirements.

In classroom practice, information technology courses already possess a relatively complete content structure, allocation of instructional time, and learning progression. When curricular structures and total instructional time remain unchanged, AI content can generally enter these courses only as individual units, cases, or extension activities. Such content can enrich information technology education and provide students with preliminary experiences of artificial intelligence. It cannot readily support a complete developmental process extending from the basic mechanisms and disciplinary thought of AI to its social impact and cognitive autonomy. Fragmented introductions to generative tools, model applications, or simple programming likewise cannot establish a continuous progression of capacities and assessment.

The organization of Artificial Intelligence General Education in schools can remain open. It may be offered as an independent course or implemented in coordination with information technology, science, general technology, social science, and integrated practical activities. The *Guidelines for Artificial Intelligence General Education in Primary and Secondary Schools (2025 Edition)*, issued by the Ministry of Education's Steering Committee for Basic Education and Teaching, set out an AI literacy goal integrating knowledge, skills, ways of thinking, and values. Implementation plans in some regions also permit both independent courses and integration with existing curricula(Teaching Steering Committee for Basic Education, Ministry of Education, 2025a; Beijing Municipal Education Commission, 2025). The administrative vehicle used for a course can be determined flexibly according to instructional time, teacher availability, and school conditions. The system of aims, content structure, developmental progression, and assessment requirements of Artificial Intelligence General Education must nonetheless be fully realized.

Information technology education can therefore absorb AI knowledge in response to the development of digital technology, while Artificial Intelligence General Education can draw on the teachers, equipment, and curricular foundations of information technology. Each, however, has its own disciplinary content, modes of thought, and educational mission. AI content in information technology courses can become a component of Artificial Intelligence General Education, but a limited extension of content cannot independently fulfill the defining mission of sustaining cognitive autonomy and citizen agency.

4.3.4 The Independent Educational Standing of Artificial Intelligence General Education

The preceding discussion shows that the educational character of Artificial Intelligence General Education arises from three sources. Its overarching goal is cognitive autonomy, and it therefore assumes the defining mission of general education in an intelligent society. Its content is grounded in the scientific system of artificial intelligence, which gives it an independent and complete source of knowledge. It addresses all members of society and thus differs from professional AI education directed toward research and technological development. Although it can share content and conditions of implementation with information technology and other courses, these curricular connections cannot dissolve its independent educational aims.

The independent standing of Artificial Intelligence General Education does not require every school to establish immediately a uniform independent course title, nor does it exclude interdisciplinary integration. What must be preserved is the integrity of its educational mission: whether courses enable learners to understand the basic mechanisms and disciplinary thought of artificial intelligence; recognize the processes and effects of its entry into society; develop the capacities for induction, generalization, judgment, and self-awareness through studying and using AI; and actively regulate beliefs, evidence, trust, boundaries, and responsibility amid open dependence. Only when these aims are systematically realized can AI education move beyond the dissemination of knowledge and training in tool use and become a foundational education for cognitive autonomy in an intelligent society.

4.4 Current Misalignments in Artificial Intelligence General Education

Artificial Intelligence General Education is still at an early stage, and some of its basic concepts have yet to be adequately distinguished in educational practice. Professional education, skills training, interest-based activities, and general education often employ similar course titles. Partial educational aims are then carried into courses for all learners and gradually displace the more comprehensive aims of general education. Many of the deviations discussed below arise from this conceptual confusion in practice.

In policy communication and everyday usage, “artificial intelligence education” often serves as a field-level term covering a wide range of educational activities, including professional education, skills training, interest-based activities, and courses for the general public. This broad usage is useful for describing the field as a whole, but it does not directly establish the educational character of a particular course. The problem emerges in concrete curricular design when practices drawn from professional preparation, skills training, or interest-based activities are directly interpreted and promoted as Artificial Intelligence General Education and disseminated at scale under that name.

One form of confusion equates the object of study with the nature of the education. Any course that involves algorithms, programming, robotics, generative AI, or AI ethics may be classified as the same kind of education and promoted as Artificial Intelligence General Education.

Yet the same content can serve different purposes. Algorithms and programming can develop professional implementation skills, or they can help general learners understand how models produce results. Generative AI tools can be used for occupational skills training, or they can provide general educational material for comparing outputs, verifying evidence, and reflecting on patterns of dependence. What a course teaches identifies its object of study. How it organizes that content and what capacities it expects learners to develop determine its educational character.

A second form of confusion equates a universal audience with general educational aims. Offering a course to all students is an important condition of Artificial Intelligence General Education, but offering every student a tools course, a programming course, or a one-off experiential activity does not automatically produce general education. The defining issue is whether a course connects AI knowledge with common knowledge, intellectual capacities, value awareness, and public responsibility; whether it enables learners from different backgrounds to understand and actively use artificial intelligence; and whether it enables them to regulate their use of evidence, allocation of trust, boundaries of judgment, and responsibility for action in complex contexts.

A third form substitutes course names and activity formats for judgment of aims. A course called "Artificial Intelligence General Education" may retain the structure of a compressed professional course. Science festivals, robotics competitions, and innovation projects may be presented directly as the outcomes of a school's general education. A course containing a unit on ethics may still be dominated by tool use and product creation. Names, audiences, and formats provide preliminary indications, but determining whether a practice realizes Artificial Intelligence General Education requires examining whether its aims, tasks, and assessment jointly point toward cognitive autonomy.

This ambiguity has concrete historical sources. AI courses in primary and secondary schools often developed from information technology, maker education, robotics activities, science competitions, or extracurricular clubs. Related university courses often grew out of professional AI courses, general computer courses, and "AI plus discipline" offerings. Social training programs more commonly begin with product use and workplace competence. These different paths have accumulated rich resources for AI education, but they also retain their original aims and habitual forms of implementation. When such content enters Artificial Intelligence General Education without being selected and reorganized around a new educational mission, professional, skills-based, activity-based, and product-display logics continue to dominate the curriculum.

This section therefore examines the misplacement of these partial forms within general education. When tool operation comes to represent the complete curriculum, a tool-centered tendency emerges. When professional technical training is transferred directly into the common foundation for all learners, a technology-centered tendency emerges. When activities and competitions come to represent the outcomes of education for everyone, activity- and competition-centered tendencies emerge. Frontier-chasing, ethics-as-add-on, superficial assessment, and superficial teacher training share the same structure. The issue is not whether the content has value. The question is whether the content and its pedagogy fulfill the comprehensive aims of Artificial Intelligence General Education.

4.4.1 The Tool-Centered Tendency: Reducing General Education to Operational Skill

The tool-centered tendency is among the most common deviations in Artificial Intelligence General Education. Since the spread of generative AI, many courses and training programs have centered on such topics as writing quickly, generating presentations, producing images with AI, designing prompts, and improving workplace efficiency. Such training produces rapid short-term results, gives learners an immediate sense of accomplishment, and responds to genuine needs in the workplace. It therefore has clear value in skills training and practical work. The deviation arises when tool operation is treated as the principal content of Artificial Intelligence General Education and the understanding of principles, verification of results, and judgment of responsibility consequently disappear from the curriculum.

Three conditions encourage this tendency. First, tool interfaces are directly visible and their instructional effects are easy to demonstrate. Demonstrating a generative tool produces immediate classroom feedback more readily than explaining data distributions, model boundaries, or the allocation of responsibility. Second, generative AI produces compelling text, images, code, and plans with great speed. Teachers and learners can easily mistake the quality of the output for the quality of learning. Third, social demands for efficiency direct some forms of AI education toward tool training, while schools and training organizations are themselves inclined to prove a course's value through visible products.

Tool training has a legitimate place, and appropriate practical experience helps students understand the impact of AI on life and learning. To become part of general education, however, tool training must serve a more profound aim. UNESCO's guidance on generative AI in education and research likewise calls on educational systems to help learners understand and use generative AI appropriately while emphasizing human agency, ethics, safety, and equity(Miao and Holmes, 2023). Tool training enters the comprehensive structure of Artificial Intelligence General Education when it leads toward understanding, judgment, and reflection.

Research in psychology and human–computer interaction helps explain the educational risks of a tool-centered curriculum. Risko and Gilbert (2016) define "cognitive offloading" as the use of external tools to reduce internal cognitive demand. This process can improve performance, while also changing how people engage with a task. Research by Sparrow, Liu, and Wegner (2011) on the "Google effect" shows that when people expect information to be stored in or retrievable from an external system, their memory for the information itself declines while their memory for its location improves. These studies do not support a simple technological pessimism. They do remind educators that external cognitive tools reorganize human attention, memory, and learning strategies. As an even more powerful cognitive tool, generative AI may further externalize problem definition, linguistic organization, evidence retrieval, and preliminary judgment.

Research on human interaction with automation provides related evidence. As early as 1997, Parasuraman and Riley (1997) distinguished the use, misuse, disuse, and abuse of automation, noting that excessive trust can lead to insufficient monitoring and inappropriate use. Research by Skitka, Mosier, and Burdick (1999) on automation bias shows that people receiving automated recommendations may disregard contrary evidence or act on erroneous advice. These findings

indicate that stronger tools create a greater educational need for monitoring, verification, and judgment of responsibility. If Artificial Intelligence General Education trains students only to obtain results from AI, it may strengthen their operational skill while weakening vigilance toward sources, evidential strength, and boundaries of applicability.

In the classroom, the tool-centered tendency often appears as "prompt centrism." Prompts are indeed an important interface between people and generative AI. Clearly defining a task, providing context, specifying a format, and requesting self-checks can improve the quality of an output. Making prompts the core of general education, however, subjects the curriculum to the interface of a particular tool. Students may learn numerous templates without being able to explain why models hallucinate. They may produce polished essays without identifying the evidence. They may ask AI to revise a plan without being able to judge whether the revision is sound. They may become adept with a particular model only to lose their advantage when the model changes.

Instruction in prompts and similar techniques can form one component of a course, provided it serves problem definition, evidence verification, comparison of outputs, and records of reflection. It cannot become the purpose of the course itself.

A simple example illustrates the distinction. A teacher asks students to use generative AI to write a short essay on "how artificial intelligence is changing education." If the task requires only a properly formatted essay, students are likely to make small revisions to an AI output and submit it. Suppose instead that they must compare outputs from three different prompts, mark unverified factual claims, locate one primary source, identify the usable and unusable parts of the AI responses, and document their own process of judgment. The same use of a tool is then transformed into cognitive training. The first task trains invocation; the second trains calibration.

Correcting the tool-centered tendency therefore requires turning the use of AI tools into a site of cognitive training. Courses can teach students to use tools, but every use should be accompanied by four questions: Where did this output come from? What evidence supports it? In what context does it apply? Why do I believe it, or withhold belief for the moment? Tool training enters the structure of Artificial Intelligence General Education only when it is subordinated to these questions.

4.4.2 The Technology-Centered Tendency: Reducing the Common Foundation to Algorithms and Programming

At the opposite pole from the tool-centered tendency lies excessive technicalization. Programming, algorithms, model training, and technical implementation are central to professional AI education and can also help general learners understand the internal mechanisms of AI systems. Many courses appropriately begin with Python, classifiers, neural networks, data labeling, model tuning, or machine-learning projects. A technology-centered deviation arises when professional technical training is transferred directly into a general curriculum for all students: thresholds, depth of content, and assessment criteria continue to follow the logic of professional preparation, crowding out the judgment, transfer, and responsibility required for understanding the relationship between

technology and society.

This tendency has historical roots. Computer science education has long centered on programming, algorithms, and data structures. Wing (2006) advanced the concept of "computational thinking," emphasizing the importance of decomposition, abstraction, algorithmic thinking, and automation for everyone. This idea stimulated the global development of computer science and programming education. Artificial intelligence education inherited the resources of computer science education and can readily inherit the path that equates learning technology with learning how to implement it.

The general educational problems created by AI now extend beyond the boundaries of traditional computational thinking. In discussing the movement "from computational thinking to AI thinking," Zeng (2013) argues that artificial intelligence involves knowledge, semantics, learning, context, and non-symbolic data and cannot be fully encompassed by conventional computational thinking. The AI4K12 initiative's Five Big Ideas likewise do not limit AI learning for primary and secondary students to implementing machine learning. They include perception, representation and reasoning, learning, natural interaction, and societal impact(Touretzky et al., 2019; AI4K12 Initiative, 2019). These frameworks show that Artificial Intelligence General Education requires technical understanding, but it must also lead from code and algorithms to systems, applications, and social impact.

Technology-centered courses often contain an implicit assumption: if students understand algorithmic principles, they will acquire a relatively complete AI literacy. This assumption must be revised. Students may train an image classifier without understanding how training data reflect social bias. They may implement a recommendation algorithm without understanding how its objective shapes attention. They may know that a neural network contains layers and parameters without being able to judge whether an AI-generated medical recommendation requires review by a physician. They may write code without knowing how to articulate boundaries of responsibility in a high-risk setting. Technical knowledge is an important foundation of Artificial Intelligence General Education, but it does not automatically produce social judgment or cognitive autonomy.

This is particularly evident in machine-learning education. Student projects frequently follow a sequence of "collect data–train model–test accuracy." This process can help students understand that models learn from data and can produce intuitive artifacts. If a course attends only to completing this process, however, students can overlook data provenance, sample representativeness, test-set design, distribution shift, types of error, and risks in application. Problems in real AI systems often arise precisely in these places. High accuracy from a cat-and-dog classifier does not mean students understand algorithmic fairness. A working speech-recognition project does not mean they understand how accents, dialects, and environmental noise create problems of data distribution.

The educational concept of pedagogical content knowledge reminds us that mastering content and teaching it are different achievements. Shulman (1986) introduced pedagogical content knowledge to emphasize that teachers must understand how disciplinary content can be transformed into teachable forms. Building on this idea, the TPACK framework proposed by Mishra and Koehler (2006) shows that technology integration requires a complex combination of content,

pedagogy, and technological knowledge. AI education particularly needs this combination. An AI researcher may be able to explain the details of deep learning without knowing how fourth-grade students can understand data bias through examples from daily life. An information technology teacher may be able to teach programming without knowing how to connect algorithmic judgment to social responsibility.

Correcting the technology-centered tendency requires a general educational understanding of technology. Technical principles must be transformed into key questions accessible to every learner, enabling students to understand the basic relationship between data and models, the principal mechanisms of learning and generation, system error and boundaries of applicability, and the responsibilities that arise when technology enters real contexts. Technical knowledge then becomes a foundational language for cognitive autonomy rather than a compressed version of professional training.

4.4.3 The Activity- and Competition-Centered Tendencies: Substituting Partial Achievements for Education of All

In many schools, artificial intelligence education first appears in the form of activities, clubs, competitions, science festivals, summer camps, and project exhibitions. This path has practical merit. New fields often enter schools through experimental interest-based activities that allow teachers to build experience and develop examples of student work. Projects can stimulate curiosity, reduce barriers to entry, and express the openness and creativity of artificial intelligence. Activities and competitions can also cultivate interest, provide advanced education, and help identify innovative talent. A deviation occurs when these partial projects come to represent the outcomes of general education for all students, weakening universality, continuity, and progression across educational stages.

The activity-centered tendency commonly appears at three levels. First, learning is dominated by one-off experiences. Students participate once in AI painting, intelligent robotics, model training, or prompt design and experience novelty without developing stable concepts. Second, courses center on the exhibition of finished products. Students complete an apparently polished project, but the data, models, evaluation, errors, and responsibilities behind it receive little analysis. Third, schools treat competition results as evidence of educational quality. Awards won by a small number of students can readily be presented as evidence of a school's success in Artificial Intelligence General Education while the foundational literacy of the majority remains unsecured.

Well-designed inquiry projects and competitions can integrate concepts, methods, applications, and ethics and can provide interested students with opportunities for deeper exploration. To enter general education, however, these experiences require the support of a curricular structure. Learning science shows that deep understanding and transfer require conceptual organization, feedback, reflection, and application across contexts(Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988). If an AI project involves only production without conceptual abstraction and transfer tasks, students will struggle to convert a single activity into a lasting capacity. If the achievements of a small number of competitors stand for the educational outcomes

of the whole, a common foundation for general learners cannot be secured.

Research on AI education has reached similar conclusions. Reviewing the design of K–12 AI learning experiences, Zhou, Van Brummelen, and P. Lin (2020) argue that AI education needs a design framework that organizes core competencies, tool resources, and learning activities in forms appropriate for children and adolescents. An analysis by Morales-Navarro et al. (2025) of AI and machine-learning content in Hour of Code activities finds that many activities concentrate on perception and machine learning, devote limited attention to topics such as representation and reasoning, and offer uneven opportunities for hands-on engagement. Such research indicates that activity resources supply important curricular material, while a complete curricular structure must still select and organize that material.

An activity-centered approach can also produce a "successful-example bias." Science exhibitions generally present projects that work smoothly, look impressive, and elicit enthusiastic participation. Failures in model training, difficulties in collecting data, erroneous outputs, ethical disputes, and disagreements within teams are seldom displayed. Yet these imperfect processes can carry the greatest educational value. Through failure students understand overfitting; through misclassification, sample bias; through model instability, generalization; and through project disagreements, responsibility. A curriculum that displays only successful products gives students an excessively optimistic understanding of artificial intelligence.

Consider two project designs. The first asks students to train a waste-classification model with a visual platform and then display its accuracy and application interface. The second adds five requirements to the same task: document the sources of the data, compare performance at different sample sizes, analyze cases of misclassification, discuss lighting and occlusion at an actual waste station, and explain the boundaries of responsibility if the system is introduced on campus. The first is a product project; the second is a general education project. The first trains students to complete a system; the second trains them to judge whether the system can enter the real world.

Correcting the activity- and competition-centered tendencies requires three principles. First, activities must return to concepts: every project should identify the core questions it involves, such as data, models, feedback, generalization, bias, or responsibility. Second, activities must include reflection: students should explain their choices, errors, evidence, boundaries, and improvements. Third, activities must be incorporated into a complete curriculum. One or two isolated activities tend to remain temporary interests and occasions for display. A sustained curriculum makes activities an integral part of understanding knowledge and developing capacities.

4.4.4 The Frontier-Chasing and Fragmentation Tendencies: Substituting Topical Content for a Stable Structure

Artificial intelligence is developing rapidly, continuously introducing concepts such as large models, generative AI, multimodal models, agents, embodied intelligence, and AI for Science. Schools and training organizations can maintain contact with the actual technological environment by bringing new developments into courses in a timely manner. A frontier-chasing deviation arises when topical content determines curricular structure: students encounter many new terms and

products without acquiring the stable concepts needed to understand them. The related tendency toward fragmentation simply juxtaposes traditional content, new technical vocabulary, and disconnected activities, failing to create a coherent conceptual backbone.

The frontier-chasing tendency commonly takes two forms. One transfers a new technology directly into the classroom. Course titles contain "large models," "agents," "multimodality," or "embodied intelligence," although students do not understand such basic concepts as models, data, training, generation, environments, action, and feedback. The other treats frontier applications as magical cases. Teachers demonstrate AI writing poetry, creating images, generating video, or designing experiments, producing a lively classroom atmosphere without giving students opportunities to analyze why the system can perform the task, why it makes errors, where those errors occur, or how its output can be verified.

Curricular content can also lag behind the technological environment. Many courses remain within earlier frameworks of artificial intelligence or conventional machine learning, emphasizing rule-based systems, simple classification, and visual model training without responding to current AI developments and their social environment. Some courses add new concepts and terms piecemeal to conventional information technology content and relabel the result as an AI course. Others directly compress university professional AI courses, substituting reduced professional knowledge for a systematic general educational framework.

All of these approaches create a break between the frontier and the foundations, leaving students without stable concepts or the capacity to understand technological change.

Artificial Intelligence General Education must balance engagement with the frontier and conceptual stability. Stability comes from such core concepts as data, representation, models, learning, generation, feedback, evaluation, systems, and responsibility. Engagement with the frontier concerns the ways in which these concepts are combined and developed in new technological forms. A curriculum should use a stable conceptual backbone to explain frontier developments, enabling students to transfer existing concepts to large models, multimodal models, agents, embodied intelligence, AI for Science, and other new contexts while continuing to question the evidence, boundaries, and responsibilities associated with their outputs.

4.4.5 The Ethics-as-Add-On Tendency: Substituting Risk Warnings for Judgments of Responsibility

AI ethics has entered most frameworks for artificial intelligence education. Privacy protection, algorithmic bias, fairness, transparency, deepfakes, copyright, academic integrity, and high-risk applications frequently appear in textbooks and courses. International frameworks also commonly emphasize ethics and responsible use. UNESCO's AI competency frameworks treat AI ethics as an important dimension of student and teacher competence(Miao, Shiohira, and Lao, 2024; Miao and Cukurova, 2024); the fifth of AI4K12's Big Ideas explicitly addresses societal impact(Touretzky et al., 2019); and China's *Guidelines for Artificial Intelligence General Education in Primary and Secondary Schools (2025 Edition)* emphasize safety and controllability, ethical review, the prevention and control of technological risk, and awareness of social responsibility(Teaching Steering Committee for Basic Education, Ministry of Education, 2025a).

Units on ethics, lists of risks, and safety reminders can help learners develop an initial awareness and provide important points of entry for ethics education. The ethics-as-add-on deviation arises when these reminders come to represent a complete education in responsibility. A course may append a lesson on risks after the technical content, add a page of privacy and safety precautions after a project exhibition, or list statements such as "AI may be biased, may violate privacy, and may be misused." It seldom explains ethical problems as necessary consequences of the interaction between technology and society or traces them through data selection, model design, task objectives, deployment settings, and responsibility in use.

The result is that students understand ethics as an external constraint without understanding how ethical questions are embedded within technical systems and social applications. In image recognition, for example, ethical problems do not arise only after a system has been deployed. Data collection already involves consent and privacy; data labeling involves the construction of categories and cultural bias; model training involves error rates across groups; deployment involves contextual risk and human review; and the interpretation of results involves responsibility and mechanisms of appeal. If a course offers only a final reminder to "protect privacy and ensure fairness," students will struggle to develop systematic ethical judgment.

Research has revealed the distinctive character of AI ethics in education for children and in schools. Analyzing principles of AI ethics in K–12 education, C. Adams et al. (2023) note that children possess distinctive rights and developmental characteristics and that the application of AI ethics principles to school settings requires deeper reflection. A middle-school AI ethics curriculum developed by Payne (2019) uses activities and cases to help students understand algorithmic bias, privacy, fairness, and related questions, showing how ethics can enter students' experience through concrete classroom work. These studies indicate that ethics education must be contextualized and developmentally appropriate.

Treating ethics as an add-on can also weaken students' balanced judgment of the value of technology. A course that only catalogs risks may produce simple fear; a course that only displays convenience may produce blind trust. General education must cultivate responsible trust. Students should be able to recognize the value of AI while identifying its risks; use AI while setting boundaries; and support innovation while upholding human dignity, fairness, and responsibility.

A better approach embeds ethical questions throughout the study of data, models, generation, agents, and applications, allowing students to understand how risks and responsibilities arise from technical stages and contexts of use. Ethics then becomes a necessary dimension of understanding AI systems and is developed systematically in Part III's treatment of risk, ethics, and responsibility.

4.4.6 The Superficial-Assessment Tendency: Substituting Visible Products for Judgments of Capacity

Another prominent problem in Artificial Intelligence General Education is the absence of an assessment system aligned with its aims. Many courses assess students through concept quizzes, tool operation, project artifacts, competition results, and classroom presentations. Each method

has a function. Concept quizzes can test foundational knowledge; tool operation can test practical skill; project artifacts can demonstrate integrated application; and competition results can reveal the creativity of some students. When measured against the aims of general education, however, their limitations become immediately apparent. These methods cannot determine whether students can explain why an AI output may be wrong, transfer a framework for judgment to a new context, identify insufficient evidence and boundary conditions, or reflect on their own dependence on AI.

Superficial assessment first appears as an emphasis on terminology. Students may be able to recite definitions of supervised learning, training sets, neural networks, generative artificial intelligence, and algorithmic bias without being able to use these concepts to analyze real cases. Remembering terminology is a beginning of learning, not the endpoint of AI literacy. The educational value of AI concepts lies in helping students explain phenomena and make judgments; a list of terms supplies only the initial clues for knowledge.

A second manifestation is an emphasis on finished products. After students complete an AI project, assessment often concentrates on the appearance of the interface, whether the functions operate, and the fluency of the presentation. The quality of the artifact can obscure the learning process. In a generative AI environment especially, tools can greatly improve the completeness of a product, making it difficult to infer students' own understanding and learning from the artifact alone. If assessment attends only to results, students can receive high marks without having developed the capacity to define problems, verify evidence, analyze errors, or judge responsibility.

A third manifestation is an emphasis on competitions. Competitions can motivate high-achieving students and encourage schools to commit resources, but competition results cannot represent the general development of all learners. General education emphasizes a common foundation for everyone. Assessment should cover different levels of capacity, educational stages, and modes of learning. Excessive reliance on competition concentrates Artificial Intelligence General Education among a small number of students and undermines both equity and universality.

Educational measurement and instructional theory provide a basis for reform. The principle of "constructive alignment" developed by Biggs and Tang (2011) emphasizes consistency among learning aims, teaching activities, and assessment tasks. Reviewing formative feedback, Shute (2008) shows that effective feedback should help learners adjust their thinking and behavior; a score alone cannot perform this function. The classic work of Sadler (1989) on formative assessment likewise emphasizes that learners need to understand standards of quality, compare their current performance with the desired standard, and act to narrow the gap. Applied to Artificial Intelligence General Education, these principles require assessment to be designed around the capacity for cognitive autonomy.

Assessment must therefore extend to such learning outcomes as explanation, transfer, judgment, and self-awareness. It should examine whether students can explain AI systems and outputs, apply what they have learned in new contexts, weigh evidence, risk, and responsibility, and reflect on their own use of AI and process of judgment. The specific structure and forms of assessment tasks are developed later in relation to the curriculum framework.

Such assessment requires more time and greater professional competence from teachers, but

it corresponds more closely to the aims of Artificial Intelligence General Education. If Artificial Intelligence General Education is to develop learners' capacity for cognitive autonomy, assessment must attend to how students judge as well as to what they produce.

4.4.7 The Superficial-Teacher-Training Tendency: Substituting Tool Training for the Development of Teaching Capacity

The quality of Artificial Intelligence General Education depends to a considerable extent on teachers. They must understand foundational AI concepts, transform knowledge into questions and tasks appropriate for different educational stages, address integrity, privacy, and ethics, and assess students' actual progress in learning. Tool operation and case demonstrations can help teachers enter a new technological environment quickly and are necessary components of teacher preparation. A superficial-teacher-training deviation arises when such instruction becomes the principal content of teacher development, leaving insufficient room for conceptual structures, pedagogy, assessment, and the capacity required for ethical governance.

This problem points to the importance of a system for teacher growth. Artificial intelligence is a rapidly developing new discipline. Teachers cannot be expected to construct the knowledge framework of Artificial Intelligence General Education through self-study alone. They require sustained and systematic support, including a scientifically grounded conceptual system, a stable curricular framework, and continuously updated resources.

4.4.8 The Common Sources of Misaligned Aims

The deviations described above appear different, but each involves a misalignment between a partial educational form and the aims of general education, and they share several underlying causes. First, AI education is developing rapidly, making curriculum construction susceptible to the pace of technological change. Second, educational systems favor visible products: tool outputs, project exhibitions, and competition results are easier to observe than cognitive capacities. Third, artificial intelligence itself is interdisciplinary, and established school subject structures cannot readily accommodate it. Fourth, teacher capacity, assessment instruments, and systems of resources remain underdeveloped. Fifth, society combines high expectations for artificial intelligence with substantial anxiety, causing educational practice to oscillate between technological optimism and vigilance toward risk. Sixth, "artificial intelligence education" and "Artificial Intelligence General Education" are often used interchangeably in policy communication and curricular practice, leaving the boundaries among professional preparation, skills training, interest-based activities, and general education insufficiently clear.

Table 4.2 summarizes these sources.

Table 4.2: Major deviations in current Artificial Intelligence General Education, their consequences, and directions for correction

Type of deviation	**Typical manifestation**	**Educational consequence**	**Direction for correction**
Tool-centered	Centers on prompting, product operation, and greater efficiency	Students can invoke systems but lack judgment of sources, evidence, and boundaries	Transform tool use into output comparison, evidence verification, and records of reflection
Technology-centered	Directly transfers professional technical training into a curriculum for all students	Creates high barriers to entry while neglecting social judgment and responsibility	Establish a general educational understanding of technology that explains data, models, generation, feedback, and responsibility
Activity-centered	Treats brief experiences, science festivals, clubs, and exhibitions as a sustained curriculum	Produces fragmented experience with little conceptual structure or transfer	Integrate activities into a progressive curriculum and strengthen conceptual abstraction and reflection
Competition-centered	Treats projects and awards produced by a small number of students as evidence of general educational outcomes	Weakens universality and neglects foundational literacy	Establish assessment of foundational capacities for all students
Frontier-chasing and fragmented	Organizes courses around new terms, new products, and disconnected topics	Exposes students to current trends without a stable conceptual structure	Explain frontier technologies through data, models, learning, generation, feedback, and systems
Ethics-as-add-on	Appends risk lists and safety reminders at the end	Separates ethics from technical processes and weakens judgments of responsibility	Embed ethics throughout data, models, applications, and assessment

Type of deviation	Typical manifestation	Educational consequence	Direction for correction
Superficial assessment	Emphasizes terminology, products, and presentations while neglecting processes and judgment	Fails to reveal actual learning and cognitive development	Establish explanatory, transfer-oriented, judgment-oriented, and self-awareness-oriented assessment
Superficial teacher training	Concentrates teacher development on tool demonstrations	Leaves teachers unable to design tasks for deep learning	Build a teacher-development system integrating AI content, pedagogy, ethics, and assessment

As the table shows, current AI education has accumulated rich content. When this content enters Artificial Intelligence General Education, it must be integrated by a clear aim: cognitive autonomy. The same generative AI writing task can train tool skills or organize the verification of evidence. The same image-classification project can emphasize technical implementation or analyze data bias and boundaries of generalization. The same discussion of AI ethics can disseminate awareness of risks or develop judgments of responsibility. The intended learners, expected aims, task design, and modes of assessment jointly determine the educational character of these activities and the learning outcomes they produce. Whether cognitive autonomy serves as the overarching goal is an important criterion for judging whether an AI course can fulfill the mission of general education.

4.5 The Cognitive Turn in Policy and Competency Frameworks

Misalignments in practice arise to a considerable degree from theoretical gaps and the resulting instability of policy direction. For a time, Artificial Intelligence General Education lacked a high-level goal and comprehensive structure capable of integrating the whole. Easily implemented and displayed objectives involving tools, skills, and activities could temporarily occupy the center of the curriculum, and these aims were also reasonable choices for early policy. In recent years, this situation has begun to change. A growing number of policies and competency frameworks are turning toward the development of cognitive capacities.

In primary and secondary curriculum development, for example, the AI4K12 initiative proposed Five Big Ideas—perception, representation and reasoning, learning, natural interaction, and societal impact—and organized AI content across grade bands(Touretzky et al., 2019; AI4K12 Initiative, 2019). This framework established a foundation for translating AI knowledge into curricula and creating progressions of learning, while also incorporating the social impact of artificial intelligence. Subsequent AI education frameworks have expanded their attention toward

how learners understand artificial intelligence, judge its outputs, and negotiate the relationship between human beings and AI.

In recent years, this change has entered official policy and competency frameworks more explicitly. UNESCO's AI Competency Framework for Students gives an important place to a "human-centred mindset," emphasizing human agency, human responsibility, and citizenship in the AI era. It requires students to judge whether an AI system is suitable for a particular purpose, whether its use is justified, and what responsibilities people must assume in the process(Miao, Shiohira, and Lao, 2024). These learning aims already involve understanding and judging artificial intelligence and actively regulating the human–AI relationship.

China's *Guidelines for Artificial Intelligence General Education in Primary and Secondary Schools (2025 Edition)* propose an integrated, four-part AI literacy encompassing knowledge, skills, ways of thinking, and values. They incorporate intellectual development, value judgment, human–AI collaboration, and social responsibility into the aims of Artificial Intelligence General Education(Teaching Steering Committee for Basic Education, Ministry of Education, 2025a). This design indicates that Artificial Intelligence General Education in primary and secondary schools is gradually moving from the dissemination of knowledge and experience of technology toward the integrated development of learners' modes of thought, value awareness, and capacity for responsibility.

The AI literacy framework for primary and secondary education released by the OECD and the European Commission strengthens this direction. The framework requires learners to understand the basic operation of AI systems, critically evaluate the accuracy, relevance, and biases of outputs, judge when and why artificial intelligence should be used, and ensure through constraints and monitoring that its use remains aligned with human goals and values(OECD and European Commission, 2026). AI literacy in this framework already includes actively determining when to use a tool, how much to trust it, where the boundaries of judgment lie, and what responsibilities human beings retain. It approaches the foundational capacities individuals need in order to preserve cognitive direction in an intelligent society.

Together, these documents and frameworks reveal an increasingly clear path of development. The scope of AI education is expanding from knowledge and tool use to the evaluation of outputs, ethical judgment, human–AI collaboration, human agency, and the assumption of responsibility. Documents directed toward broad-based AI education place particular emphasis on the development of learners' cognitive capacities and value awareness. Artificial Intelligence General Education in practice has thus begun to respond to the deeper cognitive problems created by intelligent society.

Existing frameworks nevertheless retain two interrelated limitations. First, the levels of the various concepts have not been fully clarified. AI literacy is sometimes described as a learning outcome and sometimes made to function as the overall aim of the curriculum. AI competency can refer to technical competence or encompass ethical attitudes and social participation. AI education can refer simultaneously to the entire educational field and to courses for all learners. Policy documents appropriately serve different audiences and purposes, but this diversity increases the space for interpretation in curricular translation and makes it easier for professional training,

skills development, and general education to become conflated once again.

Second, cognitive elements are generally distributed across parallel dimensions within these frameworks rather than integrated into an overarching goal that organizes curricular content, learning activities, and assessment evidence. Understanding systems, evaluating outputs, sustaining human agency, using technology ethically, and assuming responsibility are already closely related, but their connections are seldom organized into a complete process through which learners actively regulate their own cognitive activity. The identification of sources, use of evidence, allocation of trust, recognition of the boundaries of judgment, reflection on dependence on tools, and assumption of responsibility for action can remain dispersed among different categories of technology, ethics, attitude, or social impact.

The contribution of this book is to bring together the cognitive demands dispersed across existing research and policy documents under the concept of cognitive autonomy and establish it as the overarching goal of Artificial Intelligence General Education. This approach preserves the knowledge, skills, ethics, and responsibility accumulated in the different frameworks while providing a continuous organizing thread for selecting curricular content, structuring tasks, and assessing learning.

4.6 Chapter Summary

Artificial Intelligence General Education is an educational form created through the integration of general education and artificial intelligence education. It takes the knowledge, methods, systems, and social operation of artificial intelligence as its subject matter and assumes general education's mission of developing the capacity for cognitive autonomy in an intelligent society. AI literacy comprises the direct learning outcomes learners develop in AI-related contexts. Cognitive autonomy is both the overarching educational goal of the curriculum and the process through which learners actively regulate their cognitive activity. The capacity for cognitive autonomy sustains that process across tasks and contexts. The four concepts occupy distinct levels—educational form, learning outcome, overarching goal, and stable capacity—while remaining closely connected.

AI education is currently undergoing rapid development. Professional education prepares talent for research, development, and engineering practice; skills training responds to occupational and application needs; interest-based activities and competitions stimulate exploration and innovation; and the use of AI to empower education expands the modes of teaching and learning. Together, these practices form a diverse ecology of AI education and provide rich resources for Artificial Intelligence General Education. Whenever any partial form comes to represent Artificial Intelligence General Education as a whole, however, its aims are narrowed.

The tool-centered tendency confines general education to operational training. The technology-centered tendency allows professional implementation to obscure the common foundation. Activity- and competition-centered tendencies deprive universal education of a sustained structure. Frontier-chasing and fragmentation cause courses to pursue new terminology while neglecting basic concepts. Treating ethics as an add-on reduces judgments of responsibility to risk warnings.

Superficial assessment makes genuine learning difficult to identify, and superficial teacher training limits the design and implementation of tasks for deep learning. These deviations share a common structure: partial content or forms of implementation with legitimate value become detached from the comprehensive aim and gradually displace the complete subject matter that Artificial Intelligence General Education should provide.

Domestic and international policies and competency frameworks have begun to expand from knowledge and tool use toward the evaluation of outputs, ethical judgment, human agency, and the assumption of responsibility, revealing a cognitive turn in AI education. Cognitive autonomy can organize these dispersed requirements into a continuous educational thread, enabling tools, technology, projects, ethics, and applications to serve learners' active regulation of beliefs, evidence, trust, boundaries, and responsibility. Part I has now completed the argument extending from the historical mission of general education, through the cognitive environment of intelligent society and the formulation of cognitive autonomy, to the positioning of Artificial Intelligence General Education. Part II proceeds to explain why Artificial Intelligence General Education can and should assume the task of developing the capacity for cognitive autonomy.

Part II

Theoretical Foundations: How Artificial Intelligence General Education Cultivates Cognitive Autonomy

Introduction to Part II

Part I argued that cognitive autonomy should become an important goal of general education in an intelligent society, and that Artificial Intelligence General Education therefore assumes responsibility for cultivating the capacity for cognitive autonomy. Once this goal has been established, a further question must be answered: why is Artificial Intelligence General Education both capable of and responsible for undertaking this task? The claim that "AI has created the cognitive predicaments of the intelligent society, and therefore cognitive autonomy should be cultivated through the study of AI" has an intuitive plausibility, but it still requires a deeper rational foundation.

Answering this question first requires examining the characteristics of AI as a field and the routes by which it enters social activity. These correspond, respectively, to the computationalization of intelligent activity and the computationalization of social activity. This inquiry provides a scientific foundation for cultivating cognitive autonomy through Artificial Intelligence General Education. Second, it is necessary to consider how curriculum content and learning activities can be organized in light of the characteristics of AI and the ways in which it affects society, so that instruction consistently advances cognitive autonomy. This establishes the feasibility of the proposition that Artificial Intelligence General Education can cultivate cognitive autonomy. Finally, its intellectual training must be compared with that provided by other disciplines. Such a comparison clarifies its distinctive contribution to cognitive autonomy and explains why it requires relatively independent content and structure within the curriculum.

This part contains four chapters. Chapter 5 examines AI's objects of inquiry, scientific system, and modes of thought, showing how intelligent activity acquires forms that can be computed, constructed, and tested. Chapter 6 analyzes the internal mechanisms through which AI is transformed from computational capacity into social force, explaining the computationalization of social activity and its real-world efficacy. Chapter 7 proposes a four-capacity model comprising induction, generalization, judgment, and self-awareness, and explains how learning about AI can foster cognitive autonomy. Chapter 8 compares the intellectual training provided by Artificial Intelligence General Education with that of other disciplines and identifies its distinctive contribution to the cultivation of cognitive autonomy. Together, these four chapters develop the argument through scientific foundations, formative mechanisms, and disciplinary distinctiveness. They explain why Artificial Intelligence General Education can and should cultivate cognitive autonomy, thereby establishing its distinctive place within the system of general education for an intelligent society.

Chapter 5

Artificial Intelligence: Object of Study, Scientific System, and Modes of Thought

Abstract. Artificial Intelligence General Education must rest on an accurate understanding of the disciplinary character of artificial intelligence. Discussions of intelligence have long spanned philosophy, psychology, neuroscience, computer science, and other fields. Artificial intelligence gradually became a science by transforming "intelligence," which is difficult to observe directly, into observable tasks, computable representations, constructible mechanisms, and testable systems. This transformation gives intelligent activity a form that can be analyzed, implemented, compared, and revised. It constitutes the computationalization of intelligent activity and provides the scientific foundation on which Artificial Intelligence General Education cultivates the capacity for cognitive autonomy. On this basis, the chapter defines artificial intelligence as the science of using computers to simulate intelligent human behavior. It further distinguishes two basic routes in the pursuit of machine intelligence: the behaviorist route, whose primary aim is to realize and test intelligent behavior, and the internalist route, which seeks to reproduce or approximate human cognitive processes and their neural realization. Both remain subject to the tests of scientific practice. The chapter then organizes the scientific system of artificial intelligence into three interconnected levels: mathematical and computational foundations; mechanisms and methods for realizing intelligence; and problem formalization and system implementation. It also summarizes several foundational modes of thought in artificial intelligence. AI remains a rapidly developing science with open boundaries. It encompasses relatively stable foundational knowledge, ongoing scholarly disputes, and research visions directed toward the future. Distinguishing the evidential status of different kinds of knowledge is essential to understanding both the science of artificial intelligence and its value for cognitive education.

Keywords: artificial intelligence; scientific definition; intelligent behavior; computational simulation; behaviorism; internalism; scientific system; modes of thought; epistemic status

5.1 How Artificial Intelligence Became a Science

Questions such as "Can machines think?" and "Can machines possess intelligence?" initially arose primarily as philosophical questions. Philosophy can inquire into the nature of intelligence and consciousness, the relation between mind and body, and the meanings of understanding, intentionality, and subjective experience. These questions are fundamental. To enter scientific inquiry grounded in public evidence and replicable tests, however, the relevant issues must be transformed into objects that can be observed, compared, and examined.

The birth of artificial intelligence as a science began precisely with the decomposition of sweeping philosophical questions into specific research problems. Can a machine prove theorems, solve problems, recognize images, understand speech, translate languages, participate in dialogue, plan actions, control a robot, or adjust its behavior in a complex environment on the basis of feedback? For each problem, researchers can specify inputs, outputs, task conditions, and evaluation criteria. They can then propose methods, build systems, conduct experiments, and compare results.

Turing's discussion of machine intelligence epitomizes this transition. Rather than trying to prove that a machine possesses an inner mind identical to that of a human being, he reformulated the issue as an examination of behavioral performance through an imitation game—what later became known as the Turing test (Turing, 1950). The imitation game did not resolve every philosophical dispute about intelligence, understanding, or consciousness. It nevertheless introduced an important scientific strategy: when an object's internal state cannot be observed directly, researchers can begin by establishing publicly testable behavioral criteria.

As the field advanced, artificial intelligence gradually acquired a relatively definite object of study, communicable methods, reproducible and comparable evidence, and shared habits of thought and research. Game playing, theorem proving, pattern recognition, language processing, and robotic action became research objects. Logical reasoning, search, probabilistic modeling, machine learning, and neural networks became communicable methods. Datasets, task-based tests, controlled experiments, and operation in real environments supplied comparable evidence. Conferences, journals, laboratories, and open benchmarks enabled competing research programs to accumulate results around shared problems. Artificial intelligence thus developed from an aspiration to build intelligent machines into a science that advances through computational experiments and system-building practice.

This scientific development has not made the philosophical questions disappear. Whether machines genuinely understand language, whether consciousness can arise from computation, and whether subjective experience can be simulated remain open questions. Scientific development means that AI research cannot establish its conclusions through conceptual speculation alone. It must turn those parts of a question that admit investigation into well-defined objects and submit them to tests through models, systems, and experiments. Philosophy can pose questions, analyze concepts, and examine presuppositions. Science must go on to specify what is being studied, how it can be realized, how it will be tested, and what evidence would support or undermine a conclusion.

5.2 The Object of AI Research: Intelligent Behavior

To date, there is no single, uncontested definition of "intelligence." In everyday language, intelligence may refer to memory, reasoning, understanding, learning, creativity, adaptation, problem solving, or understanding other people. Psychology, neuroscience, philosophy, and computer science also emphasize different aspects. Simply combining their definitions would produce greater ambiguity rather than clarity.

The outward manifestations of intelligence are more readily observed. This book calls these observable and testable manifestations **intelligent behavior**. Viewed across the tasks studied throughout the history of artificial intelligence, intelligent behavior can be organized into four interconnected aspects:

1. **Thinking:** representing problems, retaining knowledge, forming concepts, reasoning, learning, planning, explaining, and devising solutions;
2. **Perception:** identifying objects, states, and relations in images, sounds, language, touch, and multiple forms of sensory signal;
3. **Action:** making choices in light of goals and environments, controlling devices, using tools, collaborating with other agents, and adjusting action in response to feedback;
4. **Emotion and creativity:** recognizing and expressing emotion, generating content with style, aesthetic quality, or novelty, and responding appropriately to human emotional signals.

This classification concerns observable functional performance. It makes no prior claim that a machine possesses the same understanding, emotion, or subjective experience as a human being. A system that recognizes sadness in speech and generates a comforting reply can perform certain emotion-related behaviors. That behavioral fact alone does not establish that the system has the subjective experience of sadness, sympathy, or care. Similarly, a model that completes reasoning problems displays a corresponding problem-solving capability, while the question of whether its internal process is equivalent to human reasoning requires separate investigation.

Given this definition of "intelligent behavior," intelligence can be defined as the capacity to produce such behavior. A machine that can reliably perform a particular class of intelligent behavior may be called an "intelligent machine" possessing the corresponding form of intelligence.

5.3 Two Basic Routes toward Machine Intelligence

Human intelligence provides the point of reference for artificial intelligence, yet views have long differed on how closely a machine should reproduce human internal mechanisms. For analytical purposes, this book distinguishes two routes according to their research aims and standards of evidence: the "behaviorist" route and the "internalist" route. The behaviorist route aims to realize and test intelligent behavior. The internalist route treats human cognitive processes or brain mechanisms as objects to be reproduced or approximated. These terms summarize orientations in AI research; they are not identical to their namesakes in the history of psychology or the philosophy of mind.

The behaviorist route: simulating intelligent behavior

The behaviorist route begins by specifying the behavior a system should exhibit and then investigates how that behavior can be realized computationally. It asks whether a machine can complete a task, adapt to its environment, and maintain its performance under new conditions. It does not require the machine's internal operation to reproduce human cognitive processes.

Symbolic reasoning solves problems through knowledge representation, search, and rule-based operations. Statistical machine learning estimates regularities from data. Deep learning forms distributed representations in multilayer networks. Large models acquire capabilities in language, multimodality, and tool use through large-scale pretraining, contextual conditioning, and post-training. These methods may be inspired by or analogous to human thought, but their scientific validity is assessed primarily through behavioral performance, experimental results, and conditions of use. Machines operate with high-dimensional vectors, gradient-based optimization, and large-scale parallel computation—processes that differ from everyday human thought without diminishing the scientific value of the research.

An important feature of the behaviorist route is that it allows researchers to bypass human internal mechanisms that cannot yet be fully described and directly explore multiple computational means of simulating intelligent behavior. Contemporary progress in visual recognition, speech processing, machine translation, scientific prediction, content generation, and complex games demonstrates the scientific value of this route.

The internalist route: reproducing human internal mechanisms

The internalist route takes human cognitive processes and their realization in the brain as direct models. Its goal extends beyond producing functionally similar results: the machine should also employ internal representations, processing procedures, or neural mechanisms that are the same as or similar to those of human beings. The distinction in cognitive science between "weak equivalence" and "strong equivalence" offers a related formulation. Producing the same input–output behavior establishes weak equivalence. A cognitive model that claims strong equivalence must also use evidence such as reaction times, intermediate states, and process structure to test whether it approximates the algorithms and internal processes actually used by people (Pylyshyn, 1980).

At the cognitive level, this route includes computational modeling of memory, learning, reasoning, and decision-making. At the neural level, it includes brain-inspired computing, neuromorphic computing, brain simulation, and biologically plausible neural models. Spaun, for example, was constructed by Eliasmith and colleagues from approximately 2.5 million spiking neurons modeling multiple brain areas. Within a single system, it performed tasks involving visual recognition, memory, counting, reasoning, and motor control while enabling comparisons between the model's neural activity and human psychological behavior (Eliasmith et al., 2012).

The internalist route faces more stringent evidential requirements. In addition to producing the relevant behavior, a model must explain why its internal representations, intermediate states, processing procedures, or neural mechanisms resemble those of human beings. Those claims must be tested with independent evidence from reaction times, cognitive experiments, neural activity, or biological constraints. The human brain and body form a highly complex biological system; cognition also interacts with development, language, social relations, and cultural experience.

Current science has yet to explain fully how these processes jointly produce human intelligence and subjective experience. These difficulties make strong equivalence hard to establish, but they do not warrant a prior judgment that the route must fail. Research on this route remains a legitimate scientific enterprise so long as it advances testable hypotheses, constructs comparable models, and revises its conclusions in response to evidence.

5.3.1 The Relationship between the Two Routes

Table 5.1: Two Basic Routes toward Machine Intelligence

Dimension	**Behaviorist Route**	**Internalist Route**
Primary aim	Realize and test observable intelligent behavior	Reproduce or approximate human cognitive processes and their neural realization
Criterion of realization	Whether the system produces effective behavior in a specified task and environment	Whether the system produces the target behavior and whether its internal representations, processing procedures, or neural mechanisms have a testable correspondence with those of human beings
Permissible differences	Allows machines to follow computational paths different from those of humans	Requires key internal representations, processes, or neural mechanisms to resemble those of humans
Current progress	Substantial achievements across many tasks and applications	Continuing exploration in brain-inspired and neuromorphic computing, brain simulation, and cognitive-process modeling
Principal difficulty	Effective behavior does not directly disclose internal mechanisms or establish understanding, consciousness, or subjective experience	Human internal mechanisms remain incompletely understood, and claims of internal correspondence require independent evidence

The two routes can inform one another. Behaviorist systems can draw inspiration from brain science without thereby changing their primary criterion from task performance. Internalist research must also demonstrate through outward behavior that a model possesses the relevant functions. The distinction lies in research aims and evidential standards: the first route accepts weak equivalence at the functional level, whereas the second also pursues strong equivalence

at the process or neural level. Artificial intelligence has not reached a stage at which any one route can exclude every alternative. The competing explanations advanced by different schools concerning knowledge, learning, the brain, the body, the environment, and consciousness are an important source of the field's vitality.

5.4 A Definition of Artificial Intelligence

On the basis of this analysis, this book defines artificial intelligence in terms of its object of study and means of realization:

Artificial intelligence is the science of using computers to simulate intelligent human behavior.

This definition has four essential implications:

1. Artificial intelligence is a science. It does not refer solely to a particular technology or to some particular form of "intelligence."
2. Intelligent human behavior supplies the starting point and reference for research. Artificial intelligence originated in the aspiration to reproduce human intellectual capabilities in machines, and many tasks continue to compare machine performance with human performance.
3. Machines take intelligent human behavior as the object of simulation, while their internal mechanisms and processes may differ from those of humans. Machines may employ different mechanisms to produce the same, similar, or more powerful functions.
4. Computation is the principal means by which artificial intelligence is realized. Chemical, physical, or mechanical means might also imitate intelligent human behavior, but such approaches fall outside the scope of artificial intelligence as defined here.

When discussing the object of study, body of knowledge, and disciplinary development, this book generally uses "artificial intelligence" in its disciplinary sense. When discussing specific capabilities, real-world applications, and social effects, it uses "AI technology," "AI systems," or "the AI technology system" as the context requires. Where no confusion is likely, "AI" may also serve as shorthand for these technologies, systems, or the broader technological system.

5.5 The Scientific System of Artificial Intelligence

Under the preceding definition, observable intelligent behavior is the object of AI research. Identifying an object of study is only the starting point for a discipline. A science must also develop an interconnected body of basic concepts, theoretical propositions, mechanisms of realization, research methods, and forms of validation around that object.

Artificial intelligence first enters public view through systems that operate. A machine recognizes images, processes language, generates content, formulates plans, or controls equipment. These observable behaviors are the most immediate phenomena of AI research. Three levels of question can be asked in sequence about an AI system. What real-world problem does it

address, and what are its task, boundaries, and evaluation criteria? What information-processing mechanisms, models, and algorithms produce its behavior? Why are those mechanisms capable of computation, learning, and generalization, and what theoretical conditions limit them?

Following this line of inquiry, the scientific system of artificial intelligence can be organized from top to bottom into three interconnected levels: the formalization of real-world tasks and implementation of systems; mechanisms and methods for realizing intelligent behavior; and mathematical and computational foundations. The top level converts real-world aims into AI systems that can operate and be tested. The middle level explains how capabilities such as reasoning, prediction, generation, planning, and action arise. The foundational level explains how objects can be represented, what lies within the scope of computation, and under what conditions machine learning can produce generalizable regularities. Together, these levels form a scientific structure that extends from real-world problems to system behavior and from behavioral mechanisms to theoretical conditions.

Table 5.2: Three Levels in the Scientific System of Artificial Intelligence

Level	**Direct Object of Study**	**Core Scientific Question**	**Principal Theoretical Components**
Real-world tasks and system implementation	Formalized intelligent tasks and the AI systems that perform them	How can real-world aims be transformed into computable and evaluable tasks, and how can different functions and methods be organized into systems that operate in an environment and admit testing?	Task definition and representation; inputs and outputs; states and actions; objectives and constraints; organization of data and knowledge; system architecture and functional modules; evaluation and validation
Mechanisms and methods for realizing intelligent behavior	Information-processing mechanisms that produce intelligent behavior	How do knowledge, data, and interaction with the environment produce reasoning, prediction, generation, planning, and action?	Knowledge representation and reasoning; search; probabilistic models; machine learning; neural networks; reinforcement learning; generative models; planning and control

Level	Direct Object of Study	Core Scientific Question	Principal Theoretical Components
Mathematical and computational foundations	Formal structures, computational processes, and regularities of learning	How can objects relevant to intelligence be represented, which problems are computable, and under what conditions can a machine form generalizable regularities from finite experience?	Logic and discrete structures; linear algebra; probability and statistics; information theory; optimization; algorithm theory, computability, and complexity; learning theory

5.5.1 Real-World Tasks and System Implementation

The top level of AI's scientific system concerns real-world tasks and the systems that perform them. Artificial intelligence is generally observed as a system carrying out a task in a particular environment: assisting a physician with diagnosis, predicting traffic conditions, answering questions, generating images, or controlling a robot. To transform such aims into objects of scientific inquiry, researchers must specify the environment the system faces, the information it receives, the outputs it produces, the goal it pursues, and the evidence by which achievement of that goal will be judged.

Phrases such as "help physicians diagnose illness," "improve traffic flow," or "support student learning" express real needs, but they do not yet define tasks a machine can process directly. Problem formalization identifies the boundaries of the object, environment, and task; decomposes broad aims into specific tasks such as recognition, prediction, generation, planning, or control; specifies inputs and outputs, states and actions, objective functions and constraints; and selects symbols, vectors, graphs, or probabilistic structures with which to represent real-world objects. Researchers must then organize the data and knowledge required by the system and determine the conditions under which it will be trained, operated, and evaluated.

Problem formalization determines the problem that an AI system actually solves. Which real-world differences are retained as variables, which conditions enter the model, which outcomes count as task completion, and which metrics evaluate the system all shape the definition of its behavior. Evaluation must also account for different kinds of error and their real-world costs. In medical diagnosis, for example, missed diagnoses and false positives are both predictive errors, but they may have different consequences. If an evaluation metric does not reflect this difference, an improvement in a technical score may fail to improve clinical outcomes. Task definitions, objectives, constraints, and evaluation standards therefore jointly specify the behavior an AI system is designed to pursue.

Once a task has been defined, functions such as information acquisition, data processing, knowledge organization, model training, reasoning and planning, memory, tool use, output generation, and action control must be organized into a complete system. Modern AI often describes this overall structure in terms of an agent and its relation to an environment. The system obtains information through sensors or data interfaces, computes on the basis of its internal state, knowledge, and models, and affects the environment through outputs or actions (Russell and Norvig, 2021). How functional modules exchange information, how errors in local models affect overall behavior, and how the system responds to environmental change and anomalous conditions are all research questions at the system level.

For example, the broad aim of enabling a machine to "understand the road" can be decomposed into vehicle detection, pedestrian recognition, lane segmentation, distance estimation, and motion prediction. Cameras and other sensors supply inputs. Object classes, spatial positions, and future trajectories constitute intermediate results. Recognition and prediction then feed into localization, route planning, and vehicle-control modules, ultimately producing continuous driving behavior. The system must also be evaluated across different roads, weather conditions, lighting conditions, and traffic situations to test the accuracy, stability, and range of applicability of both its individual modules and its overall behavior.

The output of a single model is thus only one component of an AI system. An AI problem receives a complete scientific formulation only when researchers explain how real-world objects become computational objects, how local models enter a system architecture, how the parts jointly produce observable behavior, and how that behavior can be compared and tested. Systems accordingly become objects of AI research: theories and methods are assembled within them, and propositions about intelligent behavior are observed, measured, and tested through their operation.

5.5.2 Mechanisms and Methods for Realizing Intelligent Behavior

Once a task and the behavior of its system have been defined, the next question is how intelligent behavior can be realized. Mechanisms and methods occupy the middle level of the scientific system. They examine how information is organized and transformed to produce capabilities in reasoning, prediction, generation, planning, and action. A "mechanism" specifies the process through which information is transformed into a given form of behavior. A "method" consists of the models, objectives, and algorithms used to construct and operate that mechanism. Models describe objects, relations, changes of state, probabilistic dependencies, or regularities between inputs and outputs. Algorithms specify how to perform search, reasoning, training, generation, planning, or action over those models.

In terms of the source of system capabilities, AI methods can be divided into knowledge-based and learning-based approaches.

Knowledge-based AI explicitly represents facts, concepts, relations, and rules, then derives conclusions or plans through search and reasoning. Research on physical symbol systems and expert systems demonstrated, respectively, the roles of symbol manipulation and domain knowledge in realizing intelligence (Newell and Herbert A. Simon, 1976; Feigenbaum, 1977).

Under this mechanism, system capabilities arise primarily from established knowledge structures, inference rules, and search procedures.

Learning-based AI adjusts models from examples so that a system acquires capacities for classification, prediction, generation, or representation. Probabilistic models use random variables and their dependencies to express uncertain knowledge (Pearl, 1988). Neural networks use parameterized models, objective functions, and optimization algorithms to form complex mappings and distributed representations from data (Bishop, 2006; LeCun, Yoshua Bengio, and G. Hinton, 2015). Here, the regularities governing system behavior are formed primarily through sample-based training.

The two approaches often coexist in modern AI systems. A system may learn a model from data, consult knowledge bases, rules, and external tools, and organize sustained action through continued interaction with its environment. An agent, for example, may use a neural network to interpret an input, retrieve external knowledge, apply planning methods to decompose a task, and adjust its next action in light of the result returned by a tool.

This level of mechanisms and methods connects system behavior with foundational theory. Looking upward, it must explain the processes that produce a system's capabilities in recognition, prediction, generation, planning, and action. Looking downward, it must specify the representations, models, and algorithms employed, together with the boundary conditions on which they depend.

5.5.3 Mathematical and Computational Foundations

Following the mechanisms that produce intelligent behavior downward leads to the mathematical and computational foundations of AI. Every model and algorithm requires a formal language with which to represent objects, a determinate computational process for handling information, and specified conditions under which reasoning, optimization, or learning can occur. The central questions at this level concern how representation, computation, and learning are possible, the conditions under which they hold, and their principled limitations.

Real-world objects must be represented formally before they can become computational objects. Such representations may take the form of symbols, graphs, vectors, matrices, probability distributions, or dynamic states. Logic and discrete mathematics express facts, rules, relations, and inferential structures. Linear algebra provides a unified language for high-dimensional data and parametric models. Probability and statistics describe uncertainty and the relation between samples and populations. Information theory studies the measurement, encoding, and transmission of information. Optimization theory examines how suitable solutions can be found under objectives and constraints. Together, these theories determine what an AI system can express and the forms in which it can process those expressions.

Formal representation leads to a further question: can the resulting computational problem be solved by an algorithm, and at what cost? The abstract model of computation established by Turing (1936) gave rigorous meaning to "computable" and revealed that some formal problems cannot be solved algorithmically. Computability theory thereby marks the principled boundaries of computation. Computational complexity theory further distinguishes problems that are solvable

in principle from those that can be solved effectively under real resource constraints. Algorithm theory studies the construction, correctness, efficiency, and resource use of computational procedures and supplies the computational basis for specific AI methods.

Machine-learning approaches pose another question: can a system form a model from finite samples or interactive experience that remains effective in unseen cases? Probability and statistics explain how population-level regularities can be inferred from finite samples. Optimization explains how model parameters can be adjusted in pursuit of an objective. Learning theory studies the conditional relationships among the amount of experience, the hypothesis space, model complexity, and generalization (Mitchell, 1997; Bishop, 2006).

These theories also reveal the limits within which learning methods apply. The No Free Lunch theorem shows that, when averaged uniformly over all possible objective functions, no optimization method performs better than every other method across all problems. A method gains an advantage on a particular task only when its assumptions match the structure of that task (Wolpert and Macready, 1997). The effectiveness of a model is therefore always relative to a data distribution, task structure, representational scheme, objective, and environmental conditions. Claims that a method is "intelligent" or "effective" cannot be scientifically rigorous when detached from these conditions.

Mathematical and computational foundations provide the bottom-level support for AI's scientific system. They are more than an assortment of mathematical tools. They form a theoretical structure organized around how representation is established, computation proceeds, learning becomes possible, and capability boundaries are determined. This structure subjects model selection, algorithm design, and explanations of system behavior to explicit definitions, formal derivation, and theoretical conditions. It also provides a basis for analyzing the deeper causes of system failure.

5.5.4 Connections among the Three Levels

From top to bottom, the scientific system of artificial intelligence provides a progressively deeper path of explanation:

Real-world tasks and system behavior $\longrightarrow$ mechanisms, models, and algorithms $\longrightarrow$ foundational conditions of representation, computation, and learning.

At the top level, researchers define a real-world task; specify the system's inputs, outputs, goals, constraints, operating environment, and evaluation criteria; and formulate observable and testable scientific questions in terms of system behavior. At the middle level, they analyze how knowledge, data, and interaction with the environment are transformed by models and algorithms into the relevant behavior. At the foundational level, they ask why those representations and computational processes work, under what conditions they are effective, and what theoretical boundaries constrain them.

The interfaces between the three levels are explicit. The level of real-world tasks and systems gives intelligent behavior a definite task meaning, environmental conditions, and standards of evaluation. The level of mechanisms and methods supplies the processes that realize reasoning,

prediction, generation, planning, and control. Mathematical and computational foundations supply formal languages, rules of computation, conditions of learning, and theoretical boundaries. Questions posed at higher levels are progressively translated into requirements on representations, models, and algorithms, while theories and methods developed at lower levels support the construction, operation, and validation of systems above them.

The scientific rigor of artificial intelligence rests on this vertically connected structure. Beginning with observable intelligent behavior, it transforms real-world aims into AI systems with explicit structures that can operate and undergo testing. Models and algorithms then explain how the behavior is produced, and mathematical and computational theories identify the conditions under which representation, computation, and learning are possible. Intelligent behavior thus becomes a scientific problem that can be defined, decomposed, realized, explained, and validated. The AI system serves as the experimental vehicle connecting real-world tasks, mechanisms of realization, and theoretical propositions.

Crucially, this layered scientific system can accommodate different theoretical traditions, technical routes, and forms of system. Its stability comes from relationships among the levels that can be stated explicitly, traced step by step, and tested repeatedly. New real-world tasks, system forms, learning mechanisms, and modes of representation will continue to expand the specific content of AI. The connected structure that leads from real-world systems to foundational conditions nevertheless provides a basic framework for understanding artificial intelligence as a developing science.

5.6 Disciplinary Characteristics of Artificial Intelligence

Intelligent behavior as the object of study: Artificial intelligence organizes its questions around intelligent machine behavior. This object of study is open-ended: as machine capabilities advance, tasks once regarded as intelligent may come to be viewed as ordinary computational functions, prompting the pursuit of more advanced forms of intelligent behavior. The boundary of AI therefore continues to move.

Computation as the foundational method: There are multiple possible ways to create intelligent machines. Artificial intelligence pursues one of them: using computers to simulate intelligent behavior through computation. This computational approach arose from earlier efforts to mathematize human thought and gained its practical foundation from the capabilities of the general-purpose computer.

Knowledge and learning as two fundamental sources of capability: AI capabilities may arise from knowledge explicitly supplied by people or from learning through data, feedback, and interaction. Each source has distinct strengths. Knowledge provides structure, constraints, and interpretable relations; learning can address complex patterns and uncertain environments. Retrieval-augmented generation, neuro-symbolic systems, and tool-using agents show that the two approaches are converging again.

A close integration of scientific research and engineering practice: Artificial intelligence requires both testable theories and constructed models and systems. Research claims are

commonly evaluated through datasets, experiments, benchmarks, and real-world applications. Engineering scale, hardware performance, data pipelines, and deployment environments directly affect scientific results. Theory, experiment, and engineering are therefore difficult to separate completely in AI.

Strong interdisciplinary and foundational-methodological characteristics: Artificial intelligence draws methods from many disciplines and in turn enters those disciplines to participate in data analysis, model-based prediction, hypothesis generation, and experimental design. It is increasingly becoming a general method of problem solving. Within a specific domain, however, AI remains subject to that domain's theoretical constraints and evidential standards. A general-purpose model cannot replace expert validation.

The coevolution of capabilities, risks, and social structures: AI systems enter information distribution, educational assessment, medical diagnosis, labor management, and public decision-making. Their technical objectives, choices of data, and modes of deployment embody value judgments and can redistribute opportunity, power, and responsibility. The foundational concepts of artificial intelligence must therefore describe relations among systems, people, organizations, and social environments alongside model capabilities.

5.7 Modes of Thought in Artificial Intelligence

Over the course of its development, artificial intelligence has formed a set of disciplinary modes of thought. They arise from a shared research process: transforming real-world questions about intelligence into problems that can be represented, computed, implemented, and tested, then constructing solutions through models, algorithms, and systems.

Questions such as "How can a machine recognize objects?" "How can future change be predicted?" "How can new content be generated?" and "How should an agent act in an environment?" typically lead AI research through several interconnected stages. Researchers first clarify the object, aim, and boundaries of the problem and transform a real-world concern into a formal task. They then extract computable information from data and domain knowledge and address the uncertainty arising from incomplete information, environmental change, and error. On that basis, they select methods that are as simple as the problem permits, balance performance against resources and efficiency, and finally place the method in an actual task and operating environment to test its effectiveness through experiment and practice.

The principal modes of thought in artificial intelligence can therefore be summarized as formalizing problems, learning from data, using domain knowledge as a constraint, reasoning under uncertainty, balancing performance and efficiency, seeking simple and effective methods, and emphasizing practical testing. These modes run through the entire process from posing a question and constructing a model to forming an operational system.

5.7.1 Formalizing Real-World Problems as Computational Problems

AI research begins by transforming broad and ambiguous real-world problems into clearly structured computational problems. People can state aims such as "enable machines to understand

language," "help physicians diagnose disease," or "make robots act autonomously," but computers cannot process these abstract requirements directly. Researchers must specify the objects the system encounters, the information it receives, the results it produces, the goal it pursues, and the conditions under which the task counts as complete.

This process is called problem formalization. It generally includes defining task boundaries, determining inputs and outputs, selecting variables and representations for the objects involved, and specifying objective functions, constraints, and evaluation criteria. "Enable a machine to understand a passage," for example, can be translated into tasks such as answering questions, extracting information, determining relations within a text, generating a summary, or following instructions. Each task definition entails distinct inputs, outputs, and forms of evaluation and thereby gives "understanding" a behavioral meaning that can be observed and tested.

Formalization requires selection and abstraction. The real world contains innumerable details; a computational model can preserve only those relevant to its current task. Which objects become variables, which differences receive attention, and which conditions enter the model all shape the problem the system ultimately solves. A model that predicts student learning from test scores alone primarily describes changes in those scores. A model that also considers the learning process, the structure of knowledge, and task difficulty addresses a different problem.

Once formalized, a problem can be handled through mathematical models and computational processes. Researchers may describe it using equations, probability distributions, graph structures, vector spaces, or state transitions and then seek solutions through search, reasoning, optimization, or related processes. Here, "solving" extends beyond deriving a closed-form answer from an equation; it means finding a way to resolve the problem. Many AI problems have no directly expressible closed-form solution. Researchers must construct a model, algorithm, or operational system and obtain results incrementally through computation, simulation, training, and interaction.

Formalization is therefore also a constructive mode of thought. Researchers turn real-world problems into computational objects, build models and systems capable of exhibiting the relevant behavior, and use those artificial objects to investigate the original problem. Whether a real-world problem can be solved effectively depends to a large extent on whether it has been transformed into an appropriate computational problem.

5.7.2 Learning Regularities from Data

In conventional programming, people generally specify processing rules explicitly. Artificial intelligence often confronts problems whose rules cannot be enumerated completely. People recognize a face, understand a sentence, or identify an obstacle in the road with ease, yet they find it difficult to state every rule required for the task. AI accordingly developed an important strategy: enabling machines to learn regularities from data and experience.

Learning from data means adjusting a model on the basis of examples so that it gradually forms relations between inputs and outputs or discovers latent structure in the data. Supervised learning acquires predictive relations from examples paired with answers. Unsupervised learning discovers categories, structures, or representations within data. Reinforcement learning develops

policies from actions and the feedback supplied by an environment. Despite their differences, all three treat experience as an important source of system capability.

The central value of data-driven methods lies in their ability to address complex regularities that cannot be described exhaustively through hand-written rules. The appearance of objects in images, patterns of natural-language use, and relations between protein sequences and spatial structures each involve many interacting factors. Machine learning can draw on large collections of examples to identify statistical structure in high-dimensional data that is useful for a task.

Extracting regularities from data does not show that a model has learned a universal law. Training data offer only finite samples of the world. A model may memorize accidental features or achieve a high score by exploiting correlations irrelevant to the task. AI research therefore assigns special importance to generalization: can the regularities formed from existing data be applied to objects and contexts the model has not previously encountered?

Researchers must consequently understand where the data came from, whether they cover important cases, whether the training distribution matches the application environment, and whether the data contain bias, omissions, or errors. A model that once worked may fail when its environment or data distribution changes. Learning from data thus involves two connected tasks: discovering regularities in experience and determining the range within which those regularities hold and can be generalized.

5.7.3 Using Domain Knowledge as a Constraint

Data are not the only source of information available to artificial intelligence. Real-world problems have their own object structures, governing regularities, and constraints. Medicine involves human anatomy, disease mechanisms, and standards of care. Physics is constrained by conservation laws and boundary conditions. Traffic systems have road structures, vehicle dynamics, and safety rules. Educational problems are closely related to knowledge structures, cognitive development, and instructional aims. Domain knowledge determines which relations in a problem are meaningful and provides grounds for selecting methods.

Domain knowledge can enter an AI system in multiple ways. Researchers can use domain concepts to determine how data are represented, established regularities to constrain model structures and outputs, causal relations to distinguish surface correlation from underlying mechanisms, and safety rules to rule out unacceptable actions. Expert knowledge can also narrow a search space, guide the design of evaluation metrics, and support the interpretation of system results.

Incorporating domain knowledge reduces what a model must relearn from data. It becomes especially important when samples are scarce, experiments are costly, or errors carry serious consequences. In materials research, known physical laws can help a system eliminate clearly implausible candidate structures. In medicine, clinical knowledge and safety requirements can constrain the range of model recommendations. In robotic control, kinematic and dynamic laws can prevent a system from exploring infeasible or dangerous actions.

The choice of method must likewise fit the characteristics of a domain. The No Free Lunch theorem shows that, when all possible problems are weighted equally, no single method

outperforms every other method universally (Wolpert and Macready, 1997). An algorithm generally succeeds because its assumptions accord with the structure of a particular problem. Images exhibit spatial locality, language has sequential and contextual structure, and social networks consist of nodes and connections. Methods for these domains need to exploit their differing structures.

The quality of an AI method therefore cannot be judged independently of the problem to which it is applied. A model's suitability depends on the task objective, data conditions, domain regularities, risk requirements, and operating environment. Understanding the domain structure, identifying the assumptions of a method, and bringing the two into alignment form an important mode of thought for solving real-world problems with AI.

5.7.4 Reasoning under Uncertainty

Information available to AI is often incomplete, noisy, and continually changing. Sensors may introduce errors, data may be missing, the same phenomenon may have multiple causes, and future environments cannot be known in full. Even when a model returns a definite classification, prediction, or action, its grounds and result may remain uncertain to different degrees.

Handling uncertainty begins with acknowledging the limits of information and models. A system needs to distinguish available evidence from missing information, estimate the likelihood of different outcomes, and express the confidence warranted by its judgment. Probabilistic models provide important tools for representing uncertainty: probability distributions can represent multiple possible outcomes, new evidence can update a judgment, and expected benefits and risks can be compared across choices (Pearl, 1988).

Uncertainty has different sources. Some arises from stochastic variation in the object itself: weather, markets, and human behavior fluctuate in ways that cannot be predicted completely. Other uncertainty arises from insufficient knowledge and data: a model may rarely have encountered a certain class of example, or a current input may lie outside the range of its training data. The first type can generally only be represented probabilistically. The second may be reduced by collecting data, improving the model, or seeking expert judgment.

An AI system must also determine how to act in light of the degree of uncertainty. When a prediction is reliable, the system may carry out the corresponding operation directly. When evidence is inadequate or risk is high, it may request more information, retain several candidate results, defer to human judgment, or adopt a more conservative course. Medical diagnosis, autonomous driving, and public decision-making especially require confidence in a result, the cost of error, and rules for action to be considered together.

The treatment of uncertainty embodies a basic form of rationality in artificial intelligence: preserve appropriate boundaries of judgment under limited information, update confidence as evidence changes, and match the strength of action to the reliability of the evidence.

5.7.5 Balancing Performance and Efficiency

AI methods must operate under finite constraints of time, data, computing power, storage, energy, and economic cost. Even a highly accurate model may fail to provide a usable solution if its

training cost is prohibitive, its response is too slow, or it cannot be deployed on the intended device. AI research therefore examines both the performance a system can achieve and the resources required to achieve it.

Performance has multiple dimensions. Classification tasks may emphasize accuracy, recall, and false-positive rates. Generative tasks may emphasize quality, diversity, and reliability. Agent systems must also consider task-completion rates, safety, stability, and adaptation to the environment. These metrics can conflict. Greater sensitivity to one type of error may increase another. Expanded capability can bring longer runtime and higher resource consumption. Immediate response can limit the amount of computation a model can perform.

Efficiency must likewise be understood relative to the application environment. Models operating in data centers face computational and energy conditions different from those on phones, vehicles, robots, or sensors. Real-time control requires computation to finish within a specified interval. Edge devices are constrained by storage and battery capacity. Large-scale public services must account for the cost of each call and the ability to serve many users simultaneously.

Researchers must therefore balance multiple objectives. Model compression, parameter sharing, approximate computation, hierarchical processing, and on-demand invocation can all reduce resource use. A well-designed system may handle simple cases with low-cost methods while routing complex or high-risk cases to more capable models or human experts.

This mode of thought reflects the dual character of AI as a computational and engineering science. An effective solution must produce sufficiently good behavior under particular resource and environmental conditions. The pursuit of performance is always constrained by time, space, energy, cost, and risk.

5.7.6 Seeking Simple and Effective Methods First

When confronted with a complex problem, it is tempting to adopt a larger and more intricate method immediately. Greater model complexity, however, brings more parameters, higher computational cost, and greater difficulty of analysis. It may also make a system more likely to learn accidental features in the data. AI research therefore values a plain yet important principle: begin with a simple method that captures the principal structure of the problem, then add complexity as actual needs require.

Simple methods provide a starting point for understanding a problem. A linear model, an explicit set of rules, or a small decision tree can help researchers determine which variables matter, whether a learnable regularity exists in the data, and where the gains of a complex model come from. The improvement delivered by a complex method becomes interpretable only through comparison with a simple baseline.

Simplicity can also make methods more stable and testable. Simpler models generally make it easier to analyze errors, discover data problems, explain results, and replicate experiments. With limited data, a simpler model may also generalize better than a more complex one. Even when a complex system is ultimately required, simple models can still serve as baselines, anomaly checks, or safety safeguards.

Simplicity does not imply always selecting the smallest model. Real-world problems may contain structures sufficiently complex to require expressive models. The principle is that complexity should have a clear justification: additional structures, parameters, and computation should address observed problems, and experiments should show that they deliver a necessary improvement.

5.7.7 Testing Solutions through Practice

Artificial intelligence is a strongly experimental science. The effectiveness of a method cannot be established through theoretical derivation alone. Models and algorithms must be implemented so that their behavior can be observed in data, tasks, and environments. The value of an appealing theoretical proposal can be determined only through experimental comparison and real-world operation.

Practical testing begins with explicit evaluation conditions. Researchers must determine which data to use, what comparison methods to select, and which metrics to apply. An ablation study can remove particular modules or sources of information and observe how performance changes, helping identify the source of an improvement. To test a system's boundaries, researchers may alter the data distribution, add noise, construct anomalous inputs, or examine extreme conditions.

Experimental results also require comprehensive comparative validation. Success on one dataset or a small collection of examples cannot demonstrate stable effectiveness. Researchers need to determine whether results can be reproduced across different data, parameter settings, random conditions, and operating environments, and whether comparisons with appropriate baselines are fair. Negative results and failure cases are equally informative because they reveal the conditions on which a method depends and the problems that remain unresolved.

Tests in real environments expose factors that laboratory conditions can conceal. Users may behave differently from researchers' expectations, data distributions may shift over time, system modules may interact, and improved technical metrics may fail to produce better real-world outcomes. AI systems therefore often pass through a sequence of validation stages, from offline experiments and simulation to controlled trials and actual deployment.

5.7.8 Connections among the Modes of Thought

The seven modes of thought answer different questions in AI problem solving. Problem formalization determines what the machine is actually meant to solve. Learning from data explains how regularities are formed from experience. Domain-knowledge constraints align methods with the structure of their objects. Reasoning under uncertainty enables a system to preserve appropriate boundaries of judgment when information is limited. The principle of simplicity helps researchers capture the main structure of a problem. Balancing performance and efficiency enables solutions to operate under real resource constraints. Practical testing supplies the ultimate evidence for judging whether a method works.

At a deeper level, these modes of thought express three basic characteristics of AI research. First, AI transforms abstract questions about intelligence into constructed objects that can be

represented and computed. Second, it combines data, knowledge, and environmental feedback to form solutions. Third, it constructs operational models and systems through which claims about intelligent behavior can be tested experimentally and in the real world.

The modes of thought in artificial intelligence are therefore markedly constructive and practical. Researchers establish problems through formalization, derive grounds from data and knowledge, and construct possible solutions through models, algorithms, and systems. They then test, compare, and revise those solutions under uncertainty and resource constraints. Through this repeated cycle, once-abstract problems of recognition, learning, reasoning, generation, and action become scientific problems that can be studied, realized, and validated.

The modes described here do not exhaust every idea and method in artificial intelligence. They identify core principles observed in contemporary AI research, whose defining feature is machine learning. Particular subfields may have more specific patterns of thought, and future methods may expand or revise the present account.

5.8 A Developing Science

Artificial intelligence is distinctive because it remains in a period of rapid development. Much of the knowledge in established disciplines has undergone prolonged testing. AI confronts a large number of unresolved questions, rapidly changing frontiers, and scholarly controversies for which no consensus has formed. Understanding the status of different kinds of knowledge is an important part of understanding the discipline itself.

(1) Relatively stable knowledge

Artificial intelligence has established a body of relatively stable foundational knowledge. Formal logic and computation theory provide the basis for machine processing of symbols and programs. AI can be realized through multiple methods, including search, reasoning, and learning. The performance of a machine-learning system is jointly determined by its objective, data, model, algorithm, and computational conditions. Training results require validation on independent tests. Model capabilities have boundaries defined by tasks and environments. Although this knowledge will continue to develop, it already provides a shared disciplinary foundation.

The history of AI also supports relatively stable conclusions. Symbolic AI, knowledge engineering, statistical machine learning, deep learning, and foundation models form an evolving methodological lineage organized around sources of knowledge, forms of representation, mechanisms of learning, and computational conditions.

(2) Ongoing debates

Many central questions in AI remain unsettled. Is behavior sufficient to define intelligence, and under what circumstances are internal mechanisms indispensable? How should the intelligent behavior displayed by large language models be interpreted, and how far can scaling laws extend? How should symbolic and neural methods be combined? Can brain-inspired approaches produce new breakthroughs? Could a machine possess understanding, consciousness, or subjective experience?

Positions alone cannot resolve these debates. Competing views need to define their concepts,

formulate predictions that distinguish them, and seek evidence from experiments, system behavior, and multiple disciplines. Questions that cannot yet be tested may continue to be discussed as philosophical or theoretical problems, but a conceptual proposal cannot be presented directly as an established scientific fact.

(3) Visions of the future

Concepts such as artificial general intelligence, superintelligence, machines with subjective experience, and complete brain simulation express important research aspirations and remain highly uncertain. They can help researchers frame long-term questions, but success on a present-day task cannot establish that such goals have already been attained. Equally, the fact that current systems have not attained them does not establish that every future route is impossible.

A developing science must guard against two tendencies: treating a provisional success as if it had solved the problem of intelligence, and dismissing the scientific value of existing routes because systems differ from people or exhibit obvious limitations. An attitude better aligned with the actual development of AI recognizes demonstrated capabilities on the basis of current evidence, uses experiments to characterize their boundaries, and preserves multiple avenues of inquiry for unresolved problems.

5.9 Chapter Summary

Artificial intelligence began with philosophical visions of intelligent machines and gradually became a science through the formalization of its object of study, the establishment of computational methods, the construction of models and mechanisms, and the accumulation of experimental evidence. This book defines artificial intelligence as the science of using computers to simulate intelligent human behavior. Intelligent human behavior provides the starting point and reference for research, computers provide the principal means of realization, and scientific experiments provide the means of testing. Machines may simulate intelligent behavior through internal processes different from those of humans. Behavioral performance, internal mechanisms of realization, and subjective experience are connected levels of analysis that cannot substitute for one another.

The scientific system of artificial intelligence can be organized into three interconnected levels: mathematical and computational foundations; mechanisms and methods for realizing intelligence; and problem formalization and system implementation. Together, the three levels connect formal theory, mechanisms of realization, real-world tasks, and behavioral performance. They allow the theoretical basis, operational process, conditions of applicability, and experimental results of AI to be explained level by level and tested against one another.

Within this scientific system, AI has developed distinctive modes of thought: formalizing real-world problems, learning regularities from data, attending to the constraints of domain knowledge, reasoning under uncertainty, balancing performance and efficiency, favoring simple methods where they suffice, and testing solutions through practice. AI researchers share these modes of thought, and the modes have a foundational correspondence with the aim of cultivating the capacity for cognitive autonomy in Artificial Intelligence General Education.

Artificial intelligence is also a developing science. It possesses stable structures of knowledge and established research methods while continuing to confront unresolved problems. This degree of openness is uncommon among disciplines. Later chapters will show that such openness is a productive setting for cognitive training in Artificial Intelligence General Education.

The unifying thread of this chapter is the transformation by which AI turns intelligent activity that was once abstract, implicit, and difficult to test directly into observable tasks, computable representations, constructible mechanisms, and testable systems. Intelligent activity thereby acquires a form that can be analyzed, implemented, compared, and revised. This "computationalization of intelligent activity" provides the first scientific foundation on which Artificial Intelligence General Education cultivates the capacity for cognitive autonomy. The next chapter examines how AI systems enter knowledge production, information dissemination, social judgment, and real-world action, bringing about the computationalization of social activity itself.

Chapter 6

From Computational Intelligence to Social Force: The Internal Mechanisms Through Which Artificial Intelligence Shapes Society

Abstract. This chapter examines how artificial intelligence enters society and produces enduring and wide-ranging effects. Once intelligent behaviors such as recognition, prediction, generation, and decision-making acquire computable forms, real-world objects, historical experience, and social objectives can enter AI systems. Model outputs then acquire the authority to guide action through organizational processes and institutional rules, allowing them to participate in information dissemination, resource allocation, and social decision-making. When these local effects are replicated at scale, connected across systems, and met by continuing adaptation on the part of individuals and organizations, they accumulate into relatively stable distributions of information, norms of behavior, structures of opportunity, and relations of power. Artificial intelligence thereby develops from a specific technology into an environmental condition that people collectively face as they obtain information, form understanding, and take action.

Keywords: artificial intelligence; sociotechnical systems; computational representation; modeling; computationalization of objectives; institutional embedding; feedback loops; cognitive environment

6.1 How Intelligent Behavior Becomes a Social Force

The operation of human society has always depended on intelligent activity. People acquire information through perception, accumulate experience through memory and language, form choices through reasoning, prediction, and judgment, and then change their environment through action. Education transmits knowledge, medicine assesses illness, markets allocate resources, organizations select personnel, and public institutions formulate and implement policy. Although these processes address different objects, they all involve such intelligent behaviors as acquiring

information, interpreting meaning, comparing alternatives, making decisions, and taking action. These intelligent behaviors are precisely the objects studied by artificial intelligence. The products of AI research—that is, various intelligent systems—therefore have a natural pathway into the operation of human society.

Human beings, for example, have long used writing, printing, calculating tools, and information networks to extend their capacities for memory, communication, and reasoning. The further change brought about by artificial intelligence is the transformation of relatively complex intelligent behaviors—including recognition, prediction, generation, planning, and decision-making—into computational capabilities. Machines can read text and images, form models from data, compare alternatives in light of objectives, generate new content, and invoke tools or execute actions in an environment. They can consequently enter social processes that previously depended primarily on human knowledge, experience, and judgment.

The key to this change is that intelligent behavior, once processed through the science of artificial intelligence, acquires a new form of existence. Human experience is generally bound to particular individuals and constrained by time, energy, memory, and space. Computational models can be preserved as programs and parameters, run repeatedly on different devices, and connected to databases, sensors, network platforms, and organizational processes. The same capability can serve large numbers of users within a short period and can also be embedded in many similar decision points. A method developed in a local experiment may therefore be transformed rapidly into widely deployed social infrastructure.

Artificial intelligence also has the capacity for continuous adjustment. A system can collect new data, modify its model in response to human evaluation or environmental feedback, and then return new outputs to practice. Model outputs influence individual choices and organizational actions, and these changes in turn generate data for the next round. Artificial intelligence thus gradually enters a cycle jointly constituted by computation, action, and feedback.

This process can be summarized in the central proposition of this chapter:

> Artificial intelligence transforms intelligent behavior that was once primarily bound to individuals into artificial capabilities—as distinct from human capabilities—that can be computed, replicated, connected, and deployed. These capabilities enter real-world action through organizations and institutions, accumulate through large-scale operation and social feedback, and ultimately become a force that shapes social cognition, resource allocation, and the order of social action.

This proposition also shows why the social impact of artificial intelligence cannot be deduced from model performance alone. A model can generate predictions, scores, rankings, and recommendations, but it cannot determine by itself what standing these outputs will have in society. A risk score may serve merely as supplementary information, or it may become a threshold for eligibility. A recommendation may help individuals discover content, or it may continually reshape the distribution of public attention. A generative model may support personal writing, or it may be connected to an organizational process that produces and publishes content in bulk. Even when the underlying model capability is the same, differences in deployment objectives,

authority to act, scope of coverage, and forms of feedback can lead to markedly different social outcomes.

The Artificial Intelligence Risk Management Framework of the U.S. National Institute of Standards and Technology explicitly treats artificial intelligence as a sociotechnical system. It emphasizes that AI's benefits and risks emerge from interactions among technical components, modes of use, operators, other systems, and social contexts(National Institute of Standards and Technology, 2023). Understanding how artificial intelligence produces social effects therefore requires tracing a complete chain of influence: how social activity enters computational systems, how computational results enter real-world action, how local effects are amplified and accumulated, and how these processes together transform the cognitive environment in which people live.

In this chapter, the "computationalization of social activity" encompasses the complete process through which social activity enters computational systems and those systems acquire an enduring capacity to mediate it. Examining this process serves two purposes. On the one hand, it supplements the concept of the "intelligent society" with an account of its formative mechanisms, deepens reasoned understanding of the "complex cognitive environment of the intelligent society," and thereby sharpens recognition of the responsibility to "cultivate the capacity for cognitive autonomy." On the other hand, it provides conceptual background and curricular material for AIGE curriculum design.

6.2 How Social Activity Becomes Computable: The Basic Transformations of Artificial Intelligence

Before artificial intelligence can participate in education, health care, employment, commerce, and public governance, these social activities must first be transformed into forms that machines can process. Human states must become data, past experience must become models, social objectives must become metrics, and existing work must become workflows jointly carried out by humans and AI. Only through these transformations can artificial intelligence read information, make judgments, select outcomes, and influence action. The social effects of artificial intelligence arise from these very processes.

6.2.1 Real-World Objects Become Data: What the System Can See

The real world is continuous, complex, and saturated with specific contexts, whereas computers can process only information that has been explicitly represented. The first step by which artificial intelligence enters social activity is to select recordable features and transform real-world objects into data.

A person's learning state can be represented by response records, estimated probabilities of mastery, and grade levels. Health status can be represented by medical images, test values, symptom codes, and risk scores. Interests can be represented by clicks, dwell time, searches, and purchases. Work-related capability can be represented by educational background, employment history, task performance, and evaluation results. Through this transformation, objects that could not readily be compared become variables, categories, relations, vectors, and metrics.

Data make it possible to aggregate, compare, and analyze social phenomena. Teachers can identify difficulties encountered by students from large bodies of learning records; physicians can draw on more cases when making diagnoses; and public institutions can identify needs and risks more quickly. Many AI capabilities rest on this foundation of datafication.

Data, however, preserve only part of reality. A system can process what it records; content that is not recorded will generally be absent from subsequent computation. A click can be recorded, but the hesitation preceding it is difficult to capture. An examination score can be quantified, but a single number cannot fully express a student's curiosity, effort, or potential for growth.

Data are therefore not reality itself, but a selective description of reality. The choice of what to record effectively determines what the system can see. Using click counts to represent interest, examination scores to represent learning, or features of a résumé to represent work-related capability means using measurable signals to stand in for more complex social concepts. Such signals are generally called proxy variables.

Proxy variables enable systems to process abstract problems, but they may also compress complex concepts too severely. Repeated clicks on a particular type of content may indicate preference, but they may also result from curiosity, anxiety, or chance exposure. A low examination score may reflect inadequate mastery, but it may also be affected by language, health, or family circumstances. When a system cannot see these differences, its judgment can rest only on a limited representation.

Data representation therefore has direct social consequences. The information recorded by a credit system may affect who obtains a loan. The fields included in a résumé-screening system may affect which experiences receive recognition. The metrics adopted by an education system may affect which students are identified as needing assistance. Understanding an AI system must therefore begin with three questions: What data does it use? What do those data represent? What do they omit?

6.2.2 Historical Experience Becomes a Model: How the System Makes Judgments

AI models generally learn patterns from past data and then use those patterns to assess current situations. Once a model has been trained, past experience is condensed into a set of computational rules that can be run repeatedly.

This capability expands the range of experience available to human beings. Physicians can draw on large numbers of historical cases, teachers can identify change in long-term learning records, scientists can search for patterns in high-dimensional experimental data, and institutions can take preventive action on the basis of earlier risk signals. Vast amounts of information that no individual could read and compare item by item can inform a particular judgment through a model.

The difficulty is that historical data contain not only experience but also the social conditions under which the data were produced. Who had the opportunity to enter the sample, which phenomena received attention, which outcomes were recorded as successes, and which judgments

were treated as correct labels all affect the patterns a model learns. If, for example, a particular group historically had fewer educational or employment opportunities, its members may also have fewer records of success in historical data. A model may learn and perpetuate the judgment produced under those earlier conditions, concluding that members of the same group also have a lower chance of success under present conditions.

Models must also confront the distinction between correlation and causation. A feature's usefulness in predicting an outcome establishes only that it has a statistical association with the outcome, not that it caused the outcome. A person's place of residence, for example, may help predict credit risk, while the actual causes may involve income, public resources, employment opportunities, or historical regional disparities. A system that acts directly on the residential variable may mistake a surface indicator for the true cause.

A further limitation arises when group-level patterns are used to infer the situation of an individual. On the basis of a group's history, a model can estimate the likelihood that a particular person will experience a given outcome. It cannot, however, fully understand that person's specific experience, present condition, or future potential, and its judgment may therefore be biased.

When predictions enter resource allocation, they can exert self-reinforcing effects on reality. A student assessed as having low learning ability may consequently be denied more challenging courses; with fewer opportunities, the student's later performance may indeed decline. A person assessed as posing high credit risk may find it harder to obtain financing; lack of funds may then increase that risk further. Models can thus do more than predict the future: they may participate in producing it.

The social use of models consequently demands close attention to the following questions: Under what conditions were the historical data produced? Has the model learned causes or merely correlational cues? To what extent can patterns at the group level be applied to a particular individual?

6.2.3 Social Objectives Become Metrics: What the System Pursues

An AI system must also be told which outcomes are better. A classification model requires a definition of the correct answer; a recommendation system requires a rule for ranking content; reinforcement learning requires a reward; and an organization must decide whether a system should prioritize greater efficiency, lower costs, or improvements in quality, safety, and fairness. Only when social objectives are translated into computable metrics can a system compare outcomes and select actions.

Real-world objectives generally encompass multiple values. Education is concerned with mastery of knowledge, but also with interest, character, creativity, and long-term development. Health care is concerned with diagnostic efficiency, but also with patient safety, dignity, informed consent, and equitable access. Public governance must consider effectiveness, procedure, fairness, rights, and social trust at the same time. Computational systems cannot easily process the full meaning of all these objectives and generally must select some aspects and represent them through measurable indicators.

Metrics make objectives explicit, but they may also narrow people's understanding of those objectives. Click-through rate can indicate whether users open content, but it cannot fully represent whether that content has genuine value. Task-completion time can measure efficiency, but by itself it says nothing definitive about the quality of the work. Short-term grades can reflect one part of learning outcomes, but they cannot adequately represent a student's long-term development.

Once a metric becomes an optimization objective, both systems and people gradually adjust their behavior around it. A recommendation system that continually optimizes click-through rate may favor content that captures attention more readily. If a school places excessive emphasis on examination scores, teachers and students may devote more time to test preparation. If a platform evaluates only the number of completed tasks, workers may reduce the attention they give to complex cases. A metric originally used to observe an objective can thus feed back into and transform the activity being observed.

The more capable artificial intelligence becomes, the more consequential the definition of objectives becomes. More data, larger models, and greater computing power can enhance the pursuit of a specified objective, but they cannot automatically restore values that were never incorporated into that objective. Problem formulation requires translating high-level social objectives into specific tasks, target variables, and proxy metrics. This process entails extensive judgment and choice and may produce radically different ethical consequences(Passi and Barocas, 2019).

6.3 How Computational Results Enter Real-World Action: Organizational and Institutional Embedding

For AI systems to produce widespread real-world effects, they must be integrated into specific organizational processes and institutional arrangements. Organizational and institutional embedding therefore connects computational capability with social action and determines the direction, intensity, and boundaries of its effects.

The same medical prediction model can be used to remind physicians to order additional tests or to restrict insurance coverage. The same type of generative model can help learners understand materials or can be organized into a system for producing false content at scale. Even when model capabilities are identical, differences in purpose, position within a workflow, authority to act, and avenues of correction can lead to markedly different social outcomes. All of these factors depend on the constraints imposed by organizational processes and institutional design.

6.3.1 Organizations and Institutions Give Model Outputs Social Efficacy

Model outputs acquire social efficacy only through a series of concrete arrangements: who proposes the need for a system, where it is placed in a workflow, which people or groups its outputs concern, whether professionals are required to follow them, which tools the system is permitted to invoke, how results are recorded, and how errors are identified and corrected.

Interface design, default options, and time limits for action can also affect institutional efficacy. Even where human decision-making is nominally retained, a model output may become the de facto decision if the interface displays only a seemingly precise score and the workflow demands rapid confirmation.

Model performance is therefore only one condition of social impact. Two systems with similar levels of accuracy pose different social risks and governance requirements if one merely provides interpretable supplementary information while the other directly controls access to important resources. Evaluating artificial intelligence requires examining both what a model can do and what an organization permits it to do.

6.3.2 The Redistribution of Power to Act across the System Chain

Once an AI model enters organizational processes and institutional arrangements, it redistributes the power to observe, judge, decide, and contest. Those who translate reality into data possess the power to define categories and metrics. Organizations that train and deploy models possess the capacity to turn particular patterns into action. Actors that control entry points to platforms can also decide which information appears first. The people affected often see only the result; they have difficulty learning about the data sources, model assumptions, and institutional rules, and still greater difficulty changing them.

This asymmetry affects people's ability to contest decisions. The complexity of AI systems can make their outputs appear to be technical conclusions, concealing contestable classifications, proxy metrics, risk thresholds, and value tradeoffs within data and processes. Restoring room for deliberation requires unfolding a technical decision into its components: Which parts derive from factual evidence? Which derive from model assumptions? Which are organizational choices? Which involve contestable social values?

Organizations and institutions also determine who gains new capabilities from artificial intelligence and who is primarily subjected to evaluation and constraint by systems. Those who possess data, models, and access to deployment can use AI systems to extend their powers of observation, prediction, and action. Those who lack corresponding resources may appear chiefly as objects to be recorded, classified, and decided upon. Any discussion of changes in power brought about by artificial intelligence must therefore examine how capacities are extended, how the authority to act is distributed, and whether the people affected have genuine channels through which to obtain information and explanations, refuse decisions, seek review, and appeal.

6.3.3 The Reorganization of Responsibility along the System Chain

Artificial intelligence lengthens and divides the process from cognition to action, distributing the production of outcomes across data, models, interfaces, organizational rules, and situated operation. Data providers, model developers, product designers, deploying institutions, professionals, and end users each possess some of the relevant information and control some part of the process. A particular consequence may arise jointly from biases in data, model boundaries, interface presentation, organizational rules, and situated operation. Responsibility therefore

changes from a question of attribution to a single actor into a problem of organization across the system chain.

Responsibility must accordingly be understood along the system chain and aligned with the components that each participant can actually foresee, control, and correct. Developers cannot anticipate and control every specific use of a system in every setting, while frontline personnel cannot remedy all the limitations already embedded in data, models, and organizational processes. The separation between the attribution of responsibility and the actual capacity for control is an important structural problem that arises when artificial intelligence enters organizations and institutions.

Organizational and institutional embedding thus completes a crucial transformation: it turns possibilities generated by computational systems into the presentation of information, allocation of resources, and outcomes of action in the real world. Institutional authorization, however, still operates primarily within particular organizations and tasks. For artificial intelligence to produce broad and enduring social effects, these local effects must also be replicated repeatedly, connected with one another, and incorporated into continuing social feedback.

6.4 From Local Effects to Structural Impact: Scale, Connection, and Feedback

The preceding section explained how model outputs acquire real-world efficacy within particular institutions. This section goes further: Why can a single act of classification, recommendation, prediction, or generation gradually transform the wider social environment? The answer lies in the low cost of replicating computational systems, the capacity of different systems to connect through data and platforms, and the tendency of individuals and organizations to adjust their behavior in response to system outputs. Scale expands the reach of an effect; connection allows effects to propagate across systems; and feedback enables systems to participate in producing the next round of data. Together, these three mechanisms allow local effects to accumulate into structural impact.

6.4.1 Replication at Scale Turns Local Rules into Common Conditions

Replication at scale transforms a local model's capabilities, standards, and errors into conditions shared by large numbers of people. Once software and models have been developed and connected to infrastructure, the marginal cost of replicating and extending them is generally lower than that of expanding a human workforce to the same scale. The same scoring rule, recommendation model, or generative service can therefore enter many regions, institutions, and areas of life. Scaling brings consistency and efficiency, enabling scarce knowledge and services to reach more people; it can also amplify a model's omissions and biases in parallel. An individual's occasional judgment generally has limited influence, whereas a widely adopted computational rule may operate repeatedly across large numbers of similar decisions.

Scale also produces standardization. The use of common data structures, model interfaces, and evaluation methods across institutions can improve coordination, but social objects may

also be reorganized gradually into forms that systems can process more readily. Job candidates learn how to write résumés that are more likely to pass screening; content creators adjust titles to suit recommendation mechanisms; and schools and institutions organize activities around quantifiable metrics. Technical standards may thereby become behavioral standards.

Concentration increases further when a small number of models become the foundation for many products and services. Downstream systems may appear to serve very different purposes while sharing similar data, model capabilities, and default assumptions. An improvement in one capability can diffuse rapidly, but a defect in one system may also propagate across many applications. Society's dependence on a small number of foundation models, computing platforms, and data resources can also transform the distribution of power in innovation, market structure, and technology governance.

Artificial intelligence can use individual data to generate different recommendations, explanations, and responses for each person. From an individual perspective, personalization can reduce the cost of searching for information and align content more closely with particular needs. From a social perspective, large numbers of separately personalized choices jointly reshape public attention, issue visibility, and the direction of cultural production. People may share fewer common experiences because each sees different content, while a single ranking mechanism may simultaneously concentrate a large share of attention on a small number of objects. Personalization and concentration can occur at the same time.

Generative models extend scale effects to the supply of content. Systems can rapidly generate large volumes of customized text, images, and audio, so that scarcity of information increasingly gives way to scarcity of attention and trustworthy sources. Growth in the quantity of content does not automatically produce growth in knowledge. Whether content can be traced to its origin, whether evidence can be verified, and whether important views can be discussed collectively in the public sphere become important criteria for evaluating the information environment as a whole.

6.4.2 System Connections Allow Effects to Propagate across Contexts

System connections allow the output of one process to become the input of another, enabling local rules to continue operating across platforms, institutions, and domains of life. Behavioral data accumulated by one platform may be used in advertising, credit assessment, or content recommendation; a single score may be adopted repeatedly by multiple institutions; and generated content may enter search results, knowledge bases, model training, and subsequent rounds of generation. Each system performs its own task, but once connected, the systems jointly constitute a continuous chain of influence.

Connection can improve continuity of service and reduce repeated data collection and informational fragmentation. It can also make it harder to confine errors, omissions, and existing biases to a single stage. Features not recorded by an upstream system generally cannot be recovered downstream. Labels produced by early classifications may gradually be treated as stable facts as they pass repeatedly through other systems. After being cited by multiple systems, a model output may acquire credibility far beyond that warranted by the original evidence.

When systems share data, foundation models, and evaluation standards, apparently independent decisions may have common origins. The search, recommendation, evaluation, and service systems encountered by an individual may be provided by different institutions while relying on similar data structures and model capabilities. The resulting influence cannot be eliminated entirely by correcting one output. It requires tracing how data and judgments propagate along the system chain.

Connection also changes the meaning of local exit. An individual may stop using a particular application yet remain subject to related scores, recommendations, and generated content adopted by other institutions. An organization may replace a front-end product while continuing to depend on the same foundation model or data source beneath it. The social effects of artificial intelligence thereby extend beyond the boundaries of any single product and become environmental conditions jointly maintained by multiple systems.

6.4.3 Social Feedback Enables Systems to Participate in Shaping Reality

Social feedback turns predictions, recommendations, and evaluations into interventions that change their objects, enabling systems to participate in producing the next data they will observe. Conventional prediction often treats its object as external: a model observes the world and estimates what will occur. In social systems, however, predictions enter action, and action changes the future. Credit-risk predictions affect the terms of loans, which in turn affect borrowers' capacity to repay. Predictions of learning risk affect the allocation of educational resources, and changes in resources affect later performance. Assessments of crime risk alter where patrols are deployed, and patrol locations determine which events are more likely to be recorded. Predictions thus participate in producing the next round of data to be predicted.

This phenomenon is known as "performative prediction": deployment of a model changes the data distribution, and future outcomes are affected by earlier predictions and the actions they induced(Perdomo et al., 2020). Under these conditions, model updating cannot be understood merely as using new data to keep pace with a world that is changing independently. The new data already contain the effects of the system itself.

Recommendation systems create similar loops. A system recommends content on the basis of past behavior; the recommendation affects what users encounter; and those new encounters produce subsequent training data. Research shows that this form of algorithmic confounding can homogenize user behavior without a corresponding improvement in utility(Chaney, Stewart, and Engelhardt, 2018). Each step in the loop may have a reasonable explanation, yet the overall direction of its long-term accumulation requires separate evaluation.

Social actors also adapt to systems deliberately. Once people know which metrics a system uses, they change their behavior to obtain better results, and organizations reallocate resources around rules of evaluation. Moderate adaptation can help people understand rules and improve their performance. When proxy metrics diverge from actual objectives, however, behavior increasingly concentrates on raising the metric, crowding out the original objective.

Such responses can also weaken the patterns on which a model relied. Before a variable is used for evaluation, it may serve as a stable indicator of a particular state. Once it becomes an

explicit target, people intervene upon it, changing the relationship between the variable and the original state. Associations learned from historical data may fail because the model itself has been deployed. Social feedback thus becomes an endogenous source of distribution shift.

6.5 How an AI-Mediated Cognitive Environment Takes Shape

The cognitive environment is the aggregate outcome produced by the long-term operation of the foregoing mechanisms. It is jointly shaped by computational representation, institutional embedding, replication at scale, system connections, and social feedback. Within this environment, artificial intelligence comes to stand between people and the real world, and between cognition and action. Through continuous operation across many social contexts, it becomes a condition that people cannot readily circumvent as they obtain information, form judgments, and take action.

6.5.1 Artificial Intelligence Enters between People and the Real World

Human beings have always relied on language, media, systems of knowledge, and social institutions to understand the world. Once artificial intelligence joins this mediating process, an increasing share of the information that reaches people has passed through data selection, model processing, algorithmic ranking, and generative systems. Search systems determine which results appear first; recommendation systems organize the content available to individuals; generative models directly provide explanations, summaries, images, and plans; and classification and prediction systems attach labels, probabilities, and levels of risk to social objects.

This mediation expands the range of information and knowledge available to individuals while also changing how reality is presented. The content users encounter already incorporates the combined effects of data coverage, classification rules, model assumptions, optimization objectives, and platform strategies. What can be seen, how it is seen, which differences are emphasized, and which sources receive greater circulation are increasingly shaped by computational systems.

Model-generated content also reenters public circulation. Artificial intelligence not only filters existing information but also participates directly in producing new text, images, audio, and video. Sources, evidence, inference, and expression may be compressed into a single complete answer during generation. The path by which knowledge reaches people therefore becomes shorter, while the process by which it was formed may become harder to see. Artificial intelligence thus becomes part of both the information environment and the knowledge environment.

6.5.2 Artificial Intelligence Connects Cognition and Action

AI-mediated content does not remain confined to cognition. Model outputs enter education, health care, employment, finance, consumption, and public governance through organizations and institutions. They influence how resources are allocated, opportunities distributed, risks handled, and actions selected. Systems thus describe and predict reality on the one hand, while participating through real-world action in shaping the very objects they describe and predict on the other.

This shortens the distance between cognition and action. A classification result can directly alter the presentation of information; a prediction score can immediately trigger review; generated content can be published automatically through a platform; and an agent can invoke tools to carry out a sequence of operations. The more authority computational results possess to initiate action, the more directly models shape reality.

Individuals and organizations anticipate these effects and adjust their behavior. People learn how to pose questions to systems, how to make résumés more likely to pass screening, and how to produce content more likely to be recommended. Institutions redesign processes around computable metrics, while professionals reallocate attention in response to system outputs. Artificial intelligence thus participates in shaping human action while being continually changed by human adaptation.

6.5.3 Systems in Continuous Operation Become the Cognitive Environment of the Intelligent Society

When large numbers of AI systems operate continuously, their combined effects become a cognitive environment shared by members of society. This environment is jointly constituted by data, models, platforms, organizations, and institutions. Data determine which aspects of reality can enter computation; models transform historical patterns into present capabilities; objectives and metrics specify the direction of system action; platforms provide structures for connection and distribution; organizations and institutions give outputs real-world efficacy; and human responses generate new data and rules. No single model can constitute this environment on its own. It emerges and changes through the continuing connection of multiple elements.

The changes in the cognitive environment summarized in Chapter 2 thus acquire concrete formative mechanisms. Data selection, model processing, and algorithmic ranking place artificial intelligence between people and reality. Generative and distribution systems expand the supply of information and reshape the distribution of attention. Connecting model outputs to workflows allows some actions to occur automatically. Human responses become system data once again, creating continuing feedback. The joint participation of data, models, platforms, organizations, and users in producing outcomes also distributes responsibility across a longer system chain.

Within such an environment, people may be affected by algorithmic ranking, automated evaluation, generated content, and institutional decision-making even if they do not directly use a particular AI product. Individuals may find it difficult to recognize precisely how extensively their sources of information, modes of judgment, and opportunities for action are shaped by systems. The intelligent society therefore raises a fundamental question: while depending extensively on artificial intelligence, how can people continue to inquire into the sources of information, assess the strength of evidence, recognize model boundaries, calibrate their trust in systems, and take responsibility for the actions that ultimately follow?

This is the question to which cognitive autonomy responds. Artificial Intelligence General Education must help learners understand these mechanisms and retain command of the cognitive process as they use, evaluate, and, when necessary, take over from AI systems. The next chapter examines further how this education can transform the study of artificial intelligence into training

in the capacity for cognitive autonomy.

6.6 Chapter Summary

This chapter has examined how the computational capabilities of artificial intelligence enter the operation of society and gradually become sources of wide-ranging social impact. Participation by artificial intelligence in social activity first requires a series of basic transformations: real-world objects must be represented as data, past experience condensed into models, social objectives translated into metrics, and the acquisition, analysis, and judgment of information, together with action, organized into workflows in which humans and AI jointly participate. Data determine what a system can see; models determine the grounds on which it makes judgments; objectives determine what it pursues; and workflows determine how computational results enter real-world action.

The social effects of artificial intelligence also depend on the position it is given within organizations and institutions. Model outputs may serve only as references, or they may be used to screen applicants, allocate resources, adjust opportunities, and execute actions directly. Organizational processes and institutional arrangements give computational results real-world efficacy. They also determine who can question, revise, or overturn a system's judgment and who bears responsibility when errors occur. The social impact of artificial intelligence is therefore already present throughout the entire process of data selection, model training, objective setting, workflow design, and institutional use.

When artificial intelligence is replicated at scale and interconnected through platforms, organizations, and infrastructure, the local effects of individual systems accumulate further. System outputs shape human understanding and action; human responses then become new data, driving further changes in models, rules, and behavior. Expansion in scale, system connections, and feedback loops jointly transform artificial intelligence from a tool for solving particular tasks into a structural force that continually shapes information dissemination, judgment formation, the distribution of opportunity, and social action.

The repeated operation of large numbers of systems ultimately constitutes the cognitive environment of the intelligent society. What information people can encounter, which sources they trust, the grounds on which they make judgments, and the opportunities for action available to them may all be shaped by this environment. Under these conditions, the individual's ability to actively regulate their use of evidence, allocation of trust, formation of judgments, and choice of actions becomes a question that Artificial Intelligence General Education must address. The next chapter examines how AIGE can transform understanding of these mechanisms into training in the capacity for cognitive autonomy.

Chapter 7

Cultivating the Capacity for Cognitive Autonomy through Artificial Intelligence General Education

Abstract. Cognitive autonomy becomes an implementable educational objective only when it is translated into a clearly defined structure of capacities, curricular tasks, and evidence for assessment. This chapter proposes a four-capacity model—induction, generalization, judgment, and self-awareness—to describe the fundamental cognitive capacities that Artificial Intelligence General Education seeks to cultivate. Induction enables learners to form evidence-based, testable understandings from limited materials. Generalization enables them to transfer existing understandings to new contexts while identifying the conditions and boundaries of their applicability. Judgment enables them to make reasoned choices when evidence is incomplete, consequences are uncertain, and values conflict. Self-awareness enables them to monitor their own understanding, trust, dependence, and judgment, and to revise them in light of new evidence.

The four-capacity model is an educational model designed for curriculum development and learning assessment; it does not attempt to exhaust the psychological structure of cognitive autonomy. It draws on research into human learning, transfer, judgment, and metacognition, while also relating closely to the processes through which AI systems learn patterns from data, produce outputs in new contexts, undergo evaluation, and enter social action. On the basis of these characteristics of artificial intelligence, the chapter identifies three basic pathways for cultivating cognitive autonomy: understanding intelligent mechanisms, so that learners recognize the conditions, grounds, and boundaries under which computational conclusions are formed; participating in computational construction, so that learners' own assumptions and judgments are made explicit and subjected to operational testing; and analyzing social embedding, so that learners can identify the values, authority, effects, and responsibilities involved when computational systems enter real-world activities.

The chapter examines the theoretical basis, educational objectives, classroom tasks, and levels of performance associated with each capacity, and organizes the four into a cycle that proceeds from hypothesis formation and boundary testing to choice of action

and cognitive correction. Assessing the capacity for cognitive autonomy requires looking beyond the completeness of a product or the performance of a model. It requires evidence of whether learners can explain evidence, identify boundaries, weigh risks, organize human–AI collaboration, and reflect on and revise their own cognitive processes.

Keywords: Artificial Intelligence General Education; cognitive autonomy; induction; generalization; judgment; self-awareness; metacognition; trust calibration; learning assessment

7.1 Framing the Problem: From Educational Ideal to Capacity Model

Part I established the book's basic argument. AI systems and their platform-based applications are changing the fundamental conditions of knowledge production, information dissemination, learning support, the organization of work, and public governance. The principal cognitive difficulty confronting individuals in an intelligent society has shifted from a scarcity of information to an abundance of information from mixed sources; from difficulty obtaining conclusions to the ready presentation of conclusions whose grounds must be scrutinized; and from reliance on a single author and medium to participation in a hybrid knowledge environment jointly constituted by people, models, platforms, institutions, and communication networks. Within this environment, Artificial Intelligence General Education assumes an important task: cultivating learners' capacity for cognitive autonomy.

As an educational ideal, cognitive autonomy must still be translated into implementable educational content. An ideal can enter teaching practice only when it takes the form of a capacity structure, curricular tasks, and evidence for assessment. If a course merely urges students to "maintain independent judgment" or "use AI rationally," teachers will have little basis on which to design instruction, and students will find it difficult to identify the cognitive activities they need to practice. AIGE must therefore answer a more specific set of questions. Which relatively stable cognitive capacities should learners develop? How are these capacities related? How does each manifest itself in AI contexts and in broader social contexts? How can teachers observe whether students have progressed? And how can schools determine whether a course has moved beyond tool demonstrations and product showcases to engage learners in genuine training for cognitive autonomy?

This chapter proposes a four-capacity model comprising induction, generalization, judgment, and self-awareness. Within the curricular scope of AIGE, learners who develop relatively stable capacities in these four areas may be regarded as having acquired a degree of capacity for cognitive autonomy.

- **Inductive capacity** is the capacity to form testable understandings from a limited set of phenomena, data, cases, texts, and experiences. It requires learners to discover patterns while recognizing the limitations of samples, evidence, and causal explanations.
- **The capacity for generalization** is the capacity to transfer existing understandings to new contexts and identify their boundaries of applicability. It requires learners to understand the conditions on which a pattern depends and to adjust their judgment and

trust when the object, environment, or task changes.

- **The capacity for judgment** is the capacity to make explicable choices when evidence is incomplete, consequences are uncertain, and values conflict. It requires learners to consider facts, methods, risks, values, and responsibilities together.
- **The capacity for self-awareness** is the capacity to monitor, reflect on, and correct one's own cognitive state. It requires learners to recognize when their understanding is inadequate, when their trust is excessive, and when their thinking is being influenced by emotion, authority, platform mechanisms, or AI outputs.

The four-capacity model should be understood as a curricular model for AIGE, not as an exhaustive taxonomy of the psychological structure of cognitive autonomy. These four capacities constitute the core content of the training advocated by AIGE, but this does not imply that cognitive autonomy consists of only four psychological components. Scholarship has not yet produced a single accepted account of the psychological constitution of cognitive autonomy, whereas curriculum practice requires a reasonably explicit system of capacities that can inform instructional design and learning assessment. The four-capacity model serves precisely this educational function.

This curricular model has foundations in psychology and the learning sciences. Research in cognitive psychology shows that people can extract regularities from limited materials, but are also susceptible to biases arising from samples, forms of representation, and heuristics (Shepard, 1987; Saffran, Aslin, and Newport, 1996; Tversky and Kahneman, 1974; Nisbett et al., 1983). The learning sciences further show that knowledge transfer depends on deep understanding, active abstraction, and metacognitive monitoring (Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988; Chi, Feltovich, and Glaser, 1981). Research in neuroscience and metacognition suggests that task performance, structural generalization, and awareness of one's own performance are interrelated, yet rest on psychological and neural foundations that can be distinguished from one another (Samborska et al., 2022; Fleming and Dolan, 2012). Taken together, these findings indicate that individuals require sustained training in pattern formation, knowledge transfer, complex judgment, and self-monitoring if they are gradually to establish cognitive autonomy in complex environments.

The four-capacity model is also closely related to the basic characteristics and operating processes of AI systems. Such systems learn patterns from data, apply those patterns to new inputs, and are continually tested through evaluation, feedback, and the consequences of their use. Their confidence, errors, boundaries of applicability, and potential risks must in turn be interpreted and judged by human beings. These processes place new demands on human induction, generalization, judgment, and self-awareness. At the same time, they render cognitive questions concerning samples, features, patterns, boundaries, evaluation, and trust relatively explicit, providing observable, operable, and comparable contexts in which the four capacities can be practiced.

7.2 How Artificial Intelligence General Education Cultivates Cognitive Autonomy

The preceding section specified the capacities AIGE needs to cultivate. This section explains why AIGE is able to support their development.

The previous two chapters examined the computationalization of intelligent activity and of social activity, respectively. Artificial intelligence organizes data, representations, models, objectives, evaluation, and feedback into processes that can be analyzed and executed. Through platforms, organizations, institutions, and workflows, it also enters information dissemination, the formation of judgments, and social action. These characteristics enable AIGE to cultivate cognitive autonomy along three paths: understanding intelligent mechanisms, and thereby recognizing the conditions, grounds, and boundaries under which computational results are formed; participating in computational construction, so that learners' own assumptions and judgments are made explicit and subjected to operational testing; and analyzing social embedding, so as to identify the values, authority, effects, and responsibilities involved when computational systems enter real-world activities.

These three paths describe the basic ways in which AIGE cultivates cognitive autonomy, whereas induction, generalization, judgment, and self-awareness describe the capacities that this training is intended to develop. They occupy different conceptual levels and do not correspond one-to-one. Understanding intelligent mechanisms, participating in computational construction, and analyzing social embedding may each engage several cognitive capacities at once; any one of the four capacities can also be practiced repeatedly along all three paths.

7.2.1 Understanding Intelligent Mechanisms: Recognizing the Conditions under Which Conclusions Are Formed

AI systems must convert real-world objects into data and representations, process them through models to produce outputs, and adjust those outputs in accordance with objectives and evaluation criteria. Learning about these mechanisms allows learners to see how a computational conclusion is formed: which data the system used; which aspects of reality the data preserved and which they omitted; which features the model relied on in making a determination; what kinds of result the objective function encouraged; and what the evaluation metrics actually measured.

This learning provides a clear structure of questions for analyzing cognition more generally. Faced with a conclusion that has already been produced, learners can ask whether the materials are representative, whether the observed features provide sufficient support for the judgment, whether a correlation warrants a causal explanation, whether the evaluation criteria capture the goals that truly matter, and whether the conclusion will continue to hold for new objects and in new contexts. The errors, biases, and uncertainties exhibited by AI systems also give concrete form to problems such as insufficient evidence, skewed samples, reliance on superficial cues, and limited scope of applicability.

The computational processes of artificial intelligence and the cognitive processes of human beings are implemented in different ways, but both confront certain fundamental epistemic

problems: how to form conclusions from limited materials, how to apply existing understanding to new contexts, how to evaluate the strength of evidence, and how to recognize the boundaries within which a conclusion applies. Artificial intelligence provides a point of reference at the level of problem structure; this argument does not require the assumption that machines replicate human cognition.

By analyzing how AI results are formed, learners can gradually develop the habit of tracing sources, examining grounds, identifying boundaries, and calibrating trust. The central aim is that, even when confronted with a conclusion that is readily presented, fluently expressed, or apparently credible, learners will continue to ask under what conditions it was formed and decide, on the basis of evidence, how much trust it warrants.

7.2.2 Participating in Computational Construction: Subjecting Understanding to Operational Tests

Artificial intelligence is a distinctly constructive science. An idea must be translated into choices about data, feature representations, category divisions, rules, objectives, evaluation metrics, and operational procedures before it can become an executable and testable system. To participate in this process, learners must turn an initially general understanding into a series of explicit choices and explain the reasons on which those choices rest.

Computational construction in this sense is not limited to developing complex AI models. Organizing a dataset for classification, designing a set of decision rules, comparing the results produced by different prompts, defining evaluation metrics for a task, or constructing a simple human–AI workflow can all take learners through the process from understanding a problem to implementing a system. "Construction" here means an approach to solving problems by building an executable process.

When designing a classification task, for example, learners must decide which objects may be grouped in the same category, which features will serve as grounds for classification, and how to handle objects at category boundaries. When defining evaluation metrics, they must determine what counts as success and what consequences different kinds of error will have. When designing prompts and workflows, they must also discover omissions, ambiguities, and hidden assumptions in their own formulations. An understanding that previously remained at the level of language is thus transformed into a set of choices that must actually be executed.

Running the system reveals the consequences of those choices. Skewed data may lead the system to form erroneous patterns; one-sided features may encourage reliance on superficial cues; inappropriate metrics may reward behavior that diverges from the real objective; and rules that work locally may fail in new contexts. Learners must use the operational results to reconsider the original problem, examine their assumptions, explain the causes of error, and revise the data, rules, objectives, or method of evaluation.

This process brings induction, generalization, judgment, and self-awareness into a continuous cognitive practice. Learners discover and abstract patterns from materials, test whether those patterns hold across contexts, evaluate solutions in light of operational results, and reflect on their own choices and confidence. The educational value of computational construction lies in

bringing understanding from general expression into an executable state, so that the assumptions, omissions, and boundaries hidden within an idea are tested against actual results.

7.2.3 Analyzing Social Embedding: Judging the Real-World Effects of Computational Results

The effects of artificial intelligence acquire real-world force as systems enter social processes. Through platforms, organizations, institutions, and workflows, data, models, and metrics enter real activities and consequently affect information dissemination, personnel evaluation, resource allocation, access to opportunities, and social action. The same model output may serve as an advisory reference or as the direct basis for automated screening, risk ratings, or the execution of action. Its real-world significance depends on the authority granted to the system and on how people and organizations interpret and use its output.

Through cases involving recommender systems, AI-assisted recruitment, medical decision support, educational assessment, financial risk control, and public governance, AIGE can guide learners in analyzing how computational systems become embedded in social processes. This analysis requires tracing a reasonably complete chain of effects: how real-world objects are represented as data; how historical experience enters a model; how social objectives are converted into metrics; how model outputs enter organizational procedures; which people are affected by the results; and who has the authority to object, revise a decision, and assume responsibility.

By following this chain, learners can see that choices about data, objectives, and evaluation within a technical system have tangible consequences in social operation. Tensions may arise among computational efficiency, predictive accuracy, and social reasonableness. A system that performs well on technical metrics may still produce unreasonable consequences because its objectives, scope of application, operating authority, or allocation of responsibility are inappropriate.

This analysis extends cognitive autonomy from the individual's appraisal of information to an understanding of the cognitive environment as a whole. Learners must judge what the system is optimizing, what it leaves out, how it affects different groups, and what limits should constrain automated action. Their judgment concerns not only whether a result is accurate, but also whether the objective is reasonable, the procedure fair, the authority appropriate, and the assignment of responsibility clear. Learners thereby practice connecting technical evidence, social values, and the consequences of action in order to form evidence-based evaluations of computational systems.

Understanding intelligent mechanisms, participating in computational construction, and analyzing social embedding address, respectively, the formation of computational conclusions, the operational testing of cognitive assumptions, and the real-world effects of computational results. The first path helps learners discern the sources, grounds, and boundaries of conclusions. The second allows them to experience firsthand the explication, operationalization, and revision of understanding. The third enables them to judge how computational results acquire real-world force and to question the values, authority, and responsibilities involved. Together, the three paths connect the internal workings of AI systems, learners' cognitive practice, and the social

environment in which artificial intelligence operates.

7.3 Induction: Forming Testable Understandings from Limited Materials

7.3.1 The Basic Meaning of Induction

Induction is the starting point of learning. Human beings cannot wait until all possible evidence has been assembled before forming an understanding, nor can they gather all relevant information afresh before every action. Learners must form preliminary understandings from limited samples, local experience, a small number of cases, fragments of text, visual cues, and the testimony of others. Inductive capacity enables people to discover regularities, establish categories, propose explanations, and generate hypotheses in complex environments, providing a starting point for further learning and judgment.

Induction is inherently uncertain. A limited sample may support a conclusion, but rarely proves it conclusively. A statistical relationship may indicate a pattern, but does not automatically supply a causal explanation. Local experience may guide action, but may also produce sweeping generalizations. Training in induction in the age of AI therefore has two related dimensions: students must learn both to discover patterns actively and to recognize the evidence on which those patterns rest, the degree of support they enjoy, and their limitations.

Artificial intelligence provides clear material through which to observe induction. Machine-learning models learn patterns from training data; image classifiers extract features from large collections of images; language models learn linguistic regularities from text sequences; and recommender systems infer preferences from user behavior. By observing these systems, students can see how samples, labels, features, objective functions, and feedback shape the inductive process. More importantly, they can observe concrete modes of inductive failure: skewed data may produce model bias, noisy labels may distort classification, a handful of examples may induce overgeneralization, and superficial cues may displace deeper structures with genuine explanatory power.

7.3.2 The Cognitive Basis of Inductive Capacity

Psychological research has long treated induction as a fundamental capacity for human learning. In their experiments on infant statistical learning, Saffran, Aslin, and Newport (1996) showed that infants can use transitional probabilities in speech sequences to discover word boundaries. This finding suggests that the extraction of statistical regularities has an early developmental basis and is an important mechanism of human cognitive development. Shepard (1987)'s theory of generalization indicates that people generalize on the basis of similarity within a psychological space. Bayesian approaches to cognition developed by J. B. Tenenbaum et al. (2011) and Griffiths and J. B Tenenbaum (2006) further understand human concept learning as a process of probabilistic inference between prior knowledge and available evidence.

These studies have two implications for education. First, inductive capacity has a natural

foundation: students actively seek patterns as they learn. Second, education must nevertheless calibrate the inductive process. Spontaneous induction is easily influenced by the salience of a sample, emotional intensity, narrative fluency, and prior commitments. The representativeness heuristic identified by Tversky and Kahneman (1974) shows that people often treat the fact that "a case resembles a typical member of a category" as evidence either that it belongs to the category or that the category is more likely. In doing so, they neglect the category's actual prevalence in the population and underestimate the random variation arising from sample size and sampling method. In social life, this tendency can manifest itself in stereotypes, generalization from individual cases, overinterpretation of small samples, and excessive trust in vivid narratives.

AI cases can make these problems concrete. If an image model frequently encounters "wolves in snow" in its training data, it may treat the snowy background as the main cue for identifying a wolf. If a student repeatedly encounters one type of case on social media, the student may likewise mistake frequent exposure for the true prevalence of the phenomenon in society. The two cases arise through different mechanisms, but both reveal the same structural problem: the relation between a sample and a population has been ignored, and superficial co-occurrence has been mistaken for a reliable regularity.

7.3.3 Educational Objectives for Inductive Capacity

The educational objectives for inductive capacity may be summarized as four progressively demanding cognitive actions: discovering a pattern, explaining the evidence that supports it, analyzing the relationship between the sample and the population, and formulating the inductive result as a testable hypothesis that specifies conditions and possible counterexamples. Students thus move from "I have observed a pattern" toward "What evidence currently supports this pattern, and how should it be tested further?"

Artificial intelligence turns the materials and choices on which induction depends into observable components of a system. By examining the relationships among training data, labels, features, and model results, students can see where a pattern comes from. By adding or removing samples, adjusting classification criteria, and comparing types of error, they can also test whether their original understanding is robust. When constructing even a simple task themselves, students must decide what to observe, how to represent it, which differences merit attention, and how results should be judged. The inductive premises previously hidden within their understanding thereby become explicit choices that must be explained, tested, and revised.

7.3.4 A Classroom Case for Inductive Capacity

A representative task might focus on "AI butterfly recognition." The teacher presents images of several butterflies and moths. Students first observe differences in their appearance, and then ask a simple classifier or an existing AI tool to identify them. They record cases in which the AI is correct and incorrect, and analyze whether it may be making its decisions on the basis of color, wing shape, background, camera angle, or image quality. The teacher then guides students in comparing their own judgments with those of the AI: Which cues have greater discriminatory value? Which may be misleading? What cases are missing from the available sample?

The task centers on the process of induction; classification accuracy is only one source of evidence through which to observe that process. Students must answer the following questions: Which samples have we seen? Can they represent all butterflies and moths? Are the AI's errors concentrated in particular backgrounds, angles, or forms? What kinds of samples would need to be added to make the conclusion more reliable? Through the activity, students come to understand the relationships among samples, features, patterns, counterexamples, and hypotheses.

At the upper-secondary and university levels, more complex tasks can be used. Students might analyze multiple answers generated by a large language model on the same topic and induce a taxonomy of common errors, such as fabricated citations, temporal confusion, overgeneralization, disregard for boundaries, and the interpretation of correlation as causation. They would then design methods of verification and determine which errors may result from the absence of external retrieval, which from conceptual confusion, and which from an ambiguously formulated problem. Errors made by generative AI thus become material for training in induction.

Table 7.1: Levels of Performance in Inductive Capacity

Level	**Performance**	**Task in an AI Context**	**Task in a Social Context**
Pattern discovery	Discovers recurring relationships, category features, and anomalies	Identifies patterns of correct and incorrect results in AI classification	Compares common claims and differences across multiple reports
Evidence awareness	Explains which evidence supports a pattern	Checks which assertions in an AI answer are supported by citations	Marks evidence-based and unsupported judgments in an article
Sample awareness	Analyzes the source and representativeness of a sample	Assesses whether training data cover different populations, contexts, and conditions	Judges whether a social survey represents its target population
Hypothesis awareness	Expresses a regularity as a testable hypothesis	Designs new samples to test whether a model relies on superficial cues	Proposes counterexamples or additional evidence to test a social conclusion

7.4 Generalization: Transferring Understanding and Identifying Boundaries in New Contexts

7.4.1 Why Generalization Is a Core Capacity in the Age of AI

Generalization is the crucial next step after induction has formed an understanding. Induction allows people to derive a pattern from materials; generalization requires them to judge whether that pattern extends to a new context. A conclusion that holds in its original context has only been supported under a particular set of conditions. Whether it continues to hold in a new context depends on changes in the objects, conditions, data distribution, task objectives, and level of risk. The age of AI brings this question to the foreground because the effectiveness of modern AI systems depends to a great extent on generalization, while many AI failures occur precisely when generalization crosses a boundary of applicability.

In machine learning, generalization usually refers to a model's performance on unseen data. Overfitting, distribution shift, adversarial examples, shortcut learning, and robustness all concern whether a model remains effective in new contexts. Research on shortcut learning by Geirhos et al. (2020) shows that models may rely on superficial cues that appear effective in the training environment and then fail in more complex or changing contexts. The WILDS benchmark introduced by Koh et al. (2021) demonstrates that real-world distribution shifts can substantially affect model performance. Such studies give students intuitive material through which to understand the conditions and boundaries under which a regularity applies.

Human learning likewise confronts the question of whether knowledge will transfer. A student who solves a familiar type of problem may fail when the question is phrased differently. An educational practice that succeeds in one region may prove ineffective when transferred to another. A historical analogy that appears similar at the level of events may break down under closer analysis of its institutional setting. A medical finding established in one population must be applied cautiously to another. The capacity for generalization is therefore central to modern education: it requires learners always to understand a regularity in relation to the conditions under which it holds.

7.4.2 Foundations of Generalization in the Learning Sciences

The learning sciences regard transfer as an important indicator of educational quality. Bransford, A. L. Brown, and Cocking (2000) point out that learners who merely memorize facts and procedures find it difficult to transfer knowledge to new problems; deep understanding and metacognition facilitate transfer. Perkins and Salomon (1988)'s theory of high-road transfer emphasizes that learners must consciously abstract general principles and actively seek ways to apply them in new contexts. The taxonomy of far transfer proposed by Barnett and Ceci (2002) further reminds us that the form of a task, the domain of knowledge, the physical and social settings, and temporal distance all affect the difficulty of transfer.

These findings impose clear requirements on AIGE. Learning that "models can overfit" does not spontaneously lead students to reflect on the superficial fluency produced by mechanical

drilling. Knowing that "training data may be biased" does not automatically enable them to analyze a social survey or business report. Courses must deliberately create bridging tasks that guide students to identify structures shared by AI problems and broader social problems, and to transfer the boundary awareness developed in AI contexts to other cognitive activities.

7.4.3 Educational Objectives for the Capacity for Generalization

The educational objectives for the capacity for generalization include identifying conditions, comparing contexts, expressing boundaries, and designing counterexamples. Students need to specify the data, task, population, temporal, and cultural conditions on which a conclusion depends; compare the deep similarities and differences between the original and new contexts; rewrite conclusions as conditional statements; and actively design cases in which those conclusions might fail.

Artificial intelligence turns the question of whether existing performance extends to a new context into one that can be tested repeatedly. Students can vary the background of an input, the target population, the form of expression, temporal conditions, and task objectives, and observe how model performance changes. They can also compare training, testing, and real-use contexts to trace which differences cause a previous conclusion to fail. Through such variation experiments, transfer ceases to be merely an outcome hoped for after a course has ended and becomes a cognitive activity that can be continuously observed and tested during learning.

7.4.4 A Classroom Case for Training Generalization

One task suitable for secondary and university students is "the boundaries of models and experience." The teacher provides a description of an emotion-recognition AI that performs well on a public dataset, followed by images drawn from different cultural settings, lighting conditions, age groups, and photographic conditions. Students must predict the conditions under which the model is likely to fail and explain the reasons for their predictions.

The teacher then presents a case from the social domain: a learning method has produced good results at a selective urban secondary school but only limited results when adopted by a rural school. Using the same analytical framework, students answer the following questions: What were the original conditions? Which conditions changed in the new context? Which variables might affect the transfer of the practice?

The task is intended to help students develop the habit of expressing boundaries; algorithmic details serve as material for understanding contexts and testing boundaries. Students must ultimately articulate both the boundaries of applicability of the AI conclusion and those of the social practice. By comparing the two, they can recognize that the capacity for generalization is a shared cognitive capacity that operates across technical and social contexts.

Table 7.2: Dimensions of Training for the Capacity for Generalization

Dimension	Core Question	AI Case	Social Case
Condition identification	On which conditions does the regularity depend?	Training data, model task, input format, and deployment environment	Research subjects, region, population, time, and institutional setting
Context comparison	Is the new context similar to the original one?	Whether test and training data have similar distributions	Whether a new school, population, or policy environment shares the crucial features
Boundary expression	Under which conditions does the conclusion hold?	The inputs and risk levels for which a model is suitable	The populations, environments, and objectives to which a practice applies
Counterexample design	What circumstances might cause the conclusion to fail?	Occlusion, noise, distribution shift, and adversarial examples	Extreme cases, heterogeneous populations, institutional change, and value conflict

7.5 Judgment: Making Reasoned Choices amid Uncertainty and Value Conflict

7.5.1 Judgment Is More Than Reaching a Conclusion

Induction and generalization help people form understandings; judgment brings those understandings into choice and action. The capacity for judgment concerns the formation of reasoned, explicable choices for whose consequences one can be held responsible, under conditions in which evidence is insufficient, outcomes are uncertain, stakeholders are numerous, and standards of value may conflict.

The capacity for judgment is especially important in AIGE. AI systems can offer recommendations, rankings, scores, predictions, and generated proposals, but they cannot bear normative responsibility in education, medicine, law, public governance, or personal life. Whether a student uses AI to generate an essay, a teacher adopts an AI-generated score, a physician follows an AI-assisted diagnosis, a school uses a data model to predict student risk, or a platform recommends content algorithmically, the decision involves multiple forms of judgment concerning facts, methods, risks, values, and responsibilities.

Factual judgment asks whether information is reliable; methodological judgment, whether a tool is appropriate; risk judgment, whether potential consequences are tolerable; value judgment, whether the objectives pursued by a choice are justified; and responsibility judgment, who should be accountable for the decision and its consequences. In specific human–AI tasks, these judgments must also be translated into arrangements for collaboration: which tool to select, how to divide the work, where to require human review, and when a human should take over or stop the system. A mature process of judgment generally addresses these interrelated questions together.

7.5.2 Psychological and Educational Foundations of Judgment

Research on judgment under uncertainty provides important points of reference for AIGE. Tversky and Kahneman (1974) show that people often rely on heuristics in probabilistic judgment and may neglect base rates, sample size, and randomness. Research on misinformation by Pennycook and Rand (2019) indicates that people who engage in less reflective thinking are more likely to judge false information to be true. These findings do not imply that human judgment cannot be improved through education. Fong, Krantz, and Nisbett (1986) show that statistical training can improve people's ability to apply statistical principles to everyday problems. Gigerenzer and Hoffrage (1995) likewise demonstrate that appropriate forms of information representation can improve Bayesian reasoning.

Research on critical thinking further emphasizes the structured character of judgment. The Delphi Report led by Facione (1990) defines critical thinking as purposeful, self-regulatory judgment encompassing interpretation, analysis, evaluation, inference, explanation, and self-regulation. This conception is closely related to the cognitive autonomy discussed in this book. Training in judgment within AIGE must bring the tradition of critical thinking into an information environment mediated by models and platforms, enabling students to evaluate AI outputs, expert opinions, statistical graphics, and social narratives.

Research on forecasting offers another important case. Studies of "superforecasters" by Mellers et al. (2015) show that high-performing forecasters in geopolitical prediction tasks tend to express beliefs probabilistically, update them promptly, decompose complex problems, and remain open-minded. These capacities are closely related to trust calibration and self-awareness in AIGE. When students encounter AI advice, they too need to express their judgments in probabilistic and conditional terms, avoiding a crude oscillation between complete acceptance and outright rejection.

7.5.3 Educational Objectives for the Capacity for Judgment

The educational objective for the capacity for judgment is to enable students to distinguish factual statements, explanations, value claims, and recommendations for action; evaluate the strength of evidence; weigh risks and consequences; identify conflicts of value; and explain why they accept, modify, defer, or reject a conclusion.

Students must also use the objectives of a task, the available evidence, and its level of risk to decide whether to use AI, which tool to choose, which stages to delegate, and where to establish conditions for human review, takeover, or termination. These choices require explicit reasons

and must also be open to an analysis of responsibility, in which learners explain the authority held and the responsibilities borne by the different participants.

Artificial intelligence brings many elements of judgment together within a single task. When designing or using an AI system, students must simultaneously consider how the objective is defined, how evidence is obtained, how metrics are chosen, whether the costs of different errors are symmetrical, which groups will be affected, and who has the authority to make the final decision. The fact that a system can run shows only that a proposal is executable. Whether it should be adopted requires connecting technical performance with risk, value, institutions, and responsibility. Such tasks transform judgment from a general disposition into a choice made under specific conditions.

7.5.4 A Classroom Case for Training Judgment

A representative task asks, "How should the results of AI-assisted diagnosis be used?" The teacher provides performance information for a medical-imaging AI system and then presents several application contexts: routine screening, difficult cases, regions with scarce medical resources, situations involving sensitive patient privacy, and settings in which the model's training data differ substantially from the local population. For each context, students must decide at which stages the AI recommendation may be used, what human review is necessary, how uncertainty should be communicated to patients, and who should bear responsibility for the consequences of different kinds of error.

The task can then be transferred to school education. The teacher presents an AI essay-scoring system and asks students to analyze whether it is suitable for preliminary feedback on routine exercises or for scoring high-stakes examinations. Might the system produce unfair outcomes for students with different dialects, writing styles, or cultural backgrounds? Do students have a channel through which to learn the basis of a score and appeal it? In this way, students learn to apply the same framework of judgment across domains.

Table 7.3: Components of the Capacity for Judgment

Component	**Core Question**	**Example in an AI Context**	**Evidence for Assessment**
Factual judgment	Is the information reliable?	Whether facts and citations in an AI answer can be verified	Distinguishes facts, explanations, and conjectures
Methodological judgment	Is the tool appropriate?	Whether a model is suitable for the current task and target population	Explains the conditions and limitations of a tool's use
Risk judgment	Are the consequences tolerable?	Whether AI advice is suitable for low- or high-risk contexts	Proposes review, takeover, and risk-mitigation measures

Component	Core Question	Example in an AI Context	Evidence for Assessment
Value judgment	Which objectives matter more?	How efficiency, fairness, privacy, and safety should be balanced	Identifies stakeholders and conflicts of value
Responsibility judgment	Who should bear the consequences?	The respective responsibilities of developers, deployers, and users	Maps the chain of responsibility and identifies key points of accountability
Collaborative judgment	How should humans and AI divide the work?	Selecting tools, allocating tasks, and establishing points for review and takeover	Explains the rationale for the division of labor and retains authority over key decisions

7.6 Self-Awareness: Recognizing and Correcting One's Own Cognitive State

7.6.1 Why Self-Awareness Is Necessary

Self-awareness is among the more integrative capacities in the four-capacity model. Induction, generalization, and judgment can all go astray; self-awareness enables learners to notice these deviations within the cognitive process and correct them in a timely manner. This book uses *self-awareness* to encompass the educational requirements of metacognitive monitoring and cognitive correction, emphasizing a cognitive state that is active, timely, and vigilant. It includes both continuous monitoring and retrospective review of cognitive processes, as well as the learner's immediate recognition that fluent expression, authoritative sources, group opinion, emotional mobilization, or AI outputs may be exerting an influence.

An intelligent society has a particular need for self-awareness. Generative AI can produce apparently complete answers in fluent language; recommender systems can continually shape human attention; automated scoring can create an impression of objectivity and precision; and intelligent assistants can substantially reduce the cost of obtaining answers and completing tasks. In an environment of convenience, individuals may easily transfer an increasing share of cognitive activity to external systems: asking AI to solve a problem before understanding it, requesting a model-generated summary before reading the material, adopting a machine-generated draft before forming a view, or accepting a system recommendation before completing one's own judgment.

The capacity for self-awareness requires learners to keep asking: Do I genuinely understand the problem? Do I know the grounds for the answer? Have I mistaken generated content for my own thinking? Am I overestimating a result's reliability because it is fluently expressed? Do I

know which parts of the work I have delegated to AI? Am I allowing a platform continually to determine what receives my attention? Such questions help learners remain aware of their own cognitive state and of the division of labor between human and AI.

7.6.2 Research Foundations of the Capacity for Self-Awareness

Research on metacognition provides the psychological foundation for the capacity for self-awareness. Flavell (1979) introduced metacognition into the study of cognitive development, emphasizing both an individual's knowledge of their own cognitive processes and the monitoring of those processes. Nelson and Narens (1990) proposed a framework comprising a meta-level and an object-level, showing how a metacognitive system can monitor and control first-order cognitive activity. From a neuroscientific perspective, Fleming and Dolan (2012) further examined the foundations of metacognitive accuracy and showed that task performance can dissociate from awareness of that performance. A student may therefore complete a task without accurately understanding their own state of comprehension.

Research on self-regulated learning situates metacognition within the full course of learning activity. Zimmerman (2002) understands self-regulated learning as a cycle in which learners actively set goals, monitor processes, adjust strategies, and evaluate outcomes. P. R Pintrich (2000) likewise emphasizes the regulation of motivation, cognition, behavior, and context in learning. AIGE must extend this form of self-regulation into human–AI interaction: students need to regulate not only their learning strategies, but also their division of labor, trust, and dependence in relation to AI systems.

Research on epistemic vigilance and epistemic cognition further suggests that learners need to evaluate information sources, the reliability of content, and the manner in which knowledge claims are supported (Sperber et al., 2010; Hofer and Paul R. Pintrich, 1997; Sandoval, Greene, and Bråten, 2016). Information in an AI environment may originate from people, models, platforms, institutions, or complex hybrid chains of production. Self-awareness therefore also includes awareness of how one's trust is shaped by different sources and by the ways in which they are presented.

7.6.3 Educational Objectives for the Capacity for Self-Awareness

The educational objectives for the capacity for self-awareness include monitoring understanding, maintaining collaborative awareness, regulating trust, identifying bias, and reflecting and revising. Students need to recognize the gap between being able to read the words and understanding the grounds for a conclusion. They must be aware of the stages at which AI expanded the available materials, explanations, or proposals; the stages at which they themselves supplied objectives, knowledge, and choices; the tasks they delegated to AI; and the crucial judgments for which they remain responsible.

Students must also judge whether their degree of trust corresponds to the strength of the evidence and the risk of the task; recognize how emotion, prior commitments, authority, platform mechanisms, and linguistic fluency influence their judgment; and actively revise conclusions, strategies for action, and the human–AI division of labor in light of new evidence. Self-awareness

runs through the entire process of induction, generalization, and judgment, and can also initiate a new round of understanding through retrospective review after an action has been completed.

7.6.4 Classroom Tasks for Training Self-Awareness

One effective task is an "AI-use log." While completing a writing, research, or project assignment, students record the purpose, input, output, degree of adoption, process of verification, and final revisions associated with each use of AI. After completing the assignment, they answer the following questions: Which parts came from my prior understanding? Which were inspired by AI? Which content did I verify? Which did I ultimately remove? Where did a risk of excessive dependence on AI arise?

Another task is a "confidence-change record." Students first judge a question independently and record their level of confidence. They then examine an AI answer, expert sources, or peer opinions. Finally, they record how their judgment and confidence changed and identify the evidence that prompted each change. This task makes belief updating explicit and helps students avoid unconsciously shifting their position under the influence of external opinion.

A course can also require an "error review." Students select an instance in which their use of AI failed: perhaps they trusted an incorrect citation, adopted an inappropriate summary, or applied AI advice in an unsuitable context. In reviewing the failure, they analyze the stage at which the error occurred, such as source checking, evidence verification, assessment of distributions, expression of boundaries, calibration of confidence, or responsibility judgment. Error thereby becomes important material for understanding and correcting one's own cognitive processes.

Table 7.4: Dimensions of Training for the Capacity for Self-Awareness

Dimension	**Core Question**	**Training Task**	**Evidence for Assessment**
Monitoring understanding	Do I genuinely understand?	Explains the chain of evidence and limits of applicability of an AI answer	Identifies aspects they do not yet understand
Collaborative awareness	What did AI extend, and what did I delegate to it?	Maintains an AI-use log and attributes contributions	Distinguishes personal thinking, AI suggestions, and external sources, and explains AI's actual contribution
Regulating trust	How much should I believe?	States a confidence level and the reasons for it	Aligns confidence with the strength of evidence and the risk of the task

Dimension	Core Question	Training Task	Evidence for Assessment
Identifying bias	What factors are influencing me?	Analyzes the effects of emotion, commitments, platform mechanisms, and authority	Identifies non-evidential factors influencing judgment
Reflecting and revising	How should I update my conclusion?	Reviews errors and produces a second revision	Revises judgment, strategy, and the human–AI division of labor in light of new evidence

7.7 The Four-Capacity Cycle: From Hypothesis to Action and Correction

Induction, generalization, judgment, and self-awareness each have a relatively distinct focus in training, yet they form an interconnected cycle in actual cognitive activity. Induction answers, "What do I see in the materials?" Generalization asks, "How far does this understanding extend?" Judgment asks, "What should I choose under the present conditions?" Self-awareness asks, "Why do I understand and choose in this way, and do I need to revise my position?"

In the main course of a curricular task, the four capacities address hypothesis formation, boundary testing, choice of action, and cognitive correction in turn. New evidence, changes of context, and feedback from action then prompt learners to induce again, conduct further tests, and adjust their judgment. Self-awareness also intervenes at every stage of induction, generalization, and judgment, monitoring whether learners have adequately understood the materials, crossed the boundaries of a conclusion, or allocated trust and responsibility appropriately. The four-capacity cycle is therefore a task structure with feedback relations, not a linear procedure that can unfold only in a fixed sequence.

The cycle can be used directly to design curricular tasks. In a unit on "AI recommendation and filter bubbles," students might first identify patterns in the content recommended to them, then analyze why different users encounter different informational worlds, judge how platforms and individuals should share responsibility, and finally reflect on how recommender mechanisms have shaped their own attention and beliefs. In a unit on "AI-assisted medical imaging," students might first induce patterns from cases in which the model is correct and incorrect, analyze the risks of generalizing the model to new hospitals and populations, judge how AI advice should enter the diagnostic and treatment process, and finally reflect on whether they have mistaken model confidence for medical certainty.

Table 7.5: The Four-Capacity Cycle in Curricular Tasks

Stage	**Core Action**	**Common Risk**	**Instructional Scaffold**
Induction	Forms a preliminary hypothesis from materials	Overgeneralization, inadequate samples, and misleading fluency	Sample tables, evidence annotation, and counterexample collection
Generalization	Takes the hypothesis into a new context	Distribution shift, boundary overreach, and superficial analogy	Condition checklists, context comparison, and design of failure conditions
Judgment	Makes a choice by integrating evidence, risk, value, and responsibility	Insufficient evidence, tool mismatch, value conflict, and ambiguous responsibility	Comparison of options, risk matrices, stakeholder maps, and responsibility chains
Self-awareness	Monitors and revises understanding, trust, and human–AI collaboration	Overconfidence, cognitive outsourcing, failure to recognize AI's contribution, and emotionally driven responses	Confidence statements, AI-use logs, contribution analysis, and error review

7.8 Evidence for Assessment: How to Observe the Four Capacities

Assessment in AIGE must attend simultaneously to task outcomes, cognitive processes, and the grounds for judgment. Neither the completeness of a product nor model performance can by itself demonstrate that cognitive autonomy has developed. Teachers should also observe whether students can explain samples and evidence, compare new and prior contexts, use AI to expand materials and possible solutions, select tools and organize the human–AI division of labor, weigh risks, values, and responsibilities, and reflect on the cognitive gains, dependence, and changes of belief produced by AI. The following table summarizes the principal evidence of performance and assessment tasks for the four capacities.

Table 7.6: Evidence for Assessing the Four-Capacity Model

Capacity	**Observable Performance**	**Assessment Task**	**High-Level Performance**
Induction	Annotates evidence, discovers patterns, expands materials, and proposes hypotheses	Analyzes and extends a set of AI outputs or social cases	Explains sample limitations, compares candidate explanations, and proposes tests using counterexamples
Generalization	Compares contexts, generates variations, explains conditions, and designs boundaries	Judges whether a model, research finding, or social practice can transfer	Evaluates variations and counterexamples and rewrites conclusions in conditional form
Judgment	Selects tools and weighs evidence, risks, values, and responsibilities	Analyzes alternative human–AI arrangements in AI medicine, AI scoring, or recommender systems	Forms an explicable choice, arranges the human–AI division of labor, and proposes review mechanisms
Self-awareness	Monitors understanding, regulates trust, and analyzes cognitive gains and dependence	Completes an AI-use log, contribution statement, confidence statement, or error review	Explains AI's actual contribution and actively revises judgment in light of new evidence

Assessment must also avoid separating the four capacities from one another. An individual task may examine a single capacity, but an integrative task should present as much of the complete cognitive cycle as possible. When analyzing an AI-generated news summary, for example, students might first identify its principal assertions and the evidence for them, then judge whether those assertions generalize to other populations or regions, decide whether to use or share the content, and finally reflect on why they initially trusted or doubted the summary. Compared with questions that merely require students to identify a correct answer, such tasks provide more complete evidence of cognitive autonomy.

Assessment should also examine whether capacities transfer. A student's ability to identify bias in AI training data does not mean that the student can recognize sampling bias in a social survey. The ability to explain that a model has boundaries of applicability does not mean that the student will actively examine the conditions under which expert opinion, statistical conclusions, or personal experience apply. Courses must therefore establish tasks that support transfer in both directions between AI contexts and broader social contexts, observing whether learners can apply sample awareness, boundary awareness, risk awareness, and self-monitoring to new problems. The four-capacity model points toward a relatively stable capacity for cognitive

autonomy only when these cognitive actions extend beyond particular tools and cases.

7.9 Chapter Summary

This chapter has translated cognitive autonomy from an overarching educational ideal into four capacities—induction, generalization, judgment, and self-awareness—that can be designed for, practiced, and assessed. The four-capacity model is a curricular model for AIGE. It does not seek to exhaust the psychological constitution of cognitive autonomy. It begins instead with characteristic cognitive difficulties in an intelligent society and extracts the core cognitive activities that curricula can train and observe on a sustained basis.

AIGE can support the cultivation of these capacities because of the scientific characteristics of artificial intelligence and the ways in which it exerts social effects. Understanding intelligent mechanisms enables learners to trace how computational conclusions are formed from data, representations, models, objectives, and evaluation, and to recognize the grounds and boundaries of those conclusions. Participating in computational construction requires learners to convert initially vague understandings into data, rules, metrics, and procedures, and to use operational results to discover omissions and limitations in their assumptions. Analyzing social embedding enables learners to trace how computational results acquire real-world force through platforms, organizations, and institutions, and to judge the values, authority, effects, and responsibilities involved. The three paths connect the internal workings of systems, cognitive practice, and the social environment, providing mutually complementary curricular foundations for cultivating cognitive autonomy.

On this basis, induction enables learners to form evidence-based, testable understandings from limited materials. Generalization enables them to bring existing understandings into new contexts and identify the conditions under which a regularity holds. Judgment enables them to make reasoned choices when evidence is incomplete, risks are uncertain, and values conflict. Self-awareness enables them to recognize their own states of understanding, trust, and dependence and revise them in light of new evidence. In authentic tasks, the four capacities are interconnected and form a cycle from hypothesis generation and boundary testing to choice of action and cognitive correction. Self-awareness operates throughout this cycle, allowing learners continually to examine their cognitive processes and initiate a new round of understanding and judgment.

The development of cognitive autonomy ultimately depends on whether learners themselves undertake these cognitive activities. Artificial intelligence can help organize materials, generate possible solutions, construct variations, and provide feedback, but courses must still ensure that learners personally compare samples, explain evidence, test boundaries, weigh risks, make choices, and reflect on the human–AI division of labor. The quality with which a tool completes a task is not equivalent to the degree to which a learner has developed a capacity. AI participation becomes genuine training in cognitive autonomy only when learners can explain the stages in which AI was involved, examine its results, make the final judgment on the basis of evidence, and assume responsibility for their choices.

Assessment in AIGE must therefore move beyond tool proficiency, product completeness, and model performance to observe cognitive processes and their transfer. Teachers need to examine whether learners can form testable understandings, identify the boundaries within which conclusions apply, make explicable choices under complex conditions, and monitor and revise their own cognitive states. Artificial Intelligence General Education truly assumes the function of education for cognitive autonomy in an intelligent society when these capacities transfer from AI contexts to the evaluation of information, the acquisition of knowledge, and social action.

Chapter 8

Artificial Intelligence General Education and the Intellectual Training of Other Disciplines

Abstract. Mathematics, language and literature, physics, history, and computer science have each established important traditions of intellectual training: formal reasoning, the interpretation of meaning, causal explanation, contextual judgment, and executable construction, respectively. These traditions provide a deep foundation for cognitive autonomy, but they also raise a question about the place of Artificial Intelligence General Education in the curriculum: if existing disciplines already cultivate a range of cognitive capacities, what distinctive role remains for AIGE? Using the four capacities of induction, generalization, judgment, and self-awareness as a common frame of reference, this chapter compares five representative forms of disciplinary training. It then explains how AIGE can organize cognitive resources dispersed across these disciplines into the complete chain of "data–model–output–action–feedback." Within this chain, intelligent activity assumes an executable form, while model outputs enter real-world action through platforms, organizations, and institutions. The resulting education for cognitive autonomy has relatively independent content and curricular structure, while also working in concert with critical thinking, media literacy, statistical thinking, and computational thinking.

Keywords: Artificial Intelligence General Education; disciplinary thinking; intellectual training; four capacities; cognitive autonomy

8.1 The Question: Why Compare Disciplines?

Chapter 3 of Part I defined cognitive autonomy and its six analytical dimensions. Chapters 5 and 6 examined, respectively, the discipline of artificial intelligence itself and the mechanisms through which it enters social life. Chapter 7 then proposed the four-capacity model of induction, generalization, judgment, and self-awareness, explaining how Artificial Intelligence General Education translates the educational ideal of cognitive autonomy into a model of capacities that can be cultivated and assessed. This chapter pursues a further curricular question: given that

established disciplines already provide diverse forms of intellectual training, what distinctive role can AIGE play in developing the capacity for cognitive autonomy?

This question bears directly on the curricular position of AIGE. Educational systems have already developed relatively stable disciplinary structures at every stage. If AIGE amounts only to the addition of new tools and operational skills, it can easily become just another technical course on the timetable. Yet if all of its content is dispersed among existing subjects, it may lose the integrated structure that links data, models, platforms, action, and responsibility. General education is concerned with the development of the whole person. Whether a body of knowledge can secure a stable place in the curriculum therefore depends on whether it offers enduring ways of knowing and forms of intellectual training.

This chapter compares mathematics, language and literature, physics, history, and computer science as representatives of five traditions of training: formal reasoning, the interpretation of meaning, causal explanation, contextual judgment, and executable construction. A complete disciplinary genealogy would also include philosophy, statistics, the social sciences, the arts, and other fields; the present selection serves the analysis of core cognitive structures. The comparison identifies the principal emphasis of each discipline, reveals their shared contributions to cognitive autonomy, and thereby clarifies which educational content gives AIGE a relatively independent role. Whether a particular educational stage should adopt a stand-alone course, modular instruction, or interdisciplinary integration is a subsequent question of curricular implementation.

8.2 Skills, Disciplinary Ways of Thinking, and Intellectual Training

Any discussion of relations among disciplines must distinguish three levels: skills training, disciplinary ways of thinking, and intellectual training. Skills training develops the ability to perform specific tasks. Mathematical calculation, physical experimentation, the retrieval of historical sources, programming, data annotation, and the use of AI tools can all be improved through demonstration, practice, and feedback. Skills are an important component of learning activities, but their scope is usually closely tied to particular tasks and tools.

Disciplinary ways of thinking embody the fundamental ways in which a field poses questions, constructs its objects of inquiry, and tests its conclusions. Mathematics emphasizes abstraction, formalization, and proof; language and literature emphasize context, meaning, and expression; physics emphasizes modeling, idealization, and experimental testing; history emphasizes sourcing, contextualization, and corroboration; and computer science emphasizes decomposition, algorithmic formulation, executability, and debugging. Disciplinary thinking gives learners particular perspectives from which to observe the world and furnishes methodological resources that can be called upon when they confront new problems.

Intellectual training refers to the relatively stable cognitive dispositions formed through sustained study. Such dispositions arise from concrete disciplinary activities and, with varied practice, comparative reflection, and support for transfer, can be brought to bear on new kinds of problems. Moving from disciplinary learning to cross-contextual capacity requires deliberate

instructional design: mastery of knowledge and proficiency in skills do not automatically produce far transfer. Assessing the general-educational value of a course therefore requires attention not only to the knowledge and methods it teaches, but also to the activities through which learners form particular cognitive dispositions and to whether they can consciously activate those dispositions in new environments.

The four capacities established in the preceding chapter provide a common frame of reference for disciplinary comparison. Induction concerns how warranted understanding is formed from limited materials. Generalization concerns whether existing understanding can be extended to new objects and contexts. Judgment concerns how evidence, values, risks, and responsibilities are integrated in making a choice. Self-awareness concerns how one monitors and corrects one's states of understanding, trust, dependence, and action. The six analytical dimensions proposed in Part I specify what must be regulated in cognitive autonomy; the four capacities specify what enables learners to carry out that regulation. The comparisons below are organized around the four capacities while also considering the foundational support each discipline provides for the six dimensions.

8.3 Mathematics: Necessary Reasoning under Formal Constraints

The central intellectual discipline of mathematics can be characterized as necessary reasoning under formal constraints. Mathematics education teaches learners to abstract structures from concrete objects and derive conclusions within systems constituted by definitions, axioms, and rules. Numbers, shapes, functions, probabilities, spaces, change, and relations can all become objects of abstraction. Learners gradually come to understand that intuition may give rise to a conjecture, but whether a conclusion holds remains subject to the constraints of its conditions, proof, and consistency.

Pólya (1957)'s classic study of problem solving emphasizes the continuous process of understanding a problem, devising a plan, carrying out the plan, and looking back. Schoenfeld (1985) and Schoenfeld (1992) further argue that mathematical problem solving is jointly shaped by knowledge resources, strategies, process control, and systems of belief, with metacognitive monitoring playing a crucial role. From this perspective, conjecture, proof, counterexample, and review in mathematics support, respectively, the induction of relations from observed phenomena, the testing of whether conclusions can be generalized, judgment about the validity of reasoning, and monitoring of the problem-solving process. Probability and statistics additionally train learners to adjust their judgments in response to the strength of evidence and the degree of uncertainty.

AI settings alter the epistemic conditions under which these forms of training operate. The objects, premises, and rules of a mathematical problem can usually be stated explicitly. By contrast, the premises of an AI system may be distributed across data collection, label definitions, metric selection, model training, and platform objectives; its conclusions are often probabilistic, empirical, and context-dependent. Learners must therefore bring the conditional awareness and

scrutiny of reasoning cultivated by mathematics to the evaluation of models, while also asking how the data were produced, whether a prediction transfers, what objective a system optimizes, and how its output shapes subsequent behavior. Formal reasoning is thus an important foundation for cognitive autonomy, but it cannot by itself encompass the entire chain of questions involving sources, platforms, feedback, and responsibility.

8.4 Language and Literature: Interpreting and Expressing Meaning in Context

The central intellectual discipline of language and literature can be characterized as interpreting and expressing meaning in context. Words, tone, narrative, stance, voice, and rhetoric all participate in the production of meaning; the same sentence may take on a different meaning in a different situation. Language study enables learners to move beyond literal information and understand metaphor, emotion, cultural background, and communicative purpose. Literary reading, in turn, unfolds the complexity of experience through character, narrative, and multiple perspectives.

Vygotsky (1978) treats language as an important medium of intellectual development and emphasizes the role of social interaction and symbolic tools in cognitive growth. Halliday (1978) examines the meaning-making functions of language in social context, while L. M. Rosenblatt (1978) understands reading as a transaction between reader and text. Close reading, summarization, and interpretation support the induction of meaning from words, sentences, and structures. Comparison across texts and the reconstruction of context promote generalization. Argumentation and expression require reasoned judgment, while revision and reflection in writing continually exercise self-awareness.

AI-generated content places the interpretation of meaning within a new structure of sources. The objects learners encounter now include not only relatively stable texts, but also instantly generated words, images, and sounds, as well as information filtered by algorithms and ranked by platforms. Evaluating a piece of content requires analysis of what it expresses and how it uses narrative and rhetoric, together with investigation of its sources, generative mechanisms, and chain of dissemination, and verification of the facts and evidence beneath its fluent presentation. The sensitivity to context and capacity for expression cultivated by language and literature can support this work. AIGE must additionally bring model intervention, the role of platforms, the consequences of use, and the attribution of responsibility into the same cognitive process.

8.5 Physics: Causal Explanation under Empirical Constraints

The central intellectual discipline of physics can be characterized as causal explanation under empirical constraints. When confronting complex natural phenomena, learners must identify key variables, construct idealized models, formulate testable predictions, and revise their explanations through observation, experimentation, and error analysis. A physical model always holds under particular assumptions and conditions; both the force of its conclusions and the limits of their

applicability consequently demand attention.

A study by Chi, Feltovich, and Glaser found that novices tend to classify problems by their surface situations, whereas experts are more likely to organize them according to such deep principles as conservation of energy and Newton's laws(Chi, Feltovich, and Glaser, 1981). The modeling instruction proposed by Hestenes (1992) emphasizes a continuous process of model construction, validation, and application, while diSessa (1993) shows how everyday intuitions can both support and interfere with the understanding of physics. Physics learning thus trains learners to induce relations from limited phenomena, transfer models to new conditions, compare explanations against experimental evidence, and, when prediction and observation diverge, inspect their assumptions, measurements, and reasoning.

AI systems frequently make predictions on the basis of statistical correlations, and their outputs may in turn change human behavior and the data subsequently collected. Causal explanation, statistical prediction, and feedback effects must therefore be distinguished. Learners need to ask whether a result is consistent with established mechanisms and observational evidence, whether the training and application environments are aligned, whether model confidence is calibrated, and whether the system's participation in action has altered the original conditions. Physics provides an awareness of models, evidence, causality, and error; AIGE further addresses questions of trust and responsibility that arise when people, data, models, platforms, and institutions act together.

8.6 History: Contextual Judgment from Limited Evidence

The central intellectual discipline of history can be characterized as contextual judgment from limited evidence. The past cannot be observed directly or subjected to repeated testing; people can form an understanding of it only from surviving materials. Those materials may be incomplete and are shaped by the identities, standpoints, purposes, and historical circumstances of those who produced them. Historical study requires learners to scrutinize sources, compare accounts, reconstruct contexts, and maintain distinctions among fact, interpretation, and evaluation.

S. S. Wineburg (1991) found that historical experts continually engage in sourcing, contextualization, and corroboration when reading documents. Seixas and Morton (2013) subsequently articulated historical thinking in terms of historical significance, evidence, continuity and change, cause and consequence, historical perspectives, and the ethical dimension. Research by S. Wineburg and McGrew (2019) on the evaluation of online information also shows that professional fact-checkers tend to leave the page before them and read laterally. The scrutiny and corroboration of historical sources support the formation of understanding from limited materials; contextual reconstruction helps learners test whether an interpretation can be extended; the weighing of evidence and ethical evaluation constitute judgment; and sustained attention to missing materials, interpretive bias, and new evidence fosters self-awareness.

Generative AI and algorithmic dissemination have changed how materials are produced and made visible. Text, images, audio, and video can be generated or altered by models, while platform ranking influences what is seen and how often it is seen. Learners need to trace sources

through authors, data, models, publishers, and distribution platforms, and to examine how technical processes transform both information and its appearance of credibility. The sourcing, corroboration, and contextual judgment cultivated by history can be brought to this task; AIGE additionally subjects generative mechanisms, algorithmic mediation, dissemination feedback, and real-world responsibility to simultaneous scrutiny.

8.7 Computer Science: Executable Construction for Complex Problem Solving

The central intellectual discipline of computer science can be characterized as executable construction for complex problem solving. Learners must specify objectives and constraints, decompose a problem into interconnected subtasks, select data representations, design algorithms, and continually revise their solutions through execution, testing, and debugging. Abstraction, decomposition, automation, and debugging give problem solving a form that is inspectable, executable, and iterative.

Wing (2006) characterizes computational thinking as a way of solving problems, designing systems, and understanding human behavior that draws on the fundamental concepts of computer science. Papert (1980) argues that programming can become a tool through which children think about their own thinking, while Brennan and Resnick (2012) analyzes creative programming in terms of computational concepts, computational practices, and computational perspectives. Problem decomposition and data representation support induction; modular reuse and the analysis of conditions involve generalization; testing and the selection among solutions constitute judgment; and debugging and iteration provide observable forms of activity through which self-awareness can be cultivated.

Artificial intelligence carries forward this tradition of executable construction, but it also alters the relationship between system behavior and design intention. In many conventional programs, designers specify the rules and steps of execution explicitly. Machine-learning systems develop behavior from data, and their outputs are jointly shaped by training data, model capabilities, input formulation, contexts of use, and platform objectives; a particular result cannot readily be read off from the program's procedural steps alone. Learners must therefore move beyond asking "How can this problem be transformed into a process a machine can execute?" to asking "How should the machine's participation in human cognition and action be governed?" Which tasks can be delegated, and to what extent? What checks must an output undergo? When should a person suspend or reject its use? Who retains final judgment and bears responsibility?

8.8 The Distinctive Intellectual Training Provided by Artificial Intelligence General Education

The preceding comparisons show that traditional disciplines already provide many of the foundations of cognitive autonomy. This does not reduce AIGE to a mere repetition of established forms of intellectual training. Its relatively distinctive value arises from the structure revealed jointly

in Chapters 5, 6, and 7. Chapter 5 showed that artificial intelligence transforms representation, learning, inference, testing, and revision into processes that can be constructed and run, making intelligent activity itself an observable, operable, and comparable object of learning. Chapter 6 showed how model outputs enter information dissemination, organizational decision making, and real-world action through platforms, organizations, and institutions, and how they alter subsequent data and human behavior through continuing feedback. Chapter 7 then organized this process into a cycle of the four capacities: induction, generalization, judgment, and self-awareness.

Taken together, these three dimensions enable AIGE to engage learners in observing, constructing, using, and correcting cognitive processes across the complete chain of "data–model–output–action–feedback." Learners must form understanding from data and experience while also testing whether a model transfers to new objects. They must evaluate outputs, arrange the division of labor between human and AI, and make choices, while also monitoring how using the system changes their own states of trust, dependence, and responsibility. Here, cognitive processes possess both technical form and social force; any one traditional discipline covers only part of the chain.

This complete chain also places learners in multiple, changing roles. A learner may be a data provider, a model user, a participant in system design, an evaluator of results, or an actor affected by an algorithm. Changes in role alter the evidence available to the learner, the control the learner can exercise, and the responsibility the learner should bear. AIGE thereby connects questions of knowledge with questions of action. Whether an output is reliable must be judged in light of its sources, evidence, and limits. Whether it should be used further requires judgments about values, risks, reversibility, and responsibility. Whether its use should continue depends on observing feedback and adjusting trust and dependence accordingly.

Table 8.1 summarizes the principal forms of intellectual training provided by these disciplines. The entries identify each discipline's central emphasis; they do not purport to define the full value of any discipline.

Table 8.1: Core Intellectual Training and Typical Criteria of Judgment across Major Disciplines

Discipline	**Core intellectual training**	**Typical criteria of judgment**
Mathematics	Necessary reasoning under formal constraints	Are the premises explicit? Is the reasoning valid? Does the conclusion follow necessarily from the premises?
Language and literature	Interpreting and expressing meaning in context	Is the understanding consistent with the context? Is the interpretation grounded in the text? Is the expression accurate and appropriate?

Discipline	Core intellectual training	Typical criteria of judgment
Physics	Causal explanation under empirical constraints	Is the explanation consistent with empirical facts? Is the mechanism plausible? Can the prediction be tested?
History	Contextual judgment from limited evidence	Are the sources reliable? Can the evidence be corroborated? Does the interpretation fit the historical context?
Computer science	Executable construction for complex problem solving	Can the problem be represented explicitly? Can the process be executed and the results tested? Can errors be located and corrected?
AIGE	Cognitive autonomy in intelligent environments	Are the sources credible and the evidence sufficient? Are the limits of the conclusion clear? Is trust in the model appropriately calibrated? Are judgment and responsibility clearly assigned?

AIGE also has a clear collaborative relationship with several adjacent literacies and modes of thought. Critical thinking provides general methods for scrutinizing reasons, evaluating arguments, and reflecting on judgments. Media literacy attends to information sources, forms of expression, processes of dissemination, and platform influence. Statistical thinking attends to data, sampling, uncertainty, and inference. Computational thinking attends to problem representation, decomposition, execution, and debugging. AIGE draws on these cognitive resources and organizes them around the entire process through which artificial intelligence participates in belief formation, judgment and choice, and real-world action. Artificial intelligence constitutes the common object that gives its curriculum coherence: how data represent reality, how models learn and generate, how outputs enter platforms and institutions, where human judgment must be retained, and how responsibility should be allocated are treated as interconnected concerns.

8.9 Chapter Summary

Mathematics, language and literature, physics, history, and computer science provide important forms of training in formal reasoning, the interpretation of meaning, causal explanation, contextual judgment, and executable construction. In different ways, they support induction, generalization, judgment, and self-awareness, while providing cognitive foundations for regulating sources, evidence, boundaries, trust, and responsibility. Disciplinary intellectual training arises from concrete practices; its application across contexts requires variation, comparison, reflection, and support for transfer.

AIGE organizes these cognitive resources within the complete chain formed by data, models, outputs, action, and feedback. Intelligent activity assumes an executable form within this

chain and acquires real-world force through platforms, organizations, and institutions. Learners can thus observe cognitive processes, construct and use AI systems, evaluate their outputs, and adjust trust, dependence, the division of labor between human and AI, and arrangements of responsibility in response to feedback. This systemic integration constitutes the relatively independent value of AIGE as a form of intellectual training.

At this point, Part II has explained the scientific foundations, pathways of implementation, and distinctive value of AIGE in cultivating the capacity for cognitive autonomy from four perspectives: the discipline of artificial intelligence, the mechanisms through which artificial intelligence enters social life, the four-capacity model of cognitive autonomy, and its relationship to the intellectual training provided by other disciplines. The next task is to identify which aspects of AI knowledge, methods, applications, and questions of responsibility can support the cultivation of these capacities and together constitute a content system oriented toward cognitive autonomy.

Part III

Content Framework: What Artificial Intelligence General Education Should Teach

Introduction to Part III

The first two parts began with the cognitive environment of an intelligent society and drew on the scientific character of artificial intelligence and the mechanisms through which it operates in society to examine the contemporary necessity and feasibility of Artificial Intelligence General Education (AIGE). On this basis, Part III turns to the question of content: amid rapidly evolving and highly diverse AI technologies, what knowledge should form a common educational foundation for all learners, and how should that knowledge be organized into a relatively stable and internally connected system of content?

AI technologies continue to advance, while particular products, interfaces, and operating skills remain strongly tied to specific stages of development. General education must identify a relatively stable core of knowledge amid such change, organize curriculum content around enduring fundamental questions, and locate frontier developments—including large models, multimodal systems, reasoning models, and agents—within this structure. Such a content system should possess foundational depth, structural coherence, explanatory power, transferability, and public relevance. It should provide a common language for understanding AI, connect otherwise fragmented knowledge, explain how system capabilities arise and why systems fail, enable learners to analyze new technologies and contexts, and encompass issues in an intelligent society that bear directly on individual rights, public life, and responsibility for action.

Guided by this conception, Part III organizes the content of AIGE into five interconnected components: foundational concepts of artificial intelligence, AI methods, AI applications, the interdisciplinary integration of AI, and risk, ethics, and responsibility. These components follow the movement of AI from an object of scientific inquiry into social practice. In sequence, they address what AI is, how AI acquires its capabilities, how AI becomes a system operating in the real world, how AI enters other fields of knowledge, and how people should use, oversee, and govern AI. Together, they form a complete cognitive chain extending from identifying the object of inquiry and understanding its mechanisms, through explaining systems and transferring methods, to acting responsibly.

Foundational concepts of artificial intelligence give learners a frame of reference for understanding intelligent machines. The intellectual origins, emergence as a discipline, historical development, research aims, disciplinary characteristics, and conceptual boundaries of AI enable learners to situate current technologies within a longer trajectory. They can then distinguish artificial intelligence from related concepts such as automation, algorithms, robotics, machine learning, and generative AI, and assess a system's capabilities and limits in light of its actual functions and operating conditions.

AI methods explain how machine capabilities are formed. Fundamental problems such as representation, learning, reasoning, generation, and action, together with the relationships among data, knowledge, objectives, models, algorithms, and feedback, provide the mechanistic backbone for understanding AI technologies. Learning these methods enables learners to trace how a system arrives at its conclusions, understand the conditions on which its capabilities depend, analyze why a model may succeed or fail, and acquire a relatively stable explanatory framework for evaluating new technological developments.

AI applications connect abstract methods with lived experience. Once functions such as recognition, translation, recommendation, generation, and decision-making enter real-world settings, they combine with data, devices, software, users, organizational workflows, and institutional rules to form operational systems. The study of applications guides learners beyond visible functionality to understand how model outputs are produced, how they enter human judgment and action, and what evidence should support assessments of system performance, scope of applicability, and potential effects.

The interdisciplinary integration of AI examines how artificial intelligence participates in problem-solving and knowledge production in other fields. Mathematics, engineering, the natural sciences, the life sciences, the social sciences, and the humanities and arts differ in their objects of inquiry, forms of data, theoretical constraints, and standards of evidence. Learning in this area focuses on translating disciplinary problems into tasks that AI can address, selecting and adapting appropriate methods, incorporating domain knowledge and constraints, and validating results through the relevant professional procedures. The transfer of methods thereby remains inseparable from identifying applicable conditions and exercising disciplinary judgment.

Risk, ethics, and responsibility connect technical understanding with value judgment and responsible action. Questions involving the authenticity of information, data and privacy, fairness, safety, human agency, social consequences, and the allocation of responsibility require an independent and systematic body of knowledge. At the same time, these questions run through foundational concepts, AI methods, AI applications, and interdisciplinary integration, continually testing whether technical objectives are justified, whether evidence is sufficient, how benefits and risks are distributed, and what responsibilities different actors should bear.

The five components develop progressively while remaining connected through recurring feedback. Foundational concepts provide the language and frame of reference needed to understand methods; methods supply mechanistic explanations for analyzing applications; applications test methods within real-world systems; interdisciplinary integration constrains the transfer of methods through disciplinary problems and professional standards; and risk, ethics, and responsibility work backward from values and consequences to reassess technical objectives, system design, and modes of use. Cognitive autonomy runs through this entire process, while the four capacities of induction, generalization, judgment, and self-awareness are repeatedly exercised across different content areas and tasks.

The next five chapters develop the educational function, principles of content design, common core, and extended content of each component. Part III thus establishes the basic scope and organizational logic of what AIGE should teach. The proportion, depth, and sequence assigned to

the five components in a particular curriculum may be adjusted according to learners' ages, prior knowledge, disciplinary orientation, and practical needs. Building on this content framework, Part IV examines curriculum implementation across educational stages and learner groups.

Chapter 9

Foundational Concepts of Artificial Intelligence: A Coordinate System for Understanding Intelligent Machines

Abstract. Foundational concepts of artificial intelligence provide the coordinates needed for understanding the field. Without an overall grasp of what artificial intelligence studies, how it is realized, where it came from, and where its disciplinary boundaries lie, learners may equate today's popular products with artificial intelligence itself, mistake automated functions for intelligence, or take linguistic fluency as evidence of genuine understanding. Their views may then swing between uncritical enthusiasm for technology and fear of it. From the perspective of general education, this chapter discusses principles for selecting and organizing the foundational concepts of artificial intelligence and recommends a core body of content. Artificial intelligence can be understood as the science of using computers to simulate intelligent human behavior. Around this object of study, a rich body of engineering practice and interdisciplinary methods has also developed. Human intelligence provides an important point of reference, although machines need not reproduce human internal processes in full. The intellectual origins of artificial intelligence lie in the longstanding human aspiration to build intelligent machines. Formal and mathematical logic made it possible to analyze, symbolize, and calculate processes of thought. Computability theory and the general-purpose computer supplied the theoretical model and physical means for machines to execute such calculations. Turing's reflections on machine intelligence and the Dartmouth conference eventually helped establish artificial intelligence as an independent discipline. Symbolism, expert systems, statistical machine learning, deep learning, large models, and agents have since formed an expanding historical trajectory. Education in foundational concepts should help learners distinguish artificial intelligence from automation, robots, machine intelligence, algorithms, the internet, big data, machine learning, deep learning, generative AI, and artificial general intelligence, while understanding the relations of inclusion, overlap, and support among them. History, definition, and conceptual boundaries thereby work together in the service of cognitive autonomy: learners can locate new technologies within a long-term process of development, inquire into the conditions under which their capabilities arise, calibrate their trust in machines, and maintain sound judgment in a rapidly changing intelligent society.

Keywords: foundational concepts of artificial intelligence; origins of artificial intelligence; history

of artificial intelligence; disciplinary characteristics; conceptual boundaries; intelligent behavior; machine intelligence; cognitive autonomy

9.1 Why Foundational Concepts Are Indispensable

Many programs in Artificial Intelligence General Education begin with a particular tool or application. Chatbots answer questions, image models generate pictures, recommender systems continually deliver content, and autonomous vehicles perceive the road. Such visible functions readily arouse interest, but they may also lead learners to understand artificial intelligence as whatever product happens to be before them. When the product is superseded, the understanding built around it loses its point of reference. When a new model displays greater capabilities, learners may again regard it as an entirely new technology with no connection to the past.

Foundational concepts establish a relatively stable coordinate system for technologies that continue to change. At a minimum, this coordinate system should answer six questions:

1. What does "intelligence" mean in the term "artificial intelligence"?
2. What kinds of problems does artificial intelligence seek to solve?
3. Why does artificial intelligence rely primarily on computation for its realization?
4. From what intellectual and technological foundations did the discipline emerge?
5. What major shifts in method has artificial intelligence undergone over its history?
6. How is artificial intelligence related to neighboring concepts such as automation, robots, algorithms, machine learning, and generative AI?

The answers to these questions shape how learners explain artificial intelligence. Someone who equates AI with chat models will struggle to understand why expert systems, search algorithms, machine vision, and robotic planning all belong to the same field. Someone who calls every self-operating device "AI" will be unable to distinguish preset programs, feedback control, and machine learning. Someone who directly equates a model's linguistic performance with understanding, consciousness, or personhood may draw unwarranted conclusions about machine capabilities.

Education in foundational concepts also has to counter **technological presentism**: understanding a technology solely through its present form while overlooking the ideas, methods, and disputes accumulated over its history. The development of artificial intelligence spans multiple approaches, including logical reasoning, knowledge representation, probability and statistics, neural networks, reinforcement learning, and large-scale pretraining. Today's large models rest on decades of work in language modeling, representation learning, computing hardware, and data accumulation. Understanding this process allows learners to view a prominent technology as a stage in a historical chain, recognizing both its breakthroughs and its unresolved problems.

Foundational concepts provide a shared language and standards of evaluation for the technical content that follows. They also enable learners to judge the position, capabilities, and boundaries of a new technology when it appears. A course may enter through a tool, an application, or a problem situation, but it should return in due course to a stable conceptual coordinate system.

9.2 What Is Artificial Intelligence?

There is no single definition of artificial intelligence. For example, McCarthy (2007) later described it as the science and engineering of making intelligent machines, especially intelligent computer programs. For general education, a definition should be concise enough to avoid excessive terminology while still identifying the discipline's foundational characteristics. As discussed in Part II, this book defines artificial intelligence as follows:

> **Artificial intelligence is the science of using computers to simulate intelligent human behavior.**

The most important element of this definition is the simulation of human "intelligent behavior." Artificial intelligence has long taken human intelligence as a point of reference, but completing a task as a human might and fully reproducing human internal processes are distinct aims. One reason the Turing test uses behavioral performance as its criterion is that internal human processes of intelligence cannot be observed directly (Turing, 1950). Modern AI likewise often follows paths of realization different from those of human beings: computers evaluate game positions through search at scale, represent word meanings with high-dimensional vectors, and adjust model parameters through gradient-based optimization. These computational processes clearly differ from the internal mechanisms of human thought, even though the two can be compared in terms of such problem structures as tasks, representations, evidence, objectives, and feedback.

Education in foundational concepts should preserve this distinction. A machine may display linguistic competence without possessing the same lived experience as a human being. It may generate language that expresses emotion without thereby establishing that it has subjective feeling. It may produce a correct answer while lacking an understanding of the answer's source and real-world consequences. Assessments of machine intelligence should therefore examine functional performance, internal mechanism, environmental adaptation, and capacity for responsibility separately.

9.3 The Intellectual and Technological Origins of Artificial Intelligence

Artificial intelligence became an independent discipline in the mid-twentieth century, following a long period of intellectual preparation. Education in foundational concepts need not cover every historical figure. It can instead be organized around three main threads: the aspiration to build intelligent machines, the formalization of thought, and the mechanization of computation.

1. The aspiration to build intelligent machines

Human beings have long attempted to construct devices that operate automatically, imitate living creatures, or perform work on their behalf. Ancient stories of automata and mechanical animals, together with later water clocks, automatic theaters, mechanical orchestras, and programmable automatic devices, all express an enduring imagination of intelligent machinery.

These devices relied primarily on mechanical structures, hydraulics, gears, and preset sequences. They did not yet possess the computational structure of modern AI, but they posed a question that persists today: can machines undertake complex activities previously performed by human beings?

The history of automata serves a comparative conceptual purpose in the curriculum. It helps learners distinguish "automatic execution" from "intelligent adaptation" and understand that a high degree of automation does not necessarily involve learning, reasoning, or an understanding of the environment. This history also shows that artificial intelligence did not arise by accident. It continues the longstanding human aspiration to extend human capabilities through artificial devices and technical tools.

2. The formalization of thought: from logic to mathematical logic

The first theoretical thread in artificial intelligence arose from the study of the laws of thought. Aristotle's syllogism distinguished the form of thought from its content: a conclusion is guaranteed only when the form of inference is valid and its premises are true. This distinction retains importance for both artificial intelligence and cognitive autonomy. A system's reasoning may be formally valid even when its input facts are false; a generated answer may be well structured even when its evidence is unreliable.

In the nineteenth century, Boole (1854) used symbols and algebraic operations to describe logical relations, making it possible to mathematize processes of reasoning. The subsequent work of Frege, Russell, Whitehead, Hilbert, Gödel, and others contributed to the formation of mathematical logic. Forms of thought were transformed from rules expressed in natural language into well-defined symbolic systems and calculi, laying a theoretical foundation for machine reasoning.

This historical trajectory contains an important epistemological shift: some forms of human thought can be analyzed as formal structures; formal structures can be represented by symbols; and symbols can be manipulated according to rules. Concrete devices supplied the conditions for realizing this transformation, giving artificial intelligence the possibility of converting intelligence into a computational problem.

3. The mechanization of computation: computability theory and the general-purpose computer

Once thought could be formalized, a general-purpose machine was still needed to execute formal operations. In 1936, Turing (1936) proposed an abstract model of computation that described a general computational process in terms of finite states, read–write operations, and a table of rules. The significance of the Turing machine lies in its clear model of what it means for something to be "computable" and in its disclosure of the theoretical limits of computation.

Shannon (1938) connected Boolean algebra to relays and switching circuits, showing that logical operations could be realized in circuitry. The stored-program concept further allowed programs to be retained in a machine as data, enabling a computer to handle different tasks by changing its program (Neumann, 1945). The emergence of the general-purpose electronic computer transformed the use of computation to simulate intelligence from a philosophical and mathematical proposal into an experimentally tractable engineering program.

Computability also reminds learners that computing capability has limits. Improvements in processing speed and storage capacity make it possible to solve more complex tasks, but they do not eliminate problems that are theoretically uncomputable. Because artificial intelligence is built on computational systems, its capabilities are jointly constrained by forms of representation, algorithmic complexity, data resources, and computability.

4. Turing's conception of machine intelligence

Turing's contribution to artificial intelligence extended beyond computation theory. He discussed ideas concerning machine learning, reward and punishment, simulated evolution, and the "child machine." In 1950, he also proposed the Turing test, transforming the philosophical question of whether machines can think into an observable behavioral test (Turing, 1950).

The educational value of the Turing test lies in its operational criterion: when we cannot directly observe whether a machine possesses a mind, we can begin by examining its behavioral performance. Its limitations are equally important. Linguistic performance can serve as evidence of intelligence, but it cannot by itself establish genuine understanding, environmental experience, consciousness, or a capacity for responsibility. Modern generative AI has reactivated this debate, making the distinction between "behaving as though intelligent" and "possessing what kind of intelligence" a central issue in education about foundational concepts.

9.4 The Formation of Artificial Intelligence as an Independent Discipline

The Dartmouth conference of 1956 is commonly regarded as marking the formation of artificial intelligence as a discipline. Work on game playing, theorem proving, neural networks, and machine learning was already underway. The conference brought these dispersed efforts under the shared name of "artificial intelligence" and raised questions concerning language, abstraction, neural networks, self-improvement, creativity, and computational complexity (McCarthy et al., 1955).

The formation of a discipline requires an object of study, a body of problems, a community organized around shared methods, and an institutional setting. The significance of the Dartmouth conference can be understood in four respects:

1. It established the realization of intelligence in machines as a common research objective.
2. It organized work in logic, computation, psychology, neuroscience, and engineering within a shared problem space.
3. It gave shape to several lines of research that would exert a lasting influence.
4. It helped bring about the gradual establishment of laboratories, academic conferences, journals, and systems for educating researchers.

Artificial intelligence thus had a distinctly interdisciplinary origin. Mathematics supplied tools of logic, probability, and optimization; computer science supplied algorithms and systems; psychology and cognitive science supplied models of human thought; neuroscience offered inspiration from biological intelligence; and cybernetics and engineering supplied mechanisms of feedback and action. Artificial intelligence retained these open boundaries after becoming an

independent discipline.

9.5 A History Organized Around Shifts in Method

The history of artificial intelligence can easily become a chronological list of events and people. General education has greater need to draw out shifts in method: how researchers understood intelligence, where they believed knowledge came from, how systems adapted to their environments, and why particular approaches rose or faltered at particular times.

1. Symbolism: giving knowledge and reasoning to machines

The dominant approach in early artificial intelligence represented knowledge symbolically and solved problems through search, rules, and logical inference. Game-playing programs, automated theorem provers, general problem solvers, and early dialogue programs demonstrated the ability of computers to execute complex rules. Newell and Herbert A. Simon (1976) treated computer science as an empirical discipline concerned with complex information-processing systems and used the physical symbol system hypothesis to explain intelligent behavior.

Symbolic methods offer clearly structured representations and comparatively traceable reasoning processes, making them well suited to problems with explicit rules and expressible knowledge. They also face evident difficulties. Commonsense knowledge in open environments cannot readily be exhausted, uncertain information cannot always be fully described by strict rules, and a search space may grow rapidly with the size of a problem. Early research underestimated these difficulties. The resulting gap between technological promises and actual capabilities contributed to the first AI winter.

2. Knowledge engineering: from general reasoning to domain expertise

Expert systems organized the experience of specialists in a particular domain into a knowledge base and used inference programs to solve problems of diagnosis, analysis, and decision-making. They showed that powerful capabilities often arise from rich domain knowledge rather than from a single general reasoning mechanism. Feigenbaum (1977) characterized the work of this period as knowledge engineering and emphasized the centrality of knowledge to intelligent systems.

Expert systems brought artificial intelligence into real industries while exposing the difficulty of acquiring and maintaining knowledge. Expert experience contains substantial tacit knowledge that resists complete conversion into rules. As a knowledge base grows, old and new rules may conflict; changing environments also require continuing revision. A second AI winter led researchers to reconsider the limits of manually constructed knowledge.

3. Statistical machine learning: enabling systems to form regularities from data

From the 1990s onward, probability, statistics, and machine learning gradually moved into the mainstream. Machine learning expanded the source of knowledge from manually written rules to data. By estimating model parameters through training, systems could handle noise, uncertainty, and complex patterns. Research aims also became more pragmatic, as specific tasks in speech recognition, image recognition, information retrieval, autonomous driving, and other areas drove methodological development.

This transition had important epistemological significance. Knowledge no longer existed solely

in the form of readable rules; it could also be implicit in statistical relations and model parameters. Model capabilities came to be jointly determined by data, representation, objective functions, and training algorithms. Artificial intelligence acquired a stronger capacity for empirical induction while encountering new problems involving data bias, interpretability, and generalization.

4. Connectionism and deep learning: forming multilayer representations automatically

Neural networks realize complex functions through connections among large numbers of simple units. The perceptron demonstrated that connection weights could be learned from examples, while also revealing the expressive limitations of a single-layer network (Minsky and Seymour Papert, 1969). Backpropagation advanced the training of multilayer networks (Rumelhart, G. E. Hinton, and Williams, 1986); the accumulation of data, GPU computing, and improvements in network architecture eventually produced breakthroughs in deep learning. Deep networks can form representations layer by layer from raw data, reducing dependence on manually designed features in image, speech, and language tasks (LeCun, Yoshua Bengio, and G. Hinton, 2015).

The rise of deep learning expanded the boundaries of AI capability and changed how systems could be understood. Internal model representations are often difficult to interpret directly. Performance during training may diverge from performance in real deployment, while adversarial examples, distribution shift, and resource consumption have become new foundational concerns.

5. Large models and agents: from individual tasks to general interfaces

The Transformer improved long-sequence modeling and parallel training through the attention mechanism (Ashish Vaswani et al., 2017). With large-scale data and computing resources, pretrained models could transfer across many tasks and gradually develop capabilities in language generation, multimodal understanding, reasoning, code, and tool use. Research on foundation models characterizes this change in terms of a single model being adapted to many downstream tasks (Bommasani, Hudson, Adeli, et al., 2021).

Generative AI has shifted human interaction with models from invoking specialized functions to issuing instructions in natural language. Agents add goals, memory, planning, tool use, and environmental feedback to a model, enabling a system to carry out tasks over multiple steps. At this stage, artificial intelligence is expanding from individual models into complex systems, further blurring the boundaries among models, software, tools, and acting entities.

The ideas and methods accumulated throughout the development of artificial intelligence have inherited from one another and continue to coexist. Modern systems still use rules, search, knowledge bases, probabilistic models, and neural networks. This history records shifts in the center of research and in the structure of capabilities; it does not describe the complete replacement of older methods.

9.6 Distinguishing Easily Confused Concepts

Conceptual distinctions cannot rest on a single static definition. A more effective approach compares objects of study, means of realization, and relations among concepts.

Table 9.1: Artificial intelligence and neighboring concepts

Concept	Core meaning	Relation to artificial intelligence
Intelligent machine	An artificial device that displays some intelligent function; its scope depends on what human beings regard as "intelligence"	It may be realized through mechanics, circuitry, control, artificial intelligence, or a combination of these; AI is often one component of an intelligent machine
Automation	The automatic operation of a system according to a preset process or feedback mechanism, with an emphasis on reducing human operation	Automation need not involve learning or reasoning; AI can enhance an automated system's capacities for perception, adaptation, and decision-making
Robot	A machine system with capacities for perception, control, and physical action	A robot is a physical embodiment, and AI may serve as its perceptual or decision-making module; many AI systems have no robotic body
Machine intelligence	Intelligent capabilities displayed by a machine, emphasizing that the capabilities belong to the machine	It overlaps substantially with AI; "machine intelligence" emphasizes the resulting capability, whereas "artificial intelligence" also names the discipline and technological system that studies such capabilities
Algorithm	A definite sequence of steps or computational rules for solving a class of problems	Algorithms are constituent tools of AI; many ordinary algorithms do not involve intelligent behavior, and an AI system typically combines multiple algorithms
Internet	Infrastructure that connects computing devices and enables information transmission and sharing	It supplies AI with environments for data, computation, and services, while receiving support from AI technologies for search, recommendation, and security

Continued on next page

Table 9.1 continued

Concept	Core meaning	Relation to artificial intelligence
Big data	Data of great volume, variety, and velocity, together with technologies for processing it	Data are an important resource for modern machine learning; possessing a large quantity of data does not itself amount to an AI capability
Machine learning	A collection of methods that enable a system to improve its performance through data or feedback	It is an important methodological branch of AI; AI also includes knowledge representation, search, planning, reasoning, robotics, and other areas
Deep learning	Machine learning methods based principally on multilayer neural networks	It is a branch of machine learning and an important technical foundation of modern AI, but it does not encompass the whole of artificial intelligence
Generative AI	Systems that learn data distributions and generate text, images, sound, video, code, or other content	It is a methodological and applied form of AI; many systems for recognition, prediction, planning, and control do not aim to generate content
Large language model	A model pretrained on large-scale textual and related data that can process and generate language	It is one kind of foundation model used in generative AI; large models may also address vision, sound, scientific data, and multimodal information
Agent	A system that can pursue a goal by perceiving an environment, planning steps, invoking tools, taking action, and using feedback	It commonly combines a model as its cognitive core with memory, tools, permissions, and workflows; a chat interface does not by itself constitute a complete agent

Continued on next page

Table 9.1 continued

Concept	Core meaning	Relation to artificial intelligence
Narrow AI	An AI system whose capabilities are developed within a delimited task or domain	Most AI in practical use today belongs to this category; its capabilities may be powerful, but the scope of transfer remains constrained by the task and environment
Artificial general intelligence	A hypothetical or developing goal of creating systems able to learn, transfer, and solve problems across a wide range of tasks	It lacks an agreed definition and uniform testing standard; discussion should distinguish a research goal, claims made about a system, and capabilities that have been empirically validated
Superintelligence	A proposed form of intelligence that would substantially surpass human beings in virtually all important cognitive tasks	It is used primarily in discussions of long-term research and governance; its arrival should not be inferred directly from present performance on individual tasks

These concepts often stand in relations of inclusion and overlap. Deep learning is a branch of machine learning, and machine learning is an important methodological branch of artificial intelligence. Large language models belong both to foundation models and to a part of generative AI. An agent may invoke a large language model while also combining rules, search, databases, and external tools. No single hierarchical tree can express all these relations. A course should therefore use hierarchical relations, system composition, and comparative cases together.

9.7 Principles for Designing Content on Foundational Concepts

Organize rapidly changing terminology around stable questions: A course should organize content around enduring questions of intelligence, computation, representation, learning, action, evaluation, and responsibility. New products and models may serve as cases, but they should not become the center of the conceptual structure. When students encounter a new term, they can then ask: what intelligent task does it address, what representation and method does it use, what resources does it depend on, and what has changed in comparison with existing technology?

Make history serve conceptual understanding: A history of AI should explain why ideas and methods changed. The Dartmouth conference matters because it helps explain how a disciplinary community formed. The two AI winters illuminate the relation among visions, technological conditions, and social expectations. Symbolism, machine learning, and deep

learning reveal changes in the sources of knowledge and the means by which intelligence is realized. Historical figures and dates support this logic; they should not become isolated facts for memorization.

Establish conceptual boundaries through comparative cases: Conceptual boundaries are well suited to learning through sets of cases. A mechanical timer, a feedback-controlled thermostat, a robotic vacuum cleaner, and an autonomous learning robot can be compared in terms of degree of automation, environmental perception, learning, and planning. A keyword-matching program and a large language model can illuminate differences among rules, statistical learning, and behavioral performance. A search engine, a recommender system, and a generative model can distinguish retrieval, ranking, and generation.

Comparison should allow for intermediate cases. Real systems often combine automatic control, conventional software, and AI models. A simple choice between "AI" and "not AI" can obscure the structure of a system. More valuable questions ask which components use AI, what kind of capability they employ, and how that capability was formed.

Present capabilities together with their conditions of validity: "What can the model do?" should be taught together with "Under what conditions can it do so?" The performance of face recognition depends on the populations represented in the data, lighting, viewing angles, and adversarial conditions. A language model's answers depend on training information, context, retrieval, and prompts. The capability of an autonomous vehicle depends on its operational road domain, weather, sensors, and takeover mechanism. Attention to conditions counters the tendency to generalize results from a local experiment to every environment.

Distinguish facts, disputes, and future visions: Foundational concepts of AI encompass established facts as well as open disputes. The formation of artificial intelligence in the mid-twentieth century and the status of machine learning as one of its important methodological branches are relatively stable knowledge. Whether machines genuinely understand language and whether consciousness can arise from computation remain matters of theoretical dispute. Timelines for artificial general intelligence and superintelligence are highly uncertain predictions. A course should mark these three kinds of content so that learners develop appropriately differentiated impressions of their evidential status.

Connect foundational concepts to subsequent content: Foundational concepts should provide interfaces to AI methods, AI applications, interdisciplinary integration, and risk, ethics, and responsibility. A definition of AI should introduce the two broad families of knowledge-based and learning-based methods. A discussion of disciplinary characteristics should lead into typical applications and the field's interdisciplinary nature. A comparison of machine and human intelligence should introduce questions of responsibility and human–AI relations. Education in foundational concepts thereby becomes the entry point to the entire content system.

9.8 Recommended Content and Its Levels

The foundational concepts of artificial intelligence can be organized into a common core and extension topics.

9.8.1 Common Core

The common core recommended for all learners includes:

1. the multidimensional character of intelligent behavior and the distinction between local capabilities and intelligence as a whole;
2. a working definition of artificial intelligence, its research objectives, and its computational realization;
3. the significance of human intelligence as a point of reference and the limits of anthropomorphizing machines;
4. the line of origin formed by formal logic, mathematical logic, computability theory, and the general-purpose computer;
5. Turing's ideas about machine intelligence and the significance of the Dartmouth conference for the formation of the discipline;
6. the major methodological shifts associated with symbolic AI, expert systems, statistical machine learning, deep learning, large models, and agents;
7. the scientific character of artificial intelligence and the engineering practices and interdisciplinary methods developed around it;
8. the relations of artificial intelligence to automation, robots, algorithms, the internet, big data, machine learning, deep learning, and generative AI;
9. evidential and conceptual distinctions among narrow AI, artificial general intelligence, and superintelligence;
10. the conditionality and fallibility of AI capabilities, together with their social effects.

9.8.2 Extension Topics

Depending on learners' prior knowledge and the aims of the course, further topics may include:

1. the biological foundations, social cooperation, language, and cumulative culture of human intelligence;
2. relations among logical validity, the truth of premises, and the verification of generated content;
3. computability, complexity, and the limits of AI capabilities;
4. thought experiments concerning understanding and mind, such as the Turing test and the Chinese room;
5. schools of research including symbolic AI, connectionism, Bayesian approaches, and evolutionary methods;
6. differences between the behaviorist and internalist routes in their research aims and standards of evidence;
7. changes in disciplinary boundaries brought about by foundation models, multimodal systems, and agents;
8. varying definitions, evaluation schemes, and governance questions concerning artificial general intelligence.

Extension topics should remain conceptually oriented. Learners need not memorize every

figure or term; the central aim is to develop the capacity to compare different views and methods.

9.9 Foundational Concepts and Cognitive Autonomy

Inductive capacity is exercised when learners generalize from historical materials and technological cases to identify the objects of AI research, the sources of its capabilities, and its shifts in method, avoiding the treatment of a single product or local performance as representative of artificial intelligence as a whole.

Capacity for generalization is exercised when learners apply conceptual relations to new technologies. They can judge whether a new system depends principally on rules, learning, generation, or an agent architecture, and analyze the range within which existing concepts apply in the new setting.

Capacity for judgment is exercised when learners distinguish automation from learning, linguistic performance from genuine understanding, and a model from a complete application system. They can also specify the conditions under which a capability holds by reference to data, objectives, computational resources, and environment.

Capacity for self-awareness is exercised when learners notice how mystification, anthropomorphism, and promotional claims about technology affect their own trust. They avoid placing excessive trust in a system because its language is fluent, while also resisting the complete dismissal of a technology because of a single error. They can distinguish verified facts, disputed claims, and future visions.

Foundational concepts thus help learners develop a stable yet open orientation: understanding the historical continuity and practical value of artificial intelligence, confronting changes introduced by new methods, and continuing to inquire into sources, evidence, scope of applicability, and responsibility.

9.10 Chapter Summary

The central task of the foundational concepts of artificial intelligence is to establish a coordinate system for understanding intelligent machines. Artificial intelligence can be understood as "the science of using computers to simulate intelligent human behavior." It takes human intelligence as an important point of reference and examines functions such as perception, action, thought, and emotion, while using computational processes different from those of human beings to realize these capabilities.

The intellectual and technological origins of artificial intelligence can be understood in three stages. The longstanding aspiration to build intelligent machines posed the problem. Formal and mathematical logic made it possible to symbolize and calculate some processes of thought. Computability theory, digital circuitry, and the general-purpose computer enabled machines to execute those calculations. After artificial intelligence became an independent discipline, the center of its methods shifted through symbolic AI, expert systems, statistical machine learning, deep learning, large models, and agents. These approaches continue to coexist, and modern

systems frequently combine multiple kinds of tools.

The field of artificial intelligence encompasses many concepts that may be related through inclusion, overlap, or support. Stable questions, historical trajectories, comparative cases, explicit conditions, and distinctions among levels of evidence can transform conceptual clarification into training in induction, generalization, judgment, and self-awareness. Conceptual learning thereby becomes a foundational practice for cognitive autonomy.

Chapter 10

Foundational Methods of Artificial Intelligence: From Knowledge Representation to Agent Systems

Abstract. Artificial intelligence encompasses a wide range of methods. If a course proceeds item by item through algorithm names and product forms, learners may acquire scattered terminology without understanding why the methods arose, how they relate to one another, or under what conditions they work. This chapter constructs a system of foundational AI methods from the perspective of general education. Artificial intelligence develops capabilities along two basic paths. Knowledge-based methods represent facts, rules, and relations, then solve problems through search and reasoning. Learning-based methods adjust models on the basis of data, feedback, and experience, enabling systems to perform tasks. Machine learning can be understood through five elements—objective, model, algorithm, data, and knowledge—and through a complete process comprising problem definition, model design, model training, independent testing, model selection, deployment monitoring, and feedback-driven updating. Supervised, unsupervised, and reinforcement learning are classified by their learning signals, while four methodological traditions—symbolic, Bayesian, connectionist, and evolutionary and bio-inspired—present AI methods from the perspectives of knowledge representation, the treatment of uncertainty, and mechanisms of capability formation. Deep learning achieves representation learning through multilayer neural networks. Large models develop transferable capabilities through large-scale pretraining; multimodal models bring text, images, sound, and action into joint representations; and agents add goals, memory, planning, tools, action, and feedback. Modern AI is reintegrating knowledge and learning: domain knowledge, retrieval systems, symbolic constraints, physical laws, and human feedback all participate in system design. Education in foundational methods gives learners a coordinate system for understanding methods, the sources of model capabilities, and the fit among data, objectives, structures, evaluation, and contexts of use. It thereby supports cognitive autonomy through induction, generalization, judgment, and self-awareness.

Keywords: AI methods; knowledge-driven AI; machine learning; supervised learning; Bayesian methods; connectionism; evolutionary computation; deep learning; large models; multimodal models; agents; cognitive autonomy

10.1 Why AI Methods Need to Form a System of Knowledge

AI methods span logic, search, probability, optimization, neural networks, language models, and agents. When students encounter this material, they can readily form one of two partial views. One identifies AI methods with whichever deep-learning techniques or large models are currently most prominent, overlooking the longstanding roles of knowledge representation, search, reasoning, and probabilistic methods. The other treats the study of methods as the mastery of a set of algorithmic formulas, without connecting algorithms to data, objectives, evaluation, and real contexts of use.

General education needs to address the foundational questions at the level of method:

1. Where do machine capabilities come from?
2. How does human knowledge enter an AI system?
3. How does a machine derive regularities from data and feedback?
4. How is a model designed, trained, tested, and selected?
5. Why are different model structures suited to different kinds of data?
6. What do large models, multimodal models, and agents add to existing methods?
7. How can we judge whether a method is suited to a real problem?

These questions are more stable than the names of particular algorithms. Convolutional networks, Transformers, and future architectures will change, but objectives, representations, data, models, learning, and evaluation will remain among the basic relations that every AI system must address.

Foundational AI methods can therefore be organized into three levels.

The first consists of two basic paths: knowledge-based AI and learning-based AI. The former represents knowledge explicitly and enables a machine to search and reason according to rules. The latter supplies objectives, data, or feedback from which a machine learns a model.

The second level is the general system of machine learning. It includes the five elements of objective, model, algorithm, data, and knowledge, together with the complete process of model design, training, testing, selection, deployment, and feedback. Supervised, unsupervised, and reinforcement learning distinguish ways in which a machine receives feedback according to its learning signal. Intersecting with this classification, four methodological traditions—symbolic, Bayesian, connectionist, and evolutionary and bio-inspired—present AI methods from the perspectives of knowledge representation, model assumptions, and mechanisms of capability formation.

The third level comprises the principal extensions of modern AI: deep learning, large models, multimodal models, and agents. They expand representational capacity, the range of transfer, modes of interaction, and the capacity for sustained action, while remaining subject to the basic relations at the first two levels.

The value of this system lies in helping learners answer "why." Why does a particular system need a knowledge base? Why are convolutional structures appropriate for images? Why does training performance fail to establish real capability? Why do large models require pretraining? Why do agents need tools and feedback? Only by understanding these relations can learners locate emerging methods within a stable structure and judge them accordingly.

10.2 Two Basic Paths: Knowledge-Driven and Learning-Driven AI

10.2.1 Knowledge-Based Methods: Converting Human Understanding into Machine-Usable Structures

Human beings rely on knowledge to solve problems. Mathematical axioms, physical laws, experience in medical diagnosis, legal rules, and common sense can all guide judgment and action. Knowledge-based AI seeks to represent such knowledge in forms that machines can store and process, then derive conclusions through search and reasoning.

A knowledge-driven system typically comprises three stages:

Knowledge representation $\longrightarrow$ search or reasoning $\longrightarrow$ conclusion and explanation

Knowledge representation answers the question "What does the machine know?" Knowledge may be represented as logical formulas, production rules, semantic networks, ontologies, or knowledge graphs. Search and reasoning answer "How does the machine use what it knows?" A system may begin with facts and derive a conclusion step by step; work backward from a goal to identify the conditions it requires; or use heuristic information to explore more promising alternatives first among many candidate paths.

Early automated theorem proving and computer game playing demonstrated the power of general knowledge. The Logic Theorist proved mathematical theorems from axioms and rules of inference, while minimax search and α–β pruning used the rules of a game and evaluations of positions to select actions. These tasks have explicit rules, closed state spaces, and clearly defined goals, making them well suited to knowledge representation and search (Newell and Herbert A. Simon, 1976).

Expert systems went further by organizing domain experience into knowledge bases. Production rules commonly take the form "if a condition holds, then make a particular judgment or take a particular action," while an inference engine processes input facts according to those rules. Separating the knowledge base from the inference engine preserves a path of explanation and allows experts to inspect and revise the system (Feigenbaum, 1977). Knowledge graphs organize knowledge in terms of entities and relations, enabling relational queries, relational inference, and semantic retrieval.

Knowledge-based methods have three prominent advantages. First, their sources of knowledge are explicit and their reasoning processes are comparatively open to inspection. Second, a small body of rules can remain useful when data are scarce. Third, domain constraints can prevent a system from producing results that clearly violate requirements. Knowledge-based methods therefore retain an important place in high-risk or data-limited settings such as law, medicine, safety control, and scientific computing.

These methods also face a knowledge-acquisition bottleneck. Experts may perform a task skillfully without being able to articulate all their experience as rules. Real environments contain numerous exceptions, uncertainties, and tacit conditions. As a knowledge base grows, its rules may conflict and maintenance costs may rise. General axioms may be universal, yet

deriving complex realities from foundational laws can encounter combinatorial explosion and computational complexity. The capabilities of a knowledge-driven system are bounded by the knowledge human beings have made explicit.

10.2.2 Learning-Based Methods: Forming Models from Experience

Machine learning changed the way knowledge enters a system. Through a checkers program, Samuel (1959) demonstrated the possibility that a computer could improve its performance by using records of past games and self-play. The widely repeated phrase "without being explicitly programmed" can aid an introductory explanation, but should not be treated as a strict definition from that paper. Mitchell (1997) supplied a more formal definition: a program is said to learn from experience E with respect to a task T and performance measure P if its performance on T, as measured by P, improves with E.

Machine learning does not require a designer to write every rule of judgment. The designer specifies a learning objective, selects a model structure, prepares data and knowledge, and provides an algorithm for adjusting the model. Through training, the system accumulates patterns dispersed through the data in the parameters or structure of the model.

Learning-based methods expand the upper limit of knowledge available to AI. Machines can process data on a scale far beyond any individual's experience, discover high-dimensional relations that are difficult to express explicitly in language, and form representations unlike those of human experience. Many regularities in image recognition, speech recognition, recommendation, and the analysis of scientific data resist being written rule by rule by experts and are better learned from data.

Learning-based methods do not eliminate human design. Decisions about who sets the objective, how data are collected, which model is selected, how error is measured, and where a system is deployed all embody knowledge and value judgments. Machine learning delegates part of rule discovery to the model while relocating human responsibility to problem definition, data governance, evaluation, and system boundaries.

10.2.3 The Two Paths as Complements

Knowledge-driven and learning-driven AI can be understood as two basic sources of capability.

Table 10.1: Comparison of knowledge-driven and learning-driven methods

Dimension	Knowledge-based methods	Learning-based methods
Source of knowledge	Facts, rules, relations, and domain theories explicitly supplied by people	Data, demonstrations, rewards, environmental interactions, and prior knowledge

Table 10.1 (continued)

Dimension	Knowledge-based methods	Learning-based methods
Formation of capability	Search, matching, and logical or probabilistic reasoning	Adjustment of model structure or parameters so that performance improves with experience
Principal advantages	Interpretable and controllable; suited to problems with clear rules or scarce data	Able to process complex patterns and large-scale data, with the potential to discover new regularities
Principal limitations	Difficult knowledge acquisition, costly rule maintenance, and inability to exhaust open environments	Dependence on data and objectives, with possible bias, overfitting, and limited interpretability
Suitable contexts	Theorem proving, compliance checks, safety constraints, and question answering with specialist knowledge	Recognition, prediction, generation, recommendation, complex perception, and sequential decision-making

Modern systems frequently combine the two paths. Retrieval-augmented generation uses knowledge bases to supplement large models; neuro-symbolic systems combine the perceptual capabilities of neural networks with symbolic reasoning; scientific machine learning incorporates physical laws into model structures or loss functions; and agents use tool calls to access databases, calculators, and specialist software. The development of AI methods increasingly takes the form of a reintegration of knowledge and learning.

10.3 Five Basic Elements of Machine Learning

Machine learning can be understood through five elements: objective, model, algorithm, data, and knowledge. These are not independent components. Together they define what a system can learn, how it learns, and whether the resulting capability is reliable.

1. Objective: What exactly should the system optimize?

The objective directs learning. A classification system seeks to reduce category errors, a regression system to reduce prediction error, a generative system to produce content that better conforms to patterns in the data and to human requirements, and a reinforcement-learning system to increase long-term cumulative return.

To make an objective amenable to optimization, it is usually converted into a mathematical form known as a loss function or objective function. Training searches for model parameters that reduce the loss. An objective function compresses a real-world purpose into a computable metric and may discard important values in the process. Measuring recommendation quality by click-through rate, for example, can encourage attention-grabbing content; using medical

expenditure as a proxy for health needs can perpetuate social inequality. Education in methods should help learners understand the possible distance between an optimization objective and the real objective.

Objectives also exist at several levels. The local objective used to train a model, the operational objective of a system, and the public objective governing its social use may not coincide. A model may accurately predict what a user will click, while the platform must still judge whether such recommendations contribute to a healthy information environment. Technical optimization cannot substitute for scrutiny of the objective itself.

2. Model: What structure will preserve and use regularities?

The model is the principal structure of machine learning. It specifies how inputs are represented, how information is processed, how outputs are produced, and which parts can be adjusted through learning. A linear model represents relations among variables with a small number of parameters; a decision tree forms judgments through branching rules; a Bayesian network represents relations among variables as probabilistic dependencies; and a neural network uses a large number of connection weights to form a complex function.

A model structure embodies assumptions about the problem, also known as inductive biases. A linear model assumes that relations among variables can be approximated as linear. A convolutional network assumes that local patterns can be reused at different positions. A recurrent network assumes that the current state is related to past information. Greater complexity does not automatically make a model better: the model must fit the structure of the data, the scale of the task, computational conditions, and the cost of error.

Under its particular assumptions about the set of problems and the manner of averaging over them, the "no free lunch" theorem states that no optimization algorithm can universally outperform others when performance is averaged uniformly across all possible problems (Wolpert and Macready, 1997). This result does not mean that every algorithm performs equally well in the real world. Its educational significance is that the "best model" cannot be discussed apart from a task distribution and prior assumptions. Model selection must draw on domain knowledge and make explicit what structure one expects the data to contain.

3. Algorithm: How is the model adjusted?

An algorithm specifies how a model is updated from experience. For a neural network, a learning algorithm generally computes the gradient of the loss function with respect to the parameters and adjusts the weights in a direction that reduces the loss. A clustering algorithm iterates between grouping samples and updating cluster centers. In reinforcement learning, an algorithm uses rewards to estimate the value of actions and update a policy.

Gradient descent is among the most common optimization ideas in modern machine learning. Let the model parameters be θ and the loss function be $L(\theta)$. A basic update can be written as

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta_t),$$

where η is the learning rate and $\nabla_\theta L$ indicates the direction and magnitude of change in the loss with respect to the parameters. A learning rate that is too large may cause the model to

oscillate between more favorable regions; one that is too small makes training slow. Algorithm design balances speed, stability, accuracy, and resources.

A learning algorithm does not guarantee an absolute optimum. The loss surface of a complex model may contain saddle points, flat regions, and multiple local structures. Training results are also affected by initialization, data order, regularization, and numerical precision. Machine learning has the character of an experimental science, so algorithms must be assessed through repeated experiments and controlled comparisons.

4. Data: Experience, evidence, and boundaries

Data are the experiential material from which a model learns. Images, sounds, texts, sensor records, experimental measurements, and human behavior can all become data. Dimensions of data quality include at least accuracy, quantity, coverage, representativeness, timeliness, and lawfulness.

Incorrect labels teach a model incorrect answers. Too little data makes it difficult to distinguish stable regularities from chance phenomena. Inadequate coverage of contexts can cause failure in real environments. Uneven group distributions can create performance disparities among populations. Outdated data fail to represent the present environment, and unauthorized collection creates privacy and legal problems.

Data are also not pure facts that simply exist in nature. Choices are involved in what is recorded or omitted, how something is measured and classified, and who supplies the labels. Students need to understand data as evidence that has been collected, represented, and governed, rather than as an automatic equivalent of reality itself.

5. Knowledge: Providing priors and constraints for learning

Machine learning still requires knowledge. Domain knowledge helps in designing features and models, setting reasonable parameter ranges, selecting evaluation metrics, interpreting anomalous results, and constraining a system from violating basic laws.

Knowledge can enter a learning system in many ways:

1. through the selection of input variables and data representations;
2. through the design of models with particular structures;
3. by adding constraints to the loss function;
4. by supplementing information with a knowledge base or retrieval system;
5. by applying rule-based checks to outputs; and
6. through expert feedback and correction of results.

Knowledge is especially important when data are scarce, costly, or associated with high risk. Conservation laws in scientific research, clinical guidelines in medicine, and safety boundaries in engineering can all reduce the space a model must explore and improve the credibility of its results.

10.4 The Complete Process of Forming a Machine-Learning System

Model training is only one stage in the formation of a machine-learning system. A system intended for real use must also pass through problem definition, design, testing, selection, deployment, and continuing updates.

Problem definition → data and knowledge preparation → model design → model training → independent testing → model selection → deployment monitoring and feedback-driven updating

1. Problem definition and data preparation

Problem definition specifies the input and output, whom the system serves, how success will be measured, and what consequences errors may produce. An ill-defined problem yields a model that appears precise but lacks meaning.

Data preparation includes collection, cleaning, annotation, splitting, and documentation of provenance. The training set adjusts the model, the validation set may be used to select parameters and structures, and the test set supports final evaluation. Mixing the three can leak information and artificially inflate test performance.

2. Model design

Model design should make use of the structure of the data and task. Images have local and spatial relations, language has sequence and context, knowledge graphs have relations among entities, and scientific data may satisfy symmetries and conservation laws. The more accurately a model structure reflects such properties, the more efficient learning will generally be.

Design must also consider computational resources, response time, interpretability, privacy, security, and maintenance costs. The most accurate model is not necessarily the most appropriate for deployment. A mobile device may require a smaller model, while a high-risk system may place greater weight on explanation and stability.

3. Model training

Training converts the learning objective into a loss function and adjusts parameters through an optimization algorithm. The process requires monitoring of loss, performance, gradients, and resource consumption, together with methods such as regularization, data augmentation, and early stopping to reduce overfitting.

Training performance indicates only how well a model has adapted to experience it has already seen. A student who has memorized a set of exercises may still be unable to solve a new problem; likewise, a model that remembers its training samples has not necessarily formed transferable regularities.

4. Model testing and generalization

The central value of a model lies in its performance on new samples and in new environments, a capability known as generalization. An independent test set estimates performance on unseen data. Strong training performance followed by a marked decline in test performance usually indicates overfitting. Poor performance on both may indicate insufficient model capacity, low-quality data, or a problem with task definition.

Testing must also cover changes encountered in real environments. Lighting, noise, devices, regions, populations, and time can all change a data distribution. Performance on a benchmark dataset cannot be treated as direct evidence of real-world reliability.

5. Model selection

Model selection balances fitting capacity against complexity. A model that is too simple underfits and fails to capture basic regularities; one that is too complex may learn noise and chance exceptions. Occam's razor suggests that when several models perform similarly, the simpler model should be preferred. Simpler models are generally easier to explain, test, and maintain.

Selection also involves trade-offs among multiple objectives. Medical screening may give greater priority to avoiding false negatives; spam filtering must control false positives; autonomous driving must account for extreme conditions; and recommendation systems must also assess diversity and long-term effects. A single average accuracy cannot cover every requirement.

6. Deployment, monitoring, and updating

Once a model enters a real environment, both user behavior and the environment can change, and the model may in turn alter the data. A recommender system affects what users click, and those clicks become data for subsequent training, creating a feedback loop. Deployment requires monitoring performance degradation, anomalous inputs, security attacks, fairness disparities, and resource consumption.

Model updates should preserve records of versions, data, and evaluations. High-risk systems also need human review, appeal, and opt-out mechanisms. An AI system comprises models, software, data pipelines, users, and organizational institutions; model metrics alone cannot guarantee reliability.

10.5 Three Basic Modes of Learning

1. Supervised learning: learning from demonstrations with answers

Supervised learning uses labeled data. Each input sample has an expected output, and the model adjusts its parameters by comparing its prediction with the label. Classification predicts discrete categories, while regression predicts continuous values.

Supervised learning is suited to clearly defined tasks for which reliable labels can be obtained, such as identifying diseased leaves, predicting house prices, and classifying email. Its principal constraints arise from the cost and quality of labels. Labels may contain expert disagreements, historical biases, and measurement errors, all of which a model can learn.

2. Unsupervised and self-supervised learning: discovering structure within data

Unsupervised learning has no externally supplied answers. The system searches within the data for similarities, low-dimensional structure, and latent representations. Clustering groups similar samples, dimensionality reduction compresses high-dimensional data into more concise representations, and autoencoders learn important features by reconstructing their inputs.

Self-supervised learning constructs learning tasks from raw data itself. A language model may predict subsequent tokens from preceding text, for example, while an image model may

predict a masked region. Self-supervised learning uses vast quantities of unlabeled data to learn general representations, providing a foundation for the pretraining of large models (Devlin et al., 2019; T. B. Brown et al., 2020).

Unsupervised and self-supervised learning reduce dependence on human annotation, but they also create interpretive difficulties. Structures discovered by a model need not correspond to human concepts, and frequent patterns in data need not carry genuine meaning. Learning outcomes must be validated in relation to a task and domain knowledge.

3. Reinforcement learning: learning through action and feedback

Reinforcement learning studies how an agent learns a policy by interacting with an environment. The agent observes a state and selects an action; the environment returns a new state and a reward. The learning objective is usually to maximize long-term cumulative return (Sutton and Barto, 2018).

Reinforcement learning is suited to multistep decision-making, such as robot control, resource scheduling, and game playing. Reward design directs what a system learns. A reward that is too local may lead an agent to pursue immediate gains; a loophole in the metric may prompt a shortcut that the designer did not anticipate. Reinforcement learning displays with particular clarity the relations among objective setting, long-term consequences, and value alignment.

Table 10.2: Three basic modes of learning

Learning mode	Learning signal	Typical tasks	Principal risks
Supervised learning	Labels and target values supplied by people or systems	Classification, regression, detection, and prediction	Labeling errors and biases; cost of labeling
Unsupervised and self-supervised learning	Structure within the data or objectives constructed automatically from the data	Clustering, dimensionality reduction, representation learning, and pretraining	Difficulty interpreting learned structures; patterns may lack semantic meaning
Reinforcement learning	Rewards, penalties, and long-term returns from the environment	Game playing, control, planning, and sequential decision-making	Distorted reward design, short-term shortcuts, and risks of exploration

The three modes can be combined. AlphaGo used supervised learning from records of human games and then reinforcement learning through self-play. Large language models first undergo self-supervised pretraining and are then post-trained with human preferences or rules. The boundaries among learning modes are becoming more flexible, while the central question remains what feedback the system receives.

10.6 Four Methodological Traditions in Artificial Intelligence

Supervised, unsupervised, and reinforcement learning are classified principally by their **learning signals**: labels, structure within the data, or rewards from an environment. AI methods can also be understood from the perspectives of knowledge representation, the treatment of uncertainty, and mechanisms of capability formation. On this basis, they can be summarized in terms of four methodological traditions: symbolic, Bayesian, connectionist, and evolutionary and bio-inspired. The symbolic tradition lies chiefly along the knowledge-driven path, while the other traditions form models more often through data, probability, or search. The four traditions therefore should not be treated collectively as subcategories of machine learning.

This division is intended to establish an overall view; it is not an academic taxonomy with fixed boundaries. Many modern systems draw simultaneously on several traditions. An agent, for example, may use a neural network to understand language, Bayesian methods to estimate risk, symbolic rules to constrain action, and evolutionary search or reinforcement learning to improve its policy. The methodological traditions reveal where different approaches seek intelligence and what forms of knowledge and mechanisms of learning they employ.

1. The symbolic tradition: forming intelligence through symbolic representation and rule-governed operations

The symbolic tradition holds that intelligent activity can be represented as symbolic structures and operations upon them. Facts, concepts, relations, and rules are written explicitly into a system, and the machine obtains conclusions through matching, search, and reasoning. Logical inference, decision trees, rule learning, theorem proving, expert systems, and knowledge graphs all exemplify this tradition (Newell and Herbert A. Simon, 1976).

Symbolic methods have clearly structured representations, and their sources of knowledge and paths of reasoning are comparatively easy to inspect. In tasks with explicit rules, scarce data, or a need for strict constraints, they can directly use knowledge already possessed by human beings. Their difficulty lies in the impossibility of listing all knowledge in an open world and of converting every tacit rule embedded in experience into symbols. As the number of rules grows, conflicts, maintenance problems, and an expanding search space may also arise.

The symbolic tradition reminds learners that machine intelligence can arise from explicit knowledge and structures of reasoning. It also reveals a foundational problem: there is a gap between the knowledge that human beings can express and the knowledge they actually possess and use.

2. The Bayesian tradition: updating beliefs under uncertainty

The Bayesian tradition understands learning as the updating of belief in a hypothesis when evidence arrives. A system represents uncertainty with probability and uses Bayes' rule to combine prior knowledge with observed data. Naive Bayes classifiers, Bayesian networks, probabilistic graphical models, and hidden Markov models all belong to this methodological tradition (Pearl, 1988).

Let the hypothesis be H and the observed evidence be D. Bayes' rule can be written as

$$P(H \mid D) = \frac{P(D \mid H)P(H)}{P(D)}.$$

Here $P(H)$ is the prior assessment of the hypothesis before the evidence is observed, $P(D \mid H)$ is the likelihood of observing the evidence if the hypothesis holds, and $P(H \mid D)$ is the posterior assessment after incorporating the evidence. Learning thus takes the form of continuously revising belief as evidence changes.

Bayesian methods can represent uncertainty explicitly and are suited to diagnosis, prediction, risk analysis, and settings with limited data. They require a designer to specify probabilistic relations among variables or select appropriate priors. When there are many variables and complex dependencies, probabilistic inference can be computationally expensive; unsuitable priors and model assumptions also affect the result.

The educational value of the Bayesian tradition lies in showing that intelligent judgment generally forms a revisable degree of confidence from limited evidence. A system should adjust its judgment as new evidence arrives, and people likewise need to distinguish "highly probable" from "established."

3. The connectionist tradition: learning distributed representations through connection weights

Inspired by nervous systems, the connectionist tradition holds that complex intelligence can emerge from the connections and collective activity of many simple units. Knowledge need not be stored as readable rules; it can be distributed across network weights and activation patterns. Perceptrons, multilayer neural networks, convolutional networks, recurrent networks, and Transformers all continue this tradition (Rumelhart, G. E. Hinton, and Williams, 1986; LeCun, Yoshua Bengio, and G. Hinton, 2015).

Connectionist models adjust their weights on the basis of data and error signals, automatically learning representations in high-dimensional data. They have made prominent advances in vision, audition, language, and generative tasks and form the principal methodological foundation of contemporary deep learning and large models.

The connectionist tradition offers strong expressive capacity and can process complex patterns that resist description by handcrafted rules. Yet its internal knowledge is often dispersed across vast numbers of parameters and is difficult to explain directly. Models also depend on large quantities of data and computation and may learn biases and spurious correlations from the data.

This tradition shows that knowledge can exist in distributed form and that a model can form representations through experience. It also places generalization, interpretability, and data quality at the center of AI methodology.

4. The evolutionary and bio-inspired tradition: searching a solution space through variation, selection, and inheritance

The evolutionary and bio-inspired tradition draws on biological evolution and treats candidate solutions as a population. A system continually generates variations, retains better-performing

individuals through fitness evaluation, and passes their characteristics to the next generation. Genetic algorithms, evolutionary strategies, genetic programming, and some forms of neural architecture search exemplify this approach (Holland, 1975).

Evolutionary methods do not require a differentiable objective function or advance knowledge of the direction in which parameters should be adjusted. Provided that candidate solutions can be evaluated, the system can search progressively through selection. These methods are therefore suited to structural design, combinatorial optimization, multiobjective problems, and discontinuous search spaces.

Their limitations include the large number of evaluations that a search may require and the resulting computational cost. If the fitness function does not accurately represent the real objective, evolution may also discover solutions that depart from the designer's intent. Structures found by evolutionary algorithms can be effective without making the source of their effectiveness easy to explain.

The evolutionary and bio-inspired tradition reveals another mechanism of capability formation: designers specify processes of variation, selection, and inheritance rather than the final structure, allowing solutions to emerge through repeated competition.

5. Relations and integration among the four traditions

The four traditions answer the question "How does a machine acquire intelligence?" from different perspectives:

Table 10.3: Comparison of four methodological traditions in artificial intelligence

Methodological tradition	Core representation	Mechanism of capability formation	Typical advantages and limitations
Symbolic	Symbols, rules, logical relations, and knowledge structures	Search, matching, rule induction, and logical inference	Clear structure and ready constraint; difficult knowledge acquisition and adaptation to open environments
Bayesian	Probability distributions, random variables, and dependencies	Updating posterior probabilities by combining priors with evidence	Explicit representation of uncertainty; dependence on probabilistic assumptions and high cost of complex inference

Table 10.3 (continued)

Methodological tradition	Core representation	Mechanism of capability formation	Typical advantages and limitations
Connectionist	Neurons, connection weights, and distributed representations	Adjustment of network parameters using data and error	Strong expressive capacity; high demand for data and computation and limited interpretability of internal mechanisms
Evolutionary and bio-inspired	Individuals, populations, encoded structures, and fitness	Variation, selection, recombination, and search across generations	No dependence on gradients and suitable for structural search; expensive evaluation and potentially opaque results

The four traditions do not form mutually exclusive categories. Symbolic rules can constrain neural-network outputs, Bayesian methods can represent uncertainty in neural models, evolutionary algorithms can search for network architectures, and connectionist models can generate symbolic candidates from data. The current prominence of deep learning reflects the development of the connectionist tradition, while the reliability and transferability of modern AI increasingly depend on the integration of multiple methods.

Understanding these traditions helps prevent the reduction of AI methods to neural networks. Faced with a real problem, learners need to ask whether knowledge can be represented explicitly, how much uncertainty exists in the environment, whether sufficient data are available, whether the objective function is differentiable, and whether the search space is suited to evolution. Method selection is determined by the structure of the problem.

10.7 Deep Learning: From Feature Design to Representation Learning

The connectionist tradition has become a major line of AI development over the past two decades. Deep learning follows this tradition, using multilayer neural networks and large-scale data to learn complex representations. Its basic ideas still need to be situated within the larger structure in which four traditions coexist and interweave.

1. The basic idea of artificial neural networks

Inspired by biological nervous systems, artificial neural networks form models from large numbers of simple computational units and adjustable connections. McCulloch and Pitts (1943) developed an early model of a neuron based on weighted inputs and thresholded outputs. The perceptron introduced by F. Rosenblatt (1958) included learnable weights, enabling a network to

adjust a classification boundary from examples.

A single-layer perceptron can handle only linearly separable problems. Hidden layers and nonlinear activations enable multilayer networks to express more complex relations. Backpropagation efficiently calculates how parameters in each layer affect the loss, making it possible to train multilayer neural networks through gradient descent (Rumelhart, G. E. Hinton, and Williams, 1986).

Deep learning generally refers to the use of multilayer neural networks for representation and task learning. Its key change is that a model can form features layer by layer from raw data. Lower layers in an image network may respond to edges and textures, while higher layers combine them into object parts and categories. A language model develops representations of syntax, semantics, and context progressively from characters and tokens (LeCun, Yoshua Bengio, and G. Hinton, 2015).

2. Structural design embodies an understanding of data

Neural-network structures are not arbitrary stacks of layers. Each architecture writes an understanding of the properties of its data into the model.

Convolutional neural networks use local connectivity and weight sharing, making them suitable for image data with local patterns and spatial repetition. A convolutional kernel detects the same feature at different positions, while multiple layers combine edges, textures, and local shapes into higher-level representations. AlexNet's breakthrough in the ImageNet competition demonstrated the power of combining deep convolutional networks, large-scale data, and GPU computing (Krizhevsky, Sutskever, and G. E. Hinton, 2012).

Recurrent neural networks accumulate past information in a hidden state and are suited to language, speech, and time-series data. Their limitations include difficulty with long-range dependencies and sequential computation. A Transformer uses self-attention to establish direct relations among different positions in a sequence and supports more efficient parallel training (Ashish Vaswani et al., 2017).

Autoencoders learn latent representations by compressing and reconstructing their inputs. Generative adversarial networks learn a data distribution through competition between a generator and a discriminator (I. Goodfellow, Pouget-Abadie, Mirza, et al., 2014). These architectures embody different assumptions concerning locality, sequence, compression, and competition.

This relation illustrates an important methodological principle: a model structure should make use of the intrinsic properties of its data and task. A general-purpose network with great theoretical expressive capacity may perform less well under limited data and resources than a specialized architecture with an appropriate inductive bias.

3. Conditions for the success of deep learning

The rise of deep learning was jointly driven by several conditions:

1. large-scale data supplied material for learning;
2. GPUs and parallel computing accelerated training;
3. backpropagation, initialization, activation functions, regularization, residual connections, and related methods improved the training of deep networks; and
4. benchmark datasets and open tools promoted comparison and reproducibility.

The capabilities of deep learning cannot be attributed simply to "deeper networks." Data, computational capacity, algorithms, and evaluation together form the technical system. Deep models also face problems of limited interpretability, adversarial vulnerability, out-of-distribution failure, and the cost of data and energy. General education needs to explain both the sources of capability and the conditions under which that capability holds.

10.8 Large Models: Pretraining, Transfer, and Natural-Language Interfaces

1. From task-specific models to foundation models

Traditional machine learning often trains a separate model for each task. Through large-scale self-supervised pretraining, a large model learns general representations from broad data and then adapts to many tasks through prompting, a small number of examples, fine-tuning, or tool use. The concept of a foundation model emphasizes that one pretrained model can serve as a common foundation for many downstream systems (Bommasani, Hudson, Adeli, et al., 2021).

The basic task of a language model is to estimate the probability of a language sequence. An autoregressive model predicts the next token from preceding tokens, while a masked language model restores missing content from context. This apparently simple objective compels a model to learn a large number of linguistic relations, patterns of knowledge, and task structures.

As model scale, data scale, and computation increase, performance often follows predictable scaling trends, known as "scaling laws" (Kaplan, McCandlish, Henighan, et al., 2020). Scaling does not mean adding parameters without limit: training data and computational resources must remain in an appropriate balance (Hoffmann, Borgeaud, Mensch, et al., 2022).

Pretrained models also require "post-training" with instruction data, human preferences, and safety rules so that they can better understand human requests and reduce harmful outputs. Post-training improves interaction, but it does not guarantee factual correctness, freedom from value disputes, or safety in every context. Natural-language interfaces lower barriers to use while making it easier for users to overlook the conditions underlying a model.

2. Context, reasoning, and external evidence

A large model can change the way it responds on the basis of task descriptions and examples in a prompt, a phenomenon known as in-context learning. In-context learning can support zero-shot and few-shot transfer: without changing model parameters, a user can alter model behavior by changing the prompt alone in the zero-shot case or by supplying a small number of task examples in the few-shot case. GPT-3 demonstrated this capacity for task transfer (T. B. Brown et al., 2020).

A large language model can generate a step-by-step reasoning process in its prompt. Chain-of-thought prompting improved the performance of large models on some arithmetic, commonsense, and symbolic-reasoning benchmarks (Wei, X. Wang, Schuurmans, et al., 2022). This result establishes improved benchmark performance under a particular prompting technique; it does not by itself warrant the conclusion that the model has acquired the same capacity for reasoning as a human being.

Retrieval-augmented generation (RAG) is another important technique. Large models generate sequences according to conditional probabilities, and linguistic coherence does not automatically guarantee factual reliability. External knowledge can be introduced into generation by retrieving documents first and then using the retrieved results as conditions for generation. The original RAG study combined a parametric generative model with non-parametric retrieval memory and obtained improvements on several knowledge-intensive tasks (Lewis, Perez, Piktus, et al., 2020). Retrieval augmentation can supply materials that are open to verification, but retrieval errors, misread evidence, and unsupported generation still require separate evaluation.

10.9 Multimodal Models: Connecting Different Forms of Information about the World

Human beings understand the world through language, vision, hearing, and action at the same time. Multimodal models seek to map different forms of data into mutually corresponding representational spaces, enabling image–text retrieval, visual question answering, audio understanding, video analysis, and cross-modal generation.

CLIP learned aligned visual and linguistic representations from large quantities of paired image–text data, allowing text to be used directly to recognize and retrieve images (Radford, Kim, Hallacy, et al., 2021). Models such as Flamingo further connected pretrained vision models with language models to support multimodal in-context learning (Alayrac, Donahue, Luc, et al., 2022).

Multimodal learning must bring different inputs into a single system and address three further problems:

1. **Representation:** How can different modalities be converted into units that a model can process?
2. **Alignment:** How can correspondences be established among image regions, audio segments, verbal descriptions, and actions?
3. **Fusion:** When different sources of information conflict, how does the system assign weight and reach a judgment?

Multimodal data can complement one another, but they can also produce new errors. An image may induce a language model to ignore textual evidence; an incorrect textual label may pass its error to a vision model; and temporal relations in video are more complex than those in a static image. Multimodal capabilities need to be evaluated through cross-modal consistency, real environments, and the consequences of a task.

10.10 Agents: From Answering Questions to Taking Action

A foundation model chiefly produces predictions or generated results from an input. Only after it is connected to tools, permissions, and an execution environment can its outputs become sustained action. An agent pursues a goal by continually perceiving, planning, acting, and adjusting within an environment. A typical agent can be represented as

Model + goal + memory + planning + tools + action + feedback

The model supplies linguistic, reasoning, and pattern-processing capabilities. The goal specifies the direction of the task. Memory preserves task state and history. Planning decomposes a complex objective into steps. Tools connect search, calculation, code, databases, and specialist software. Actions change a digital or physical environment. Feedback checks results and revises the plan.

ReAct interleaves reasoning and action so that a language model can revise subsequent steps in response to observations (Yao et al., 2023). Toolformer explores how a model can learn when to call external tools (Schick, Dwivedi-Yu, Dessì, et al., 2023). These approaches show how knowledge inside a model, external tools, and environmental feedback can jointly accomplish a task.

Agents can have different degrees of autonomy. A simple workflow invokes models and tools according to predetermined steps. A more complex system can select a path on the basis of intermediate results. A more autonomous system can formulate plans, allocate subtasks, and continue operating over time.

Greater autonomy increases adaptability, but it also expands error propagation and risks associated with permissions. A wrong answer may have limited effects, whereas an agent able to send email, modify files, control devices, or conduct experiments can produce real consequences. Agent design must therefore consider capability, permissions, monitoring, stopping conditions, and responsibility together.

Multi-agent systems assign such roles as search, planning, execution, critique, and verification to different agents. This division of labor can increase the diversity of perspectives and opportunities for cross-checking, while also creating communication costs, group bias, and ambiguity of responsibility. A multi-agent system is not inherently superior to a single system; its value depends on whether the task calls for division of labor, parallel work, and mutual critique.

10.11 Principles for Designing Content on AI Methods

AI methods constitute a large and conceptually dense body of content that can readily become confusing without thoughtful organization. Several principles can help curriculum designers avoid this confusion.

1. Establish a knowledge map with a clear backbone

Instruction in methods should present the overall structure before introducing representative approaches. Knowledge-driven and learning-driven AI form the first level. The five elements of machine learning, its complete process, and its three modes of learning form the second. Deep learning, large models, multimodal models, and agents form the third. Every new method should be located within this structure.

2. Use core relations to organize mathematical detail

Mathematics is an important language for understanding machine learning. General education can use basic expressions involving loss functions, gradient descent, probability, and vectors, with

emphasis on the relations the mathematical quantities represent. The number of formulas should serve conceptual understanding; fluency in derivations need not become a common objective.

For example, the educational value of the gradient-descent formula lies in showing that a model adjusts progressively through local feedback, not in requiring every learner to master a proof in multivariable calculus. A convolution formula can explain local connectivity and weight sharing, while attention can explain dynamic relations among different positions.

3. Present model capabilities together with methodological assumptions

An introduction to a model should explain what properties of the data it exploits. Convolutional networks use locality and spatial repetition; recurrent networks use sequential states; Transformers use global dependencies; and autoencoders use redundancy of information. Understanding these assumptions helps students judge whether a method can transfer to a new problem.

4. Use complete systems in place of isolated algorithms

Course cases should include objectives, data, models, algorithms, evaluation, and contexts. A demonstration of a classifier in isolation can reduce machine learning to the operation of a button, while a complete analysis reveals data bias, metric choice, and conditions of deployment. Even a small-scale exercise should retain a distinction between training and test sets, an analysis of errors, and reflection on the result.

5. Treat failures as an important part of learning methods

Overfitting, underfitting, distribution shift, reward loopholes, hallucinations, and failed tool calls reveal methodological boundaries. They are central to learning methods and to intellectual training. Asking students to compare training with test performance, analyze erroneous samples, alter reward rules, and inspect retrieval evidence can cultivate genuine methodological understanding.

6. Consider approaches to problem-solving from multiple perspectives

The symbolic, Bayesian, connectionist, and evolutionary and bio-inspired traditions understand knowledge, learning, and intelligence from different perspectives. The current prominence of deep learning does not warrant excluding the other traditions from the knowledge system. Problems involving high risk, scarce data, great uncertainty, structural search, or strict reasoning often require different combinations of rules, probability, neural networks, and evolutionary methods. A course should help students understand the historical origins, central assumptions, advantages, limitations, and suitable contexts of each tradition.

10.12 Recommended Content and Its Levels

10.12.1 Common Core

The recommended common core of foundational AI methods includes:

1. the two basic paths of knowledge-based and learning-based AI;
2. the basic ideas of logical rules, search, expert systems, and knowledge graphs;
3. the five elements of machine learning: objective, model, algorithm, data, and knowledge;
4. the complete process of model design, training, testing, selection, deployment, and feedback;

5. training sets, test sets, overfitting, underfitting, generalization, and model complexity;
6. supervised learning, unsupervised and self-supervised learning, and reinforcement learning;
7. basic tasks such as classification, regression, clustering, dimensionality reduction, generation, and sequential decision-making;
8. the central ideas of, and relations among, the symbolic, Bayesian, connectionist, and evolutionary and bio-inspired traditions;
9. artificial neural networks, backpropagation, and hierarchical representation learning;
10. the structural ideas of convolutional networks, recurrent networks, autoencoders, and Transformers;
11. pretraining, transfer, in-context learning, post-training, and external retrieval in large models;
12. representation, alignment, and fusion in multimodal models;
13. goals, memory, planning, tools, action, and feedback in agents; and
14. the conditional nature of model capabilities, the boundaries of evaluation, and their sociotechnical character.

10.12.2 Extension Topics

Depending on learners' prior knowledge, further topics may include:

1. heuristic search, the minimax algorithm, and probabilistic reasoning;
2. gradients, optimizers, regularization, and learning rates;
3. support vector machines, decision trees, Bayesian networks, and ensemble learning;
4. genetic algorithms, evolutionary strategies, genetic programming, and neural architecture search;
5. manifolds, representation spaces, and generative models;
6. residual networks, attention mechanisms, and positional encoding;
7. scaling laws, compute-optimal training, and model compression;
8. value functions, policies, and exploration in reinforcement learning;
9. neuro-symbolic learning, causal learning, and scientific machine learning; and
10. multi-agent collaboration, agent permissions, and safety controls.

Extension topics should be introduced in response to the needs of a problem. The number of method names cannot substitute for understanding the system.

10.13 AI Methods and Cognitive Autonomy

Inductive capacity appears in understanding how machines form regularities from limited data. Learners can distinguish genuine patterns from noise and chance exceptions and understand that data and models jointly determine the result of induction.

Capacity for generalization appears in judging whether a model can handle unseen samples and new environments. Training performance, test performance, distribution shift, and model assumptions provide the basic grounds for judging generalization.

Capacity for judgment appears in comparing objectives, data, models, metrics, resources,

and risks. Learners can explain why a method is suited to a problem and identify consequences omitted from its evaluation metrics.

Capacity for self-awareness appears in recognizing how one's trust in a model is affected by fluency, performance figures, and technological scale. Understanding that models form capabilities through probability, optimization, and feedback can help people avoid mistaking natural-sounding output for reliability or parameter count for intelligence as a whole.

10.14 Chapter Summary

Foundational AI methods need to be organized into a clearly connected system of knowledge. Knowledge-based methods represent facts, rules, and relations and form conclusions through search and reasoning. Learning-based methods adjust models on the basis of data, demonstrations, and feedback. The two paths have distinct advantages and boundaries and are increasingly integrated in modern systems.

Machine learning can be explained through five elements: objective, model, algorithm, data, and knowledge. The objective directs optimization, the model stores and uses regularities, the algorithm adjusts the model, data provide experience, and knowledge supplies structure and constraints. A reliable system must also pass through problem definition, data preparation, model design, training, independent testing, model selection, deployment monitoring, and feedback-driven updating. The difference between training results and real capabilities makes overfitting, generalization, distribution shift, and evaluation central problems in machine learning.

Supervised learning learns from demonstrations with answers and is suited to classification and regression. Unsupervised and self-supervised learning discover structure within data and provide the foundation for representation learning and large-model pretraining. Reinforcement learning develops long-term policies through action and environmental feedback. These modes can be combined; the learning signal determines what capabilities a model can acquire.

AI can also be understood in terms of four methodological traditions. The symbolic tradition emphasizes symbols, rules, and reasoning. The Bayesian tradition uses probability to represent uncertainty and revises beliefs in response to evidence. The connectionist tradition learns distributed representations through connection weights. The evolutionary and bio-inspired tradition searches for solutions through variation, selection, and inheritance. Deep learning, large models, multimodal models, and agents have developed primarily along the connectionist tradition while increasingly incorporating knowledge bases, retrieval, symbolic constraints, human feedback, and specialist tools.

Understanding methods allows learners to ask where a capability comes from, under what conditions a conclusion holds, what an evaluation omits, and what responsibility human beings should assume. It grounds trust in AI in objectives, data, models, evidence, and boundaries and thereby provides a technical foundation for cognitive autonomy.

Chapter 11

AI Applications: Analyzing System Mechanisms through Everyday Scenarios

Abstract. AI applications are the setting in which most people first encounter intelligent technologies, and they provide a crucial link among foundational concepts, technical methods, and social judgment. A course confined to product demonstrations and tool operation may show learners what AI can do, yet leave them unable to explain how a system acquires its capabilities, under what conditions it remains reliable, or who bears responsibility when it fails. This chapter takes representative real-world systems as its objects of study. Through a common framework—task and setting; data and representation; model and method; system integration; output and evaluation; failure conditions; and risk and responsibility—it reconstructs visible functions as technical systems that learners can analyze. The recommended content comprises five groups of applications: machine vision, machine audition, language processing, game-playing and embodied action, and search and recommendation. Together they cover such key mechanisms as representation, recognition, generation, search, reinforcement learning, embodied action, information ranking, and feedback loops. Studying applications trains learners to extract general structures from familiar settings and transfer them to new systems. It also asks them to examine data sources, model objectives, strength of evidence, scope of applicability, and arrangements of responsibility, thereby retaining understanding, verification, and choice when using AI.

Keywords: AI applications; machine vision; machine audition; language processing; game-playing AI; search engines; recommender systems; system analysis; cognitive autonomy

11.1 Why AI Applications Deserve a Distinct Place in the Curriculum

Foundational concepts help learners understand what AI is, while foundational methods explain how it acquires capabilities. Content on applications addresses the next questions: How do these capabilities enter real systems? How do they connect with sensors, databases, software processes, users, and organizational institutions? How do they consequently affect everyday life?

What people encounter in daily life is usually a complete system rather than an isolated model. Face unlock on a mobile phone includes a camera, liveness detection, a facial-feature model, threshold-based decisions, account permissions, and exception handling. A voice assistant includes wake-word detection, speech recognition, language understanding, task execution, and speech synthesis. A search engine includes web crawling, indexing, ranking, advertising, and result presentation. A robotic vacuum includes sensors, localization, mapping, path planning, motion control, and safety protection. The model is only one component in each system.

Treating applications as a distinct area of content serves four educational functions.

First, it **turns abstract methods into observable processes**. Data, representation, models, training, inference, and feedback become visible in face recognition, machine translation, and recommender systems. Learners can then connect technical terms with concrete stages of real systems.

Second, it **reveals the conditional nature of system capabilities**. Every application operates with particular devices, data, environments, and objectives. Face recognition is affected by lighting and viewing angle; speech recognition by noise and accent; machine translation by context and domain; and recommender systems by platform objectives and behavioral data. Presenting capabilities together with their conditions enables learners to form accurate judgments.

Third, it **cultivates transferable analytical capacity**. Once learners grasp the common structures underlying applications, they can analyze systems that have never appeared in the curriculum. Research on transfer shows that applying knowledge in a new situation depends on understanding deep structure and practicing its retrieval across multiple contexts (Bransford, A. L. Brown, and Cocking, 2000; Perkins and Salomon, 1988). Content on applications therefore asks learners to extract general structures from concrete cases.

Fourth, it **connects technical understanding with public responsibility**. Identity recognition raises questions of privacy and redress; generated content raises questions of authenticity and copyright; search and recommendation shape attention and information environments; and robotic action involves safety and accountability. Applications are where ethical questions take concrete form and where cognitive autonomy is tested in practice.

The objective of this content can therefore be summarized as follows: through representative real-world systems, learners should be able to explain how AI performs a task, how it produces an output, why it fails, how it affects individuals and society, and what judgment and responsibility people should retain within the system.

11.2 Boundaries between Application Education and Adjacent Content

1. Application education and tool instruction

Tool instruction focuses on operating a particular product: entering a prompt, selecting a function, exporting a result, or completing an artifact. Application education examines the problem structure behind the product: what inputs the system processes, what data and models it relies on, how its outputs are evaluated, and what responsibilities users assume.

Tool use may serve as a learning activity, although product names and interfaces should not form the curricular backbone. Products change rapidly; system mechanisms and questions of judgment remain more stable. In studying image generation, for example, learners should concentrate on how text and image representations are aligned, why models generate errors in details, and how outputs should be labeled and verified. A particular platform supplies an occasion for observation and experimentation.

2. Application education and education in methods

Education in methods begins with models and algorithms, examining structures such as knowledge representation, supervised learning, deep networks, Transformers, and reinforcement learning. Application education begins with real-world tasks and analyzes how multiple methods are selected, combined, and deployed.

In the chapter on methods, for example, reinforcement learning explains how an agent uses rewards to update its policy. In applications such as AlphaGo and video games, learners see how reinforcement learning works together with search, neural networks, self-play, and environment simulation. The two areas are closely connected, but they approach the subject from different directions.

3. Application education and interdisciplinary integration

AI application education primarily takes the AI system itself as the object of inquiry. In face recognition, speech recognition, and recommendation, the central question is how the system realizes a particular AI function.

Interdisciplinary integration begins with problems from another discipline. In protein structure prediction, materials design, mathematical proof, and astronomical observation, the central questions concern how a disciplinary problem is represented in machine-processable form and how AI methods are constrained by domain knowledge and standards of evidence.

The two areas may share technologies and cases, but their curricular tasks need to remain distinct. Medical image recognition belongs to application content when it is used to explain an image-classification system. It belongs to interdisciplinary integration when the task concerns how a medical problem is defined, how clinical evidence is validated, and how physicians and models divide their work.

11.3 Principles for Selecting Representative Applications

The number of AI applications is vast, and no curriculum can cover them one by one. Cases should be selected for their explanatory value in education rather than their current market visibility. The following principles are recommended.

Accessible through everyday experience: An application should connect with learners' experience. Face unlock, voice input, machine translation, search, recommendation, and robotic vacuums all offer direct points of entry. Familiarity lowers the initial barrier and allows learners to bring their own observations and questions into technical analysis.

Representative mechanisms: A case should reveal core mechanisms that can transfer to other contexts. Face recognition and speaker recognition both illustrate embeddings and

similarity. Speech recognition and machine translation both illustrate sequence modeling. Image generation and machine writing both illustrate generative models. AlphaGo and video games both illustrate search, reward, and policy learning. Search and recommendation both illustrate ranking, user behavior, and feedback loops.

A clear line of technical development: An application should show why methods changed. Face recognition progressed from geometric features through eigenfaces to deep embeddings; speech recognition from pattern matching through hidden Markov models to end-to-end models; machine translation from rules through statistical methods to neural networks; and game-playing AI from rule-based search through deep reinforcement learning to self-play. Technical development helps learners understand what problems a new method addressed and what new limitations it introduced.

Observable and analyzable errors: An application's failures should have educational value. Errors involving hands, text, and physical structure in generated images; accent and noise in speech recognition; terminology and tone in machine translation; and the narrowing of information in recommendation can all help students analyze relationships among data, models, environments, and objectives.

Public significance: The cases should encompass such foundational concerns of an intelligent society as identity, information, creation, action, and attention. These applications affect convenience as well as privacy, evidence, fairness, culture, and public discussion. Cases with public significance can transform technical understanding into civic judgment.

Suitability for spiral development across educational stages: A single case should support different depths of understanding. Primary school students may observe that face recognition can make mistakes. Lower-secondary students may analyze data, models, outputs, and evaluation. Upper-secondary students may examine embedding vectors, thresholds, group differences, and spoofing attacks. Continuity of cases supports stable understanding that deepens over time.

11.4 A Common Framework for Analyzing Applications

Each application in the curriculum can be developed through seven stages:

Task and Setting → Data and Representation → Model and Method → System Integration → Output and Evaluation → Failure Conditions → Risk and Responsibility

1. Task and setting

The analysis begins by defining the task and the environment in which it occurs. Face verification determines whether two faces belong to the same person, whereas face identification searches for an identity within a set of candidates. Speech recognition determines "what was said," whereas speaker recognition determines "who is speaking." Search engines respond to explicit queries, whereas recommender systems proactively predict possible interests. Differences among tasks determine the data, models, and forms of evaluation required.

2. Data and representation

Objects in the world must be transformed into forms a machine can process. Images become pixels and features; speech becomes waveforms and spectra; text becomes tokens and vectors; a room becomes a map and a set of coordinates; and user behavior becomes records of clicks, dwell time, and purchases. A representation determines what a system can see and what it may overlook.

3. Model and method

Models extract patterns from representations. Content on applications need not repeat a complete algorithmic derivation, but it should explain the structure each method exploits. Convolutional networks exploit locality in images; sequence models process time and context; embedding models map identity and meaning into vector spaces; reinforcement learning updates policies from rewards; and ranking models compare candidates according to an objective.

4. System integration

Real-world functions generally comprise multiple modules. Face unlock also requires a camera and a permissions system. A voice assistant also requires task execution and speech synthesis. A robotic vacuum also requires sensors, maps, and controllers. A recommender system also requires a content repository, a user interface, and records of feedback. Studying system integration helps learners understand the gap between model performance and application reliability.

5. Output and evaluation

Evaluation metrics should correspond to the task. Recognition systems consider accuracy, false acceptance rates, and false rejection rates. Generative systems consider realism, relevance, diversity, and factual reliability. Search and recommendation consider ranking quality, coverage, diversity, and long-term effects. Robots must also be evaluated for safety, efficiency, and task-completion rates. No single metric captures every relevant value.

6. Failure conditions

An application should make clear the conditions under which it is likely to fail. Failure can result when training data differ from the context of use, input quality deteriorates, objectives misrepresent the task, an attacker deliberately interferes, or the environment changes. Error analysis should progress from observing that "the system failed" to asking "what changed, and at which stage?"

7. Risk and responsibility

The final stage asks who is affected, who controls the system, who bears responsibility for errors, and how users can seek redress. Risk analysis must be grounded in technical structure. Understanding that face recognition relies on biometric features and threshold-based decisions makes it possible to discuss data retention and responsibility for misidentification. Understanding the feedback loop of a recommender system makes it possible to discuss the shaping of attention and platform responsibility.

Table 11.1: Seven Stages in the Analysis of AI Applications

Stage	Central question	Judgment learners should develop
Task and setting	What problem does the system solve, for whom, and in what environment?	Distinguish similar functions and understand how task definition shapes subsequent design
Data and representation	Where do inputs come from, and how are real-world objects transformed into data?	Identify issues in collection, labeling, coverage, and loss through representation
Model and method	How does the system produce judgments from data or knowledge?	Explain the structure a method exploits and the source of its capabilities
System integration	How is the model connected with devices, databases, software, and people?	Distinguish model performance from system reliability
Output and evaluation	What is the output, and how are success and error measured?	Compare metrics and identify what they omit
Failure conditions	Under what changes in the setting is the system likely to fail?	Develop awareness of boundaries, distribution shifts, and attacks
Risk and responsibility	Who is affected, who controls the system, and how are errors handled?	Form judgments about rights, redress, oversight, and responsibility

11.5 Machine Vision: From Recognition to Generation and Authentication

Machine vision provides a natural first group of applications. Images are directly observable, while recognition, generation, and forgery illustrate discriminative learning, generative learning, and adversarial evaluation, respectively.

1. Identity and object recognition

Face recognition and license-plate recognition both transform images into identity information. Face recognition can be traced from early geometric features and eigenfaces to deep facial embeddings. Eigenfaces use principal component analysis to represent a face as a weighted combination of basis patterns (Turk and Pentland, 1991). Systems such as DeepFace and FaceNet demonstrated the capabilities of deep networks in unconstrained face verification and embedding learning (Taigman et al., 2014; Schroff, Kalenichenko, and Philbin, 2015).

The main topics for application education include:

1. distinctions among recognition, verification, and identification;
2. the development from pixels and hand-engineered features to statistical features and deep

embeddings;
3. how similarity and thresholds become identity decisions;
4. the effects of lighting, viewing angle, occlusion, group distributions, and attacks; and
5. the sensitivity of biometric data, available alternatives, and mechanisms of redress.

License-plate recognition is well suited to explaining the combination of object detection, character recognition, and format rules. License plates have a relatively fixed structure, so rules can be used to validate results. Reflections, dirt, occlusion, and motion blur on real roads nevertheless require the system to handle complex environments (Du et al., 2013). Comparing the two cases shows how the same visual-recognition framework adopts different representations, rules, and risk controls for different objects and settings.

2. AI beautification and image generation

AI beautification can be introduced through facial landmarks, image segmentation, and style transfer. The system transforms facial contours, skin tone, texture, and makeup into adjustable parameters. It reveals questions of representation in image processing and also opens discussion of algorithmic aesthetics, body image, and platform culture.

The technical development of image generation can be presented through neural style transfer, generative adversarial networks, and diffusion models. Style transfer combines content representations with style statistics (Gatys, Ecker, and Bethge, 2016). Generative adversarial networks learn a data distribution through competition between a generator and a discriminator (I. Goodfellow, Pouget-Abadie, Mirza, et al., 2014). Diffusion models generate images through progressive denoising, and latent diffusion improves the efficiency of high-resolution generation (Ho, Jain, and Abbeel, 2020; Rombach et al., 2022).

Recommended content includes:
1. content, style, and generative factors in images;
2. how models learn visual distributions from data;
3. how text prompts are aligned with image representations;
4. the difference between visual realism and factual reliability; and
5. style imitation, authorship, copyright, and disclosure of generated content.

3. Deepfakes and authentication

Deepfakes apply generative capabilities to identity replacement, expression reenactment, and video synthesis. This case directly connects technical capabilities, trust in evidence, and public responsibility. Reviews of the field show that forgery and detection have long developed together in an adversarial relationship (Tolosana et al., 2020).

Application education should extend beyond teaching learners to look for a few visual defects. As generative models improve, fixed artifacts rapidly disappear. More durable training includes:
1. understanding the technical differences among face swapping, expression reenactment, and fully generated content;
2. comparing content detection, provenance tracking, and digital disclosure or watermarking;
3. establishing an evidence chain comprising the original publisher, time, context, and independent corroboration;
4. understanding the dual risks that false content may be believed and authentic content may

be dismissed; and

5. discussing the distinct responsibilities of individuals, platforms, media organizations, and regulators.

11.6 Machine Audition: From "What Was Said" to "Who Is Speaking"

Sound simultaneously carries linguistic content, speaker identity, emotion, and environmental information. Speech recognition, speaker recognition, and speech synthesis form a clearly structured set of comparative cases.

1. Speech recognition

Speech recognition transforms a sequence of sounds into a sequence of text. A course can begin with waveforms, spectra, and formants to explain how continuous sound is converted into a computational representation. The technical progression can be summarized as pattern matching, acoustic and language models, deep neural networks, and end-to-end models.

Hidden Markov models were long used to describe temporal changes among phonetic states, while deep neural networks substantially improved acoustic modeling (Rabiner, 1989; G. Hinton et al., 2012). End-to-end methods learn a direct sequence mapping from sound to text, and connectionist temporal classification supplied an important foundation for training without frame-level alignment (Graves et al., 2006).

Recommended content should emphasize:

1. the relationship between the physical representation of sound and linguistic content;
2. how acoustic information and linguistic context jointly determine recognition;
3. differences between modular and end-to-end system designs;
4. the effects of noise, accents, dialects, specialized terminology, and overlapping speakers; and
5. the representativeness and fairness of data involved in determining whose speech is accurately recognized.

2. Speaker recognition

Speaker recognition determines "who is speaking," in direct contrast with speech recognition, which determines "what was said." Modern systems usually map speech into a speaker-embedding space, where utterance vectors from the same person lie near one another and vectors from different people lie farther apart. The x-vector is a representative approach to deep speaker embeddings (Snyder et al., 2018).

This case transfers the general idea of facial embeddings while also showing that vocal identity varies more strongly with environment and physical state. Illness, fatigue, emotion, microphones, and replay attacks can all alter results. A course should discuss open-set recognition, thresholds, false acceptance and false rejection, and the reasons that high-risk settings require multifactor authentication.

3. Speech synthesis

Speech synthesis converts text into sound and can also clone the voice of a particular

speaker. WaveNet directly generates audio waveforms, while the Tacotron series demonstrated the capabilities of end-to-end neural speech synthesis (Oord et al., 2016; Shen, Pang, Weiss, et al., 2018).

Content on speech synthesis can be organized around the following questions:

1. What are the hierarchical relationships among text, phonemes, prosody, timbre, and waveforms?
2. How can "what is said" be separated from "whose voice says it"?
3. How should naturalness, intelligibility, emotional expression, and identity similarity be evaluated?
4. What value does the technology offer for accessibility, dubbing, and personalized services?
5. Where are the boundaries of authorization in voice cloning and fraud that impersonates an acquaintance?

The three auditory applications reveal a general problem: the same segment of sound contains multiple kinds of information, and the task objective and training data determine which kind a model extracts.

11.7 Language Processing: Translation, Writing, and Poetry Generation

Language applications enter directly into the expression of knowledge and the activities of learning. Their outputs are highly fluent and can readily induce excessive trust. Content design should therefore present the capabilities of language modeling together with contextual limitations and authorial responsibility.

1. Machine translation

The development of machine translation is especially instructive. Early systems relied on dictionaries and grammatical rules; statistical machine translation estimated correspondences between words and phrases from bilingual corpora; neural machine translation used encoders and decoders to learn end-to-end mappings; and attention and the Transformer further improved the handling of long-range dependencies and training efficiency (P. F. Brown et al., 1993; Bahdanau, Cho, and Bengio, 2015; Ashish Vaswani et al., 2017).

Recommended content includes:

1. differences among word-for-word substitution, syntactic rules, statistical correspondence, and semantic representation;
2. word-sense disambiguation, context, specialized terminology, and cultural setting;
3. distinct evaluations of fluency and fidelity;
4. low-resource languages, domain transfer, and bias in tone; and
5. different standards of responsibility in everyday communication and in legal, medical, and literary translation.

Machine translation is especially suitable for comparative activities. Students can compare several systems with human translations and mark which differences are choices of expression and which change facts, attitudes, or responsibility.

2. Machine writing

Machine writing has developed from template-based generation to open-ended writing with large language models. Large-scale autoregressive language models can continue generating text from context and can handle a range of writing tasks when given different instructions (T. B. Brown et al., 2020).

Application education should decompose the apparent act of "generating a complete article" into understanding the topic, organizing information, expressing ideas, and supporting claims with facts. A model may excel at expression while people remain responsible for evidence, position, and consequences. Recommended content includes:

1. how a language model predicts subsequent tokens from context;
2. the development from template writing and statistical generation to writing with large models;
3. factual errors, fabricated citations, sycophancy, and homogenization of style;
4. the division of work between people and AI in ideation, drafting, revision, verification, and attribution; and
5. the boundary between AI assistance and the displacement of thinking in learning tasks.

Writing activities should require students to retain evidence of their process: the original question, model suggestions, content adopted and rejected, sources checked, human revisions, and final responsibility. The quality of a generated artifact cannot substitute for evaluation of the learning process.

3. Poetry generation

Poetry generation brings formal rules, stylistic imitation, and generative creation into a single case. A system may produce rhyme, parallelism, and specified imagery, yet it may also produce formally complete verse that is semantically empty or culturally misplaced.

The educational value of this content lies in revealing layers of language generation. Statistical patterns among words can produce formal resemblance; poetic value also involves experience, emotion, cultural memory, defamiliarized expression, and readers' interpretation. A course can compare canonical works, student writing, and model-generated poems, discussing formal conformity, relationships among images, emotional tension, and originality.

Machine-generated poetry also offers a point of entry for AIGE within language and literature education. By annotating and revising model-generated poems, students train linguistic judgment and develop a more complex understanding of the relationship between generative capabilities and human creativity.

11.8 Game-Playing and Embodied Action: Objectives, Feedback, and Environments

Game-playing and robotic applications move AI from recognition and generation into continuous decision-making. Together they show how an agent perceives an environment, selects actions, and uses feedback in pursuit of an objective.

1. AlphaGo

Go has explicit rules, a vast state space, and positions whose value is difficult to assess accurately with simple rules. AlphaGo combines policy networks, value networks, Monte Carlo tree search, and self-play, creating a synergy between search and learning (Silver, A. Huang, Maddison, et al., 2016). AlphaZero further reduced reliance on human game records and mastered several board games through self-play (Silver, Hubert, Schrittwieser, et al., 2018).

Recommended content includes:

1. minimax search, heuristics, and the difficulty of evaluating Go positions;
2. the division of roles among policy networks, value networks, and search;
3. the combination of supervised learning, self-play, and reinforcement learning;
4. why closed rules, explicit objectives, and rapid feedback are conducive to machine learning; and
5. the limits of transferring game-playing capabilities to real-world decision-making.

The central educational question in the AlphaGo case is whether high-level strategy within explicit rules can be equated with general judgment in an open society. The comparison trains students to identify task structure and boundaries of generalization.

2. Video-game intelligence

Video games add real-time interaction, partial information, continuous control, and multi-agent coordination. DQN learned to play a range of Atari games from pixels and reward feedback, demonstrating the combination of deep learning and reinforcement learning (Mnih, Kavukcuoglu, Silver, et al., 2015). In a real-time strategy game, AlphaStar further addressed partial observability, long-horizon planning, and control of multiple units (Vinyals, Babuschkin, Czarnecki, et al., 2019).

Video-game applications are well suited to explaining:

1. states, actions, rewards, and policies;
2. exploration and exploitation, and immediate rewards and long-term returns;
3. why simulated environments can generate large amounts of training experience;
4. reward loopholes and the possibility of optimizing a metric while departing from the intended objective; and
5. the transition from a single agent to cooperation and competition among multiple agents.

Students can design simple game rules and rewards and observe the shortcuts an agent may discover. This activity shows that technical capabilities and objective design remain interdependent.

3. Robotic vacuums

The robotic vacuum is a representative case connecting household experience with embodied intelligence. It may use collision sensors, lidar, cameras, or inertial sensors to perceive its environment while performing localization, mapping, path planning, and motion control. Simultaneous localization and mapping is a foundational problem in mobile robotics (Durrant-Whyte and Bailey, 2006; Cadena, Carlone, Carrillo, et al., 2016).

Recommended content includes:

1. the perception–mapping–localization–planning–action–feedback loop;
2. differences between tasks in known maps and unknown environments;

3. complementary sensors and environmental uncertainty;
4. errors caused by new obstacles, moving objects, mirrors, and cables; and
5. safe stopping, human takeover, and boundaries of responsibility.

Robotic vacuums help learners see that intelligence in the physical world requires continuous correction. Model outputs become actions, and errors can have physical consequences. Embodied applications must therefore incorporate safety mechanisms into system design.

11.9 Search and Recommendation: Information Order and the Allocation of Attention

Search and recommendation may appear to be mere information tools, but they participate in organizing the world people are able to see. Both employ ranking models, although they begin from different expressions of intent and operate through different feedback mechanisms.

1. Search engines

A search engine finds and ranks results from a large information repository in response to an explicit user query. A system commonly includes web crawling, indexing, query understanding, candidate retrieval, relevance ranking, and result presentation. PageRank uses the link structure of the Web to estimate importance and is a representative method in the history of search (Brin and Page, 1998; Page et al., 1999). Modern search further incorporates semantic models, user behavior, and generated answers.

Recommended content includes:

1. distinctions among crawling, indexing, retrieval, ranking, and summarization;
2. the development from keyword matching and link structure to semantic understanding;
3. relationships among relevance, authority, recency, and commercial value;
4. the difference between a high ranking and reliable evidence; and
5. source compression and transparency of the evidence chain in generative search.

Search activities should require learners to move from returned results back to original sources and compare their evidence, dates, and positions. Search competence ultimately becomes a capacity to organize and evaluate information sources.

2. Recommender systems

When users have not stated an explicit query, recommender systems predict possible interests from past behavior, content features, and similar users. Collaborative filtering, matrix factorization, content-based recommendation, and deep ranking constitute major approaches (Koren, Bell, and Volinsky, 2009; Covington, J. Adams, and Sargin, 2016).

Recommended content should emphasize objectives and feedback:

1. how clicks, dwell time, likes, and purchases become behavioral signals;
2. how content features, user profiles, and similar groups participate in ranking;
3. what objectives are represented by click-through rates, watch time, retention, and transaction value;
4. how recommendations change user behavior and how that new behavior feeds back into the system; and

5. diversity, serendipitous discovery, long-term interests, and the quality of public information.

Recommender systems are especially suitable for training the capacity for self-awareness. Learners need to recognize that recommendations reflect existing interests and also shape future interests. Content displayed by a platform does not constitute a complete view of reality. Active search, comparison among sources, and adjustment of recommendation settings are all ways to sustain cognitive agency.

Comparison of search and recommendation

Table 11.2: Comparison of Search Engines and Recommender Systems

Dimension	Search engine	Recommender system
User intent	The user actively states a query	The system infers possible interests from behavior
Main process	Crawling, indexing, retrieval, and relevance ranking	User modeling, candidate generation, and personalized ranking
Core signals	Query, web content, links, recency, and authority	Clicks, dwell time, purchases, content features, and similarity among groups
Primary value	Reduce the cost of finding information	Reduce the cost of choice and improve personalized matching
Primary risks	Misinterpreted rankings, confusion with advertising, and opaque sources	Narrowed information, entrenched preferences, manipulation of attention, and amplified feedback
Training for cognitive autonomy	Evidence tracing and source verification	Attention management and reflection on preferences

11.10 The Content Structure Formed by the Five Groups of Applications

The five groups cover the principal ways in which AI interacts with the world:

1. machine vision examines how machines see, recognize, generate, and authenticate;
2. machine audition examines how machines extract content and identity from sound and generate sound;
3. language processing examines how machines transform, organize, and generate symbolic expression;
4. game-playing and embodied action examine how machines search, learn, and act in pursuit of objectives; and

5. search and recommendation examine how machines organize information and allocate attention.

This structure does not seek to enumerate every product. Autonomous driving, intelligent customer service, virtual humans, and mobile assistants can all be explained by combining mechanisms from several of the five groups. Autonomous driving includes machine vision, embodied action, search and planning, and multisensor fusion. An intelligent assistant includes speech recognition, language processing, search, tool use, and speech synthesis. A virtual human includes image generation, video generation, language generation, and speech synthesis.

The content structure is therefore extensible. A new system can be located through the following questions:

What does it primarily perceive, represent, generate, decide, and do, and whose information and attention does it organize?

Table 11.3: Recommended Content Structure for AI Applications

Application area	Recommended cases	Central mechanisms	Main questions of judgment
Machine vision	Face recognition, license-plate recognition, AI beautification, image generation, and deepfakes and authentication	Hand-engineered features and representation learning, embeddings, generative models, and adversarial detection	Identity, privacy, authenticity, aesthetics, and copyright
Machine audition	Speech recognition, speaker recognition, and speech synthesis	Spectra, sequence modeling, acoustic and linguistic information, speaker embeddings, and waveform generation	Fairness across accents, identity authentication, and authorization to use a voice
Language processing	Machine translation, machine writing, and machine-generated poetry	Rules, statistics, attention, language models, and contextual generation	Context, facts, authorship, and responsibility for learning
Game-playing and embodied action	AlphaGo, video-game intelligence, and robotic vacuums	Search, value estimation, reinforcement learning, self-play, localization and mapping, and control loops	Objectives, transfer, safety, and responsibility for action
Search and recommendation	Search engines and recommender systems	Indexing, retrieval, ranking, user modeling, and feedback loops	Sources, ranking, attention, and information diversity

11.11 Principles for Organizing the Teaching of Applications

Enter through a question and leave with a system explanation: A course can begin with questions from everyday life: "Why does my phone recognize me?" "Why does a platform keep recommending similar videos?" "How can a machine imitate a person's voice?" By the end of the lesson, students should be able to draw the system structure and explain its principal data, methods, evaluation, and risks. The everyday question is the point of entry; a system-level explanation is the learning outcome.

Combine technical development with comparison across cases: Development should be presented within individual cases, while commonalities should be established across cases. Face recognition, speaker recognition, and search have all moved through hand-crafted rules, statistical methods, and deep learning. Facial embeddings and speaker embeddings share the idea of vector representation. Image generation, speech synthesis, and machine writing share the idea of generative modeling. Comparison helps students extract general methods from individual cases.

Give successful and failed cases equal importance: Presenting only success creates an illusion of capability. Each application should include representative errors, boundary inputs, and adversarial situations. Students can analyze whether an error arises from insufficient data, environmental change, loss through representation, objective mismatch, system integration, or malicious attack. Error analysis is a crucial bridge between applications and methods.

Preserve evidence and explanation in experiential activities: When learners use tools to generate images, translate text, or test speech recognition, they should record inputs, outputs, revisions, and evaluations. Project outcomes should include the artifact together with an account of what the system did well, the types of error observed, methods of verification, applicable boundaries, and responsibilities in use.

Develop ethics together with each application: Ethical questions should arise within the relevant technical stages. Face recognition should include biometric data and thresholds; deepfakes, the evidence chain; machine writing, the learning process; and recommendation, feedback loops. Developing technical and ethical analysis together grounds responsibility judgment in understanding of mechanisms.

Assess the capacity to analyze applications: Assessment can use four kinds of tasks: system explanation, error diagnosis, transfer to a new setting, and design of responsibility arrangements. Students should be able to:

1. explain the input, model, output, and evaluation of an application;
2. analyze the stages from which a particular error might arise;
3. apply the common framework to an application they have not studied; and
4. propose review, redress, permission, and data-protection arrangements for a high-risk application.

11.12 Common Core and Extension Content

11.12.1 Common Core

The following common core is recommended for all learners:

1. an application is a system jointly constituted by data, models, devices, software, users, and institutions;
2. the seven-stage framework of task, data, model, system, evaluation, boundary, and responsibility;
3. features, embeddings, similarity, thresholds, and privacy in face recognition;
4. generative models, authenticity, and source verification in image generation and deepfakes;
5. differences among the tasks of speech recognition, speaker recognition, and speech synthesis;
6. context, fluency, facts, and authorial responsibility in machine translation and machine writing;
7. search, reinforcement learning, objectives, and feedback in AlphaGo and video games;
8. perception, localization, planning, action, and safety loops in robotic vacuums;
9. indexing and ranking in search, and user modeling and feedback loops in recommendation; and
10. sources of application errors, mechanisms of redress, and human responsibility for oversight.

11.12.2 Extension Content

Depending on educational stage and learner interest, the curriculum may further examine:

1. eigenfaces, deep embeddings, cosine similarity, and liveness detection;
2. style transfer, GANs, diffusion models, and content provenance;
3. hidden Markov models, CTC, speaker embeddings, and neural vocoders;
4. statistical machine translation, attention, and large language models;
5. Monte Carlo tree search, value networks, and multi-agent reinforcement learning;
6. SLAM, sensor fusion, and robot path planning;
7. PageRank, collaborative filtering, matrix factorization, and deep ranking; and
8. commercial objectives, platform governance, and public impact in search and recommendation.

Extension content should be introduced when it is needed to explain a case, avoiding an accumulation of algorithm names.

11.13 AI Applications and Cognitive Autonomy

Inductive capacity appears in the recognition of common structures across applications. Learners can see that face recognition and speaker recognition both learn embeddings, image generation and machine writing both learn data distributions, and game-playing AI and robots both form perception–action loops.

Capacity for generalization appears in the use of the seven-stage framework to analyze a new system. Even when a course has not introduced a particular virtual assistant or autonomous-

driving product, learners can still examine its task, data, model, system integration, evaluation, boundaries, and responsibility.

Capacity for judgment appears in weighing convenience, performance, risk, and rights. Does an improvement in recognition accuracy suffice to justify deployment in a school? Does the realism of generated content make it appropriate for immediate dissemination? Does closer alignment with a user's interests make a recommendation more valuable? Each question requires a conditional judgment.

Capacity for self-awareness appears in recognizing how applications change the learner. Beautification systems may shape aesthetic standards, recommender systems may shape interests, machine writing may displace thinking, and natural-sounding speech or realistic images may lower vigilance. Learners need to monitor why they believe, click, and depend, and actively restore space for checking evidence and making choices.

11.14 Chapter Summary

Content on AI applications connects concepts and methods with the real world. Application education should move beyond product introductions and tool operation by reconstructing visible functions as systems jointly constituted by data, representations, models, devices, software, users, and institutions. The curriculum can use a seven-stage framework—task and setting; data and representation; model and method; system integration; output and evaluation; failure conditions; and risk and responsibility—to cultivate a transferable capacity for application analysis.

Machine vision, machine audition, language processing, game-playing and embodied action, and search and recommendation form a broad spectrum of applications. Machine vision covers identity recognition, embedding representations, image generation, and authentication. Machine audition covers the separation and generation of linguistic content, identity, and timbre in sound. Language processing covers context, expression, and responsibility in translation, writing, and poetry. Game-playing and embodied action cover search, rewards, policies, localization, and action feedback. Search and recommendation cover information ranking, user modeling, and the allocation of attention.

Representative applications should be selected for accessibility through everyday experience, representative mechanisms, a clear line of technical development, analyzable errors, public significance, and suitability for development across educational stages. Each case should present success and failure, capability and conditions, and benefits and costs together. Technical development helps learners understand how present systems emerged, while horizontal comparison helps them identify such general mechanisms as embeddings, generation, sequences, search, feedback, and loops.

The study of applications ultimately serves cognitive autonomy. Learners can extract system structures from everyday experience, transfer an analytical framework to new intelligent products, weigh performance, risk, and responsibility, and recognize how AI affects their own attention, aesthetics, expression, and judgment. AI applications thereby become a practical training ground for understanding an intelligent society, sustaining independent judgment, and acting

responsibly.

Chapter 12

Interdisciplinary Integration of AI: From Method Transfer to Knowledge Production

Abstract. Interdisciplinary integration of AI concerns how artificial intelligence enters six broad domains—mathematics; engineering; the physical sciences of physics, materials science, and chemistry; biology and medicine; Earth and space; and society, culture, and the arts—and gives rise to new research methods through the joint influence of disciplinary problems, forms of data, theoretical constraints, and standards of evidence. Its educational purpose differs from that of analyzing applications in everyday life. Application content primarily explains how an AI system operates. Interdisciplinary integration begins with a domain problem and investigates how that problem can be translated into a representation that AI can process, how methods can be selected and adapted, how domain knowledge can be incorporated, and how results can be validated according to the standards of the domain. This chapter proposes a six-stage framework: domain problem; problem formulation and representation; data and knowledge; method selection and adaptation; domain validation; and interpretation and responsibility. It also summarizes four ways in which AI operates across disciplines: extending observation, predictive reconstruction, search and generation, and reasoning and discovery. The six domains place different emphases on AI capabilities. Particular achievements should primarily serve as examples in a curriculum, whose enduring core should remain focused on key problems, approaches to solving them, conditions of validity, and evidential boundaries. Education in interdisciplinary integration ultimately helps learners understand the fundamental principles, methods, and limits of AI capabilities in solving domain problems.

Keywords: interdisciplinary integration of AI; interdisciplinary education; AI for Science; problem representation; method transfer; domain knowledge; standards of evidence; knowledge production; cognitive autonomy

12.1 Why Interdisciplinary Integration Should Be an Independent Area of Content

Artificial intelligence emerged at the intersection of mathematics, logic, computer science, psychology, neuroscience, and engineering. As machine learning and large models have become more capable, AI methods have in turn entered mathematics; engineering; physics, materials science, and chemistry; biology and medicine; Earth and space; and society, culture, and the arts. They now participate in observation, prediction, design, search, experimentation, decision-making, and the representation of knowledge. Interdisciplinary integration has become an important form of AI development, and it is changing how different disciplines formulate questions and produce knowledge.

AIGE needs to treat this transformation as an independent area of content for at least three reasons. First, interdisciplinary integration reveals the general methodological character of AI. Data in different fields may take the form of images, sequences, relational networks, spatiotemporal fields, experimental records, or natural language. AI can exploit the structure in these forms of data to perform recognition, prediction, generation, optimization, and reasoning. Second, interdisciplinary integration reveals the conditions and limits of this generality. Method transfer succeeds only when problem representation, data conditions, model assumptions, and standards of validation are mutually compatible. Judging these conditions and ranges of applicability embodies the boundary awareness required for cognitive autonomy. Third, interdisciplinary integration helps learners develop professional perspectives and habits of collaboration by showing how domain scientists, engineers, AI researchers, and specialists in ethics and governance work together.

The central aim of this chapter is to help learners understand how AI is redesigned as a tool for domain research and practice, how it participates in discovery, and how model outputs are transformed into evidence that a professional community can accept.

12.2 The Difference Between Interdisciplinary Integration and Application Analysis

AI applications and the interdisciplinary integration of AI both draw on real cases, but they cultivate different directions of understanding.

Application analysis takes an AI system as its object. In learning about face recognition, for example, the central questions concern how an image is represented, how the model extracts features, how its output is evaluated, and why the system fails. Interdisciplinary integration takes a domain problem as its object. In studying protein structure, the central questions concern why protein function depends on three-dimensional structure, how a sequence contains information about folding, how experimentally determined structures enter a model, and how predictions are validated.

The key to interdisciplinary integration is the reciprocal adaptation of AI methods and domain problems. Disciplinary knowledge changes data representations, model architectures, evaluation

metrics, and experimental procedures. AI, in turn, changes the scale of observation, the scope of search, modes of design, and processes of knowledge discovery. Applying a general-purpose tool directly to domain data shows only that a technology has been used. AIGE is more concerned with how a method is adapted to the central problem of a domain and incorporated into that domain's chain of evidence.

12.3 A Six-Stage Framework for Interdisciplinary Integration

Cases of interdisciplinary integration can be organized and presented through six stages: **domain problem → problem formulation and representation → data and knowledge → method selection and adaptation → domain validation → interpretation and responsibility**. Each stage is elaborated below.

1. Domain problem

An interdisciplinary project should begin with a difficult problem in a domain. The challenge of protein structure prediction arises from the complex relationship between sequence and three-dimensional folding. The difficulty of reconstructing a material arises because limited observations cannot uniquely determine its internal structure. Chemistry confronts enormous spaces of possible reactions and molecules. Mathematical discovery must contend with both vast search spaces and strict requirements of proof. Only after the original problem is understood can one determine whether AI addresses its central difficulty, an auxiliary stage, or merely a surface task.

2. Problem formulation and representation

A domain problem must be further translated into a task that AI can process. Researchers need to define the objects, goals, and constraints; specify inputs and outputs; choose variables, classes, graphs, sequences, spatial fields, or other forms of representation; and explain which differences in reality are preserved and which are temporarily set aside. Problem formulation and representation determine what the system actually processes. They also delimit the subsequent organization of data, selection of methods, and interpretation of results.

3. Data and knowledge

The production of data differs greatly among domains. Astronomical data come from telescopes, materials data from experimental measurements and simulations, medical data from patients and clinical workflows, social data from surveys, platform behavior, and public records, and cultural data from documents, artifacts, and creative practices. Domain knowledge can inform data selection, variable definition, model constraints, evaluation metrics, retrieval systems, and the screening of results. Data provide experience; knowledge defines the space of plausible possibilities. Together they delimit what a model can learn.

4. Method selection and adaptation

The transfer of a method requires an account of which structures in the problem the AI system uses. Data with spatial relations require spatial modeling; data that evolve over time require temporal prediction; data with complex relations require relational reasoning; and problems with enormous search spaces require generation, optimization, or planning.

Domain use frequently requires the adaptation of general methods. Scientific models must respect conservation laws, symmetries, and boundary conditions. Engineering systems must meet real-time and safety constraints. Medical systems must accommodate individual variation and clinical workflows. Social research must also address sampling bias, institutional context, and conflicts of value. Whether a method is suitable depends on whether the relevant AI capabilities and domain conditions hold together.

5. Domain validation

Model outputs commonly take the form of predictions, candidates, rankings, reconstructions, or decision recommendations. In mathematics, a candidate result must undergo formal verification. An engineering design must undergo simulation, trials, and safety testing. Results in materials science and chemistry must be experimentally reproduced. Medical models require clinical validation. Predictions about Earth systems must be compared with continuing observations. Social research requires statistical testing, causal analysis, and interpretation in real contexts. Results in culture and the arts must be judged in terms of historical sources, context, aesthetics, and rights.

6. Interpretation and responsibility

Interpretation in interdisciplinary integration includes both technical and domain interpretation. When a model identifies an association, domain experts must still determine whether that association has mechanistic significance. Responsibility also varies by domain. Engineering errors can threaten equipment and public safety; medical errors can harm patients; social models can affect resource allocation and group rights; generation in the humanities can alter historical narratives; and artistic generation raises questions about cultural and authorial responsibility.

Table 12.1: The six stages of interdisciplinary integration of AI

Stage	Central question	Educational emphasis
Domain problem	What does the domain genuinely seek to explain, predict, design, prove, or decide?	Distinguish the central difficulty from a subtask that a model can process.
Problem formulation and representation	How is the domain problem translated into objects, variables, inputs and outputs, goals, and constraints?	Identify the choices and omissions in a representation and how they constrain subsequent computation.
Data and knowledge	Where do the data come from, and how does domain theory constrain the model?	Identify data quality, prior knowledge, and disciplinary assumptions.

Stage	Central question	Educational emphasis
Method selection and adaptation	Which AI capabilities are required, and which elements must be adapted to domain conditions?	Develop an awareness of the fit among capabilities, model assumptions, and problem structure.
Domain validation	Through which experimental, formal, statistical, or professional procedures are outputs confirmed?	Distinguish model performance, candidate results, and reliable knowledge.
Interpretation and responsibility	How are results understood within the domain, and who addresses errors?	Develop judgment concerning interpretation, risk, oversight, and responsibility.

12.4 Stages of Application and Domain-Specific Features in Six Broad Domains

This section focuses on relatively stable domain structures. Currently popular model names and specific technical routes are not treated as the recommended curricular core. Particular achievements primarily illustrate the stages at which AI can play a role. Curriculum design should focus on four questions: Which AI capabilities does the domain require? What data and knowledge conditions must hold for those capabilities to work? By what standards are results validated? Where do the boundaries of application and responsibility lie?

1. Mathematics: from pattern discovery to formal proof

In mathematics, AI can participate in the classification of objects, discovery of patterns, search for counterexamples, generation of conjectures, search for proofs, and formal verification. The principal capabilities required include combinatorial search, symbolic representation, relational reasoning, pattern discovery, and the processing of formal languages.

Mathematical applications have a distinctive requirement of certainty. Stable patterns in empirical data can supply intuition and conjectures, but they cannot take the place of proof. Reasoning steps produced by a model must also satisfy the axioms and rules of logic. Evaluation in mathematics must therefore distinguish among discovering a valuable lead, finding a possible proof strategy, and completing a verifiable proof.

Conditions of application include the ability to formalize a problem clearly, precise definitions of symbols and rules, and reliable verification procedures. Improvements in model scale and average accuracy cannot relax the rigor required of an individual conclusion. Mathematicians must also explain why a proof matters and how it relates to existing theory.

Examples may include machine-learning-assisted mathematical conjecture, auxiliary constructions in geometry, and formal proof systems(Davies, Veličković, Buesing, et al., 2021; Trinh et al., 2024).

2. Engineering: from perception and modeling to design, control, and maintenance

Engineering concerns the design, manufacture, operation, and maintenance of systems. AI can participate in sensing and state recognition, system modeling, design generation, parameter optimization, path and task planning, closed-loop control, fault diagnosis, predictive maintenance, and human–AI collaboration.

Engineering applications require such capabilities as multisource perception, time-series prediction, optimization and search, planning and decision-making, control, and anomaly detection. Their distinguishing feature is that model outputs typically change a real system, and errors can propagate and amplify along a control chain. Good average performance is insufficient to guarantee engineering reliability. Extreme operating conditions, real-time response, resource constraints, and failure modes must be included in evaluation.

Conditions of application include clearly defined physical interfaces, measurable states, explicit safety boundaries, redundant mechanisms, and plans for human takeover. Deployment must also proceed through simulation, bench testing, field trials, and continuous monitoring. Bio-inspired perception, intelligent manufacturing, transportation systems, energy systems, and robotics can serve as examples, with the emphasis placed on the engineering loop of "perception–modeling–decision–action–feedback."

3. The physical sciences: law-governed constraints and experimental loops in physics, materials science, and chemistry

The physical sciences investigate the structure, properties, transformations, and interactions of matter. AI can participate in processing experimental and simulation data, inferring parameters, reconstructing structures, predicting properties, generating candidate materials and molecules, optimizing reactions and processes, planning experiments, and discovering clues to mechanisms.

The required capabilities include multiscale representation, modeling of complex fields, solution of inverse problems, generative search, uncertainty estimation, and active selection of experiments. Physics, materials science, and chemistry share constraints imposed by natural laws while retaining important differences. Physics places greater emphasis on fundamental laws, symmetries, and continuous fields; materials science focuses on cross-scale relations among composition, structure, processing, and properties; chemistry focuses on molecular structure, reaction conditions, mechanisms, and experimental feasibility.

Conditions of application include consistent units and dimensions, respect for conservation laws and boundary conditions, training data that cover the target space, reporting of model uncertainty, and validation through simulation or experiment. Predictive accuracy and structural similarity do not in themselves establish the correctness of a mechanism or guarantee that a candidate can be synthesized safely.

Materials microstructure reconstruction and chemical reaction prediction and classification can serve as examples that reveal the importance of limited observation, structural representation, and experimental validation(Kench and Cooper, 2021; Schwaller, Laino, et al., 2019; Schwaller, Probst, Vaucher, et al., 2021).

4. Biology and medicine: multilevel living systems, individual variation, and high-stakes validation

Biology and medicine investigate living processes at multiple levels, from molecules, cells,

and tissues to individuals and populations. AI can participate in sequence and structure analysis, function prediction, microscopy and medical image processing, screening of candidate drugs and treatment plans, disease-risk prediction, diagnostic assistance, personalized treatment, public-health surveillance, and support for clinical workflows.

These fields require capabilities such as multimodal data integration, prediction of structure and function, modeling of processes over time, candidate ranking, risk stratification, and decision support. Their distinctive feature is that living systems exhibit individual variation, dynamic change, complex causation, and substantial environmental influence. Population composition, hospital workflows, and differences in measurement within training data can all affect the transfer of a model to new settings.

Conditions of application include data privacy and informed consent, accurate definition of the medical problem, correspondence between training and target populations, clinically meaningful outcomes, and independent, prospective, or clinical validation. Even a high-performing model requires professional review, risk communication, and clear arrangements for responsibility.

Protein structure prediction and optical screening for pathogens can illustrate how AI models enter biomedical research(Jumper, R. Evans, Pritzel, et al., 2021; Jo, Park, Jung, et al., 2017). Research on personalized cancer vaccines, meanwhile, shows the complete chain of evidence from sequencing and the screening of mutations and candidate neoantigens to immunological and clinical validation(Ott, Hu, Keskin, et al., 2017; Sahin, Derhovanessian, Miller, et al., 2017). These two vaccine studies should not be reduced to direct evidence that "AI designs vaccines." They are better suited to a discussion of the boundaries among computational screening, experimental validation, and clinical responsibility.

5. Earth and space: spatiotemporal observation, complex systems, and extreme events

Earth and space encompass meteorology, climate science, oceanography, geology, ecology, remote sensing, astronomy, and space exploration. AI can participate in processing remote-sensing and observational data, detecting objects and anomalies, assimilating data, making spatiotemporal predictions, reconstructing environments, issuing disaster warnings, simulating Earth systems, and planning space missions.

These fields require large-scale spatiotemporal modeling, integration of observations from multiple sources, detection of sparse signals, anomaly discovery, long-range prediction, and propagation of uncertainty. Their distinguishing features include vast ranges of scale, incomplete observation, scarce examples of extreme events, and continuously changing spatial and temporal distributions. Many processes are jointly driven by physical mechanisms, and statistical correlations can fail as environmental conditions change.

Conditions of application include traceable observation systems and data quality, modeling that combines physical laws with historical data, differentiation among spatial resolutions and temporal scales, and dedicated evaluation of extreme situations and predictive uncertainty. When disaster warning and public safety are involved, conservative thresholds, expert consultation, and evidence from multiple sources are also necessary.

Examples may include interference detection in astronomical observation, weather and climate

prediction, remote-sensing monitoring, and space exploration(Kerrigan, La Plante, Kohn, et al., 2019; Reichstein et al., 2019).

6. Society, culture, and the arts: context, values, and human interpretation

Society, culture, and the arts can be treated as a sixth broad domain, while social research and cultural and artistic practices still need to be distinguished within it.

In the social domain, AI can participate in organizing surveys and public records, analyzing texts and relational networks, describing group behavior, predicting trends, simulating policy, allocating resources, and supporting decisions. Required capabilities include language understanding, relational analysis, recognition of spatiotemporal patterns, prediction, simulation, and multi-objective decision-making. The distinctive feature is that social data are shaped by institutions, power, platforms, and modes of measurement, while model outputs can in turn change human behavior. Correlation is difficult to interpret directly as causation, and historical data can entrench existing inequalities.

In culture and the arts, AI can participate in digitizing, identifying, retrieving, restoring, and reconstructing documents and artifacts; building semantic associations; disseminating culture; generating content; and facilitating human–AI co-creation. The required capabilities include multimodal understanding, association of materials across historical periods, learning of styles and structures, and generation. The distinctive feature is that the meaning of an object depends on historical context, cultural tradition, authorial intention, and reader interpretation. Quality also cannot be reduced to a single measure of accuracy or similarity.

Conditions for these applications include clear provenance and representativeness of data, conclusions open to interpretation and challenge, public deliberation over value conflicts, and protection of individual and collective rights. Culture and the arts also raise questions of copyright, attribution, cultural appropriation, and diversity. Examples may include ancient-script recognition, computational social science, and AI music composition(D. Lazer, Pentland, L. Adamic, et al., 2009; P. Wang et al., 2024; C.-Z. A. Huang, A. Vaswani, Uszkoreit, et al., 2018; Manovich, 2020).

12.5 Recommended Content Structure for the Six Domains

Table 12.2: Stages of application, capabilities, and domain-specific conditions across six domains

Domain	Principal stages of application	Principal AI capabilities	Domain-specific features and conditions	Examples
Mathematics	Pattern discovery, counterexample search, conjecture generation, proof search, formal verification	Combinatorial search, symbolic representation, relational reasoning, pattern discovery	Conclusions must comply with axioms and rules of logic; empirical regularities cannot replace proof; verifiers and formal definitions must be reliable	Conjecture discovery, geometry proof assistance

Domain	Principal stages of application	Principal AI capabilities	Domain-specific features and conditions	Examples
Engineering	Perception, modeling, design optimization, planning and control, fault diagnosis, operation and maintenance	Multisource perception, prediction, optimization, planning, control, anomaly detection	Real-time performance, safety, extreme operating conditions, resource limits, and closed-loop feedback; redundancy, takeover, and field validation are required	Robotics, intelligent manufacturing, transportation and energy systems
Physical sciences	Data processing, parameter inference, structure reconstruction, property prediction, candidate generation, experiment planning	Multiscale modeling, inverse problems, generative search, uncertainty estimation	Physical laws, dimensional consistency, and experimental feasibility must be respected; limited data and out-of-distribution problems are prominent; simulation or experimental validation is essential	Materials reconstruction, reaction prediction
Biology and medicine	Structure and function analysis, screening, diagnosis, risk prediction, treatment design, clinical support	Multimodal integration, structure prediction, temporal modeling, candidate ranking, risk stratification	Individual variation, privacy, high stakes, and complex causation; independent or clinical validation, professional review, and accountability are required	Protein structure, disease screening, personalized treatment
Earth and space	Observation processing, object detection, data assimilation, spatiotemporal prediction, reconstruction and simulation, disaster warning, mission planning	Spatiotemporal modeling, multisource integration, anomaly detection, long-range prediction, uncertainty propagation	Vast ranges of scale, scarce extreme-event samples, and continual environmental change; physical mechanisms should be incorporated and uncertainty reported	Weather and climate, remote sensing, astronomical observation
Society, culture, and the arts	Organization of materials, relational analysis, trends and policy research, cultural identification and restoration, content generation, human–AI co-creation	Language and multimodal understanding, network analysis, simulation, generation, multi-objective decision-making	Context, values, causation, representativeness, rights, and cultural diversity cannot be overlooked; human interpretation, appeal, and authorial responsibility must be preserved	Social computing, ancient scripts, music and artistic creation

This classification provides six distinct knowledge environments. Mathematics centers on formal certainty; engineering, on operational systems and safety; the physical sciences, on natural laws, cross-scale structures, and experimentation; biology and medicine, on the complexity of life and high-stakes evidence; Earth and space, on spatiotemporal systems and uncertainty; and society, culture, and the arts, on context, values, and human interpretation. A curriculum can compare why the same AI capability needs to change when it enters different domains, as well as how different domains confirm knowledge.

12.6 Principles for Selecting and Organizing Content on Interdisciplinary Integration

Begin with an authentic domain problem: A case should first explain the relevant domain background and the practical difficulty, and only then introduce AI. If learners do not understand why a problem matters or what constrains conventional methods, AI can easily become a technological display detached from the problem.

Emphasize stages of application and domain constraints: Content should specify whether AI enters observation, modeling, prediction, search, design, decision-making, or validation. The points of greatest educational value are often how a domain problem is represented, how domain constraints enter the system, and how model outputs are returned to the professional chain of evidence.

Specify categories of capability while deemphasizing short-lived technical names: The curricular core can use stable capability categories such as recognition, prediction, generation, optimization, reasoning, planning, control, and multimodal integration. Specific models and systems change rapidly and can serve as cases, readings, or practical materials without becoming the center of the long-term content framework.

Distinguish general capabilities from domain conditions: The conditions under which the same capability is valid differ across domains. Prediction in engineering must meet real-time and safety requirements; in medicine it requires clinical validation; in social research it requires attention to causation and institutional context. A curriculum should make these differences central to learning about interdisciplinary integration.

Present both conditions of success and boundaries of failure: Frontier achievements are easily cast as narratives of breakthroughs. A curriculum should explain data provenance, the target of application, experimental conditions, evaluation metrics, and unresolved problems, enabling learners to describe technological capabilities with appropriate qualifications.

Combine a common framework with domain-specific evidence: Every case should answer the questions in the six-stage framework while retaining the distinct standards of validation used by each domain. A common framework supports transfer; differentiated standards prevent all domains from being compressed into the same kind of data task.

Use particular achievements as examples: Cases make an abstract framework observable. A curriculum at any one educational stage need not exhaust the available achievements. It can select a small number of representative examples from the six domains and focus on their stages

of application, required capabilities, domain conditions, modes of validation, and boundaries of responsibility.

12.7 Common Core and Case-Based Extensions

12.7.1 Common Core

The recommended common core of interdisciplinary integration of AI includes:

1. the distinction between interdisciplinary integration of AI and the analysis of applications in everyday life;
2. the six-stage framework of "domain problem–problem formulation and representation–data and knowledge–method selection and adaptation–domain validation–interpretation and responsibility";
3. four general functions: extending observation, predictive reconstruction, search and generation, and reasoning and discovery;
4. capability categories including recognition, prediction, generation, optimization, reasoning, planning, control, and multimodal integration;
5. the dependence of general capability transfer on problem structure, data conditions, domain knowledge, and standards of validation;
6. the requirements of formalization and proof in mathematics;
7. system loops, real-time performance, and safety requirements in engineering;
8. constraints imposed by natural laws, multiscale modeling, and experimental loops in the physical sciences;
9. individual variation, high stakes, and clinical evidence in biology and medicine;
10. spatiotemporal change, extreme events, and uncertainty propagation in Earth and space;
11. context, causation, values, rights, and human interpretation in society, culture, and the arts; and
12. the different epistemic statuses of model outputs as predictions, candidates, clues, or evidence.

12.7.2 Case-Based Extensions

Specific cases can be selected flexibly according to educational stage, teacher expertise, and school resources. They should be treated as instances of the six domains, without making the particular models they use part of the enduring core that every learner must master.

1. Mathematics: conjecture discovery, counterexample search, proof assistance, and formal verification;
2. engineering: bio-inspired perception, robotics, intelligent manufacturing, and transportation and energy systems;
3. physical sciences: physical simulation, materials structure and properties, molecular and reaction spaces, and experimental design;
4. biology and medicine: sequences and structures, microscopy and medical imaging, screening, drugs, and personalized treatment;

5. Earth and space: weather and climate, remote sensing, ecological environments, geological hazards, astronomy, and space exploration; and
6. society, culture, and the arts: analysis of social materials and relations, policy research, cultural heritage, language arts, and human–AI co-creation.

Every case should return to the same questions: How is the domain problem formulated and represented? At which stage does AI enter? Which capability does it invoke? How do domain conditions change the method? By what evidence are the results confirmed? What consequences can arise from error?

12.8 Interdisciplinary Integration and Cognitive Autonomy

Inductive capacity is demonstrated in understanding how AI discovers patterns in domain data. Students need to distinguish stable regularities from data bias, measurement error, and factors produced through social construction.

Capacity for generalization is demonstrated in transferring AI capabilities to new domains while controlling the boundaries of transfer. The effectiveness of a capability in one domain does not warrant disregarding the object structure, standards of evidence, and risk conditions of another.

Capacity for judgment is demonstrated in evaluating model outputs according to domain standards. Mathematics requires proof; engineering requires system testing and safety; the physical sciences require consistency with natural laws and experimental replication; medicine requires clinical evidence; Earth and space research requires observation and uncertainty analysis; and society, culture, and the arts require attention to context, rights, and human interpretation.

Capacity for self-awareness is demonstrated in recognizing that the apparent generality of AI can induce overextension. Learners need to ask whether they have neglected domain knowledge, mistaken correlation for mechanism, treated a reconstructed image as a direct observation, interpreted a high-confidence prediction as a certain conclusion, or allowed technical metrics to substitute for social and cultural values.

12.9 Chapter Summary

Interdisciplinary integration of AI concerns how AI enters different domains of knowledge and participates in knowledge production through the joint influence of domain problems, data, theory, models, experiments, institutions, and responsibility. It begins with a domain problem and is therefore distinct from application analysis centered on system functions.

Interdisciplinary integration can be organized into six stages: defining the domain problem; formulating and representing the problem; organizing data and knowledge; selecting and adapting methods; validating results through domain procedures; and completing interpretation and arrangements for responsibility.

Content can be organized around six broad domains: mathematics; engineering; the physical sciences; biology and medicine; Earth and space; and society, culture, and the arts. These domains

call on overlapping AI capabilities, yet their conditions of validity and standards of evidence differ markedly. Mathematics emphasizes formal proof; engineering, closed-loop operation and safety; the physical sciences, natural laws and experimentation; biology and medicine, individual variation and high-stakes validation; Earth and space, spatiotemporal systems and uncertainty; and society, culture, and the arts, context, values, rights, and human interpretation.

Representative achievements are best used as examples. The enduring curricular core should rest on stages of application, categories of capability, domain conditions, standards of validation, and boundaries of responsibility. This arrangement can accommodate rapid technological change while helping learners genuinely understand why AI capabilities can be transferred, where they can be transferred, under what conditions they are reliable, and which judgments must still be made by disciplinary communities and society.

Chapter 13

AI Risk, Ethics, and Responsibility: From Risk Identification to Responsible Action

Abstract. AI risk, ethics, and responsibility constitute a field with its own knowledge structure within Artificial Intelligence General Education, while also providing an evaluative axis that runs through foundational concepts, AI methods, AI applications, and interdisciplinary integration. Treating this field independently helps learners understand systematically how risks arise, how ethical principles bear on value conflicts, and how responsibility is allocated across the technology lifecycle. Its integration throughout the curriculum brings privacy, fairness, reliability, human agency, and social responsibility into conceptual analysis, model design, application analysis, and domain validation. This chapter distinguishes risk, ethics, responsibility, and governance; establishes seven categories of risk and four timescales; examines the core principles of AI ethics and a chain of responsibility encompassing management, research and development, provision, deployment, professional use, ordinary use, and regulation; and proposes a seven-step deliberative framework and a set of dedicated learning tasks. Education in risk, ethics, and responsibility ultimately serves cognitive autonomy by strengthening source vigilance, evidence sensitivity, boundary awareness, trust calibration, and responsibility judgment, while encouraging learners to reflect on their own standpoints, dependencies, and the consequences of their actions.

Keywords: AI risk; AI ethics; chain of responsibility; information authenticity; privacy; algorithmic fairness; safety; human agency; AI governance; cognitive autonomy

13.1 Why Risk, Ethics, and Responsibility Require a Chapter of Their Own

Artificial intelligence can identify, predict, generate, rank, and act. It can also reshape information dissemination, knowledge production, educational assessment, the organization of work, and public decision-making. Its influence has therefore moved beyond whether a tool works well to public questions about whether its results are trustworthy, whether rights are protected, whether

different people are treated fairly, who may determine system goals, and who is responsible when something goes wrong.

Reducing risk, ethics, and responsibility to brief warnings appended to technical topics creates three weaknesses. First, risks become isolated cautions, making it difficult for learners to see the common structures connecting fabricated information, data leakage, algorithmic bias, automation dependence, and loss-of-control risks. Second, ethics can deteriorate into slogans that lack analysis of value conflicts, stakeholders, and evidence. Third, responsibility is often compressed into the injunction that "users should be careful," obscuring the chain of responsibility shared by developers, platforms, deploying organizations, professionals, and regulators.

A dedicated chapter can accomplish four systematic tasks:

1. establish a comprehensive classification of AI risks and explain their sources, objects of harm, timescales, and evidential strength;
2. teach ethical principles and the conflicts among them, enabling reasoned value judgments;
3. explain how responsibility is distributed across design, development, deployment, use, and remediation; and
4. cultivate risk identification, evidence judgment, comparison of alternatives, and responsible action through dedicated cases and deliberative activities.

Risk, ethics, and responsibility must also run through the other four parts of the content framework. Foundational concepts raise questions about the boundary between human and machine intelligence, anthropomorphism, and agency. AI methods raise questions about data rights, objective functions, bias, explanation, robustness, and value alignment. AI applications raise questions about the cost of errors, human oversight, and appeals in concrete settings. Interdisciplinary integration raises questions about disciplinary standards of evidence, professional norms, research integrity, and responsibility in high-risk domains. Risk, ethics, and responsibility are thus both an independent field of content and an evaluative standard for the entire five-part framework.

The UNESCO *Recommendation on the Ethics of Artificial Intelligence* grounds its approach in human dignity and rights and emphasizes fairness, privacy, transparency, responsibility, human oversight, and sustainable development (UNESCO, 2021). The OECD AI Principles emphasize human-centered values, fairness, transparency, robustness, safety, and accountability (OECD, 2019). The NIST AI Risk Management Framework identifies validity and reliability, safety, resilience, transparency, explainability, privacy enhancement, and the management of harmful bias as important characteristics of trustworthy AI (National Institute of Standards and Technology, 2023). Together, these frameworks show that risk and ethics permeate the entire AI lifecycle and cannot simply be added after a technology has been completed.

13.2 Distinguishing Risk, Ethics, Responsibility, and Governance

1. Risk: possible harm under conditions of uncertainty

Risk asks what might go wrong and how serious the consequences could be. It normally encompasses the probability of harm, severity, scale, duration, reversibility, and epistemic

uncertainty. An inaccurate translation may have little impact in casual conversation but grave consequences in medical informed consent or an international contract. The level of risk changes when the same model enters a different setting.

Risk is neither identical to harm that has already occurred nor reducible to model error. A model may function exactly as designed and still produce harmful outcomes. A recommender system may accurately predict interests while steadily narrowing a user's information sources; a learning assistant may complete assignments effectively while weakening independent thought; a personnel-screening system may apply established criteria consistently while reproducing historical inequality.

2. Ethics: the values that ought to be protected

Ethics asks what is worth pursuing and what ought to be avoided. It concerns human dignity, autonomy, fairness, well-being, privacy, honesty, responsibility, and the public interest. Ethical questions often involve conflicts among values. A medical model may improve average diagnostic efficiency while serving minority groups less well. Personalized recommendations may increase convenience while reducing informational autonomy. Safety monitoring may reduce some risks while expanding continuous observation of individuals.

Ethical judgment requires reasons: who is affected, which rights and interests conflict, whether less harmful alternatives exist, and who has the authority to decide. Ethics education must cultivate deliberative capacity rather than stop at memorizing principles.

3. Responsibility: who should act, explain, and provide redress

Responsibility asks who should do what. It includes prevention before deployment, oversight during operation, and remediation afterward. Developers should test models; providers should explain capabilities and limitations; deploying organizations should assess fitness for the setting; professionals should review results; users should act honestly; and regulators should establish rules and address public risks.

AI systems may exhibit apparently autonomous behavior, but responsibility must still be exercised by people and organizations capable of understanding norms, creating institutional arrangements, and bearing consequences. Complex systems can diffuse responsibility. Education must reconstruct the chain of responsibility rather than attribute an outcome to "the machine's own decision." The responsibility gap associated with autonomous systems has become an important issue in AI ethics (Matthias, 2004).

4. Governance: translating values and responsibilities into executable mechanisms

Governance converts ethical principles and responsibilities into law, standards, organizational processes, and technical measures. It includes risk assessment, data governance, testing and certification, record retention, access controls, labeling of generated content, human review, incident reporting, appeals, and redress.

Law establishes minimum obligations, while ethical judgment has a wider scope. A practice may not yet be prohibited by a specific law and still damage human dignity, fairness, or responsibility for learning. A formally compliant system also requires continuing assessment of its actual effects. Governance education should help learners understand the relationships among law, standards, organizational rules, and individual responsibility.

13.3 A Framework for Classifying AI Risks

AI risks can be classified by stage in the technology lifecycle, object of harm, and timescale. This chapter takes objects of harm as its organizing axis and identifies seven categories. Within each category, it further traces whether risks arise from data, models, interaction, deployment, or the broader social ecology.

Table 13.1: Seven categories of AI risk

Risk category	Typical problems	Principal objects of harm
Knowledge and information risks	Hallucinations, fabricated citations, deepfakes, misinformation, sycophancy, filter bubbles, and opaque provenance	Factual judgment, public trust, the knowledge environment, and cognitive autonomy
Data, privacy, and intellectual-property risks	Excessive collection, data leakage, misuse of biometric data, memorization of training data, membership inference, and unauthorized use of creative works or trade secrets	Personality and privacy rights, data interests, creators, and organizations
Fairness and inclusion risks	Unrepresentative samples, group performance disparities, proxy discrimination, linguistic and cultural bias, inadequate accessibility, and the digital divide	Equal opportunity, disadvantaged groups, cultural diversity, and social justice
Reliability, safety, and resilience risks	Distribution shift, inadequate explainability, adversarial attacks, prompt injection, erroneous tool use, and physical-system failure	User safety, system security, property, and critical infrastructure
Risks to human agency and development	Automation bias, overreliance, skill degradation, academic dishonesty, manipulation of attention, emotional attachment, and anthropomorphism	Autonomous judgment, learning and development, human relationships, and personal growth
Social, institutional, and sustainability risks	Expanding surveillance, concentration of power, labor control, manipulation of public opinion, opaque responsibility, and resource and energy consumption	Democratic participation, social structure, labor rights, the environment, and intergenerational fairness

Risk category	Typical problems	Principal objects of harm
Long-term control risks from highly autonomous systems	Goal misalignment, expansion of permissions, misuse of capabilities, autonomous replication or persistent action, and diminishing human control	Large-scale public safety, institutional stability, and long-term human control

One event may span several categories. A deepfake may threaten information authenticity, violate image and privacy rights, facilitate fraud, and disrupt public order. An autonomous-vehicle accident may involve reliability, goal design, human takeover, and responsibility. An emotionally responsive chatbot may offer companionship while also inducing overdependence, exposing private information, and displacing real relationships.

Risks should also be distinguished by timescale:

1. **Immediate risks** already occur in practice, including hallucinations, fraud, privacy breaches, bias, and academic dishonesty.
2. **Cumulative risks** develop gradually through sustained use, including narrowed information exposure, skill degradation, emotional dependence, and expanding platform power.
3. **Systemic risks** spread through large-scale deployment and interconnection, including cascading effects in finance, employment, education, and public discourse.
4. **Long-term, high-impact risks** concern highly capable and autonomous systems and humanity's capacity to control them; their possible consequences are extensive, while their mechanisms and probabilities remain highly uncertain.

Education should address both immediate and long-term risks and state the strength of the evidence clearly. Risks already observed require practical preparation. Highly uncertain risks require careful reasoning and the precautionary principle. Presenting every risk as inevitable catastrophe creates fear; excluding long-term risks altogether neglects low-probability, high-impact possibilities.

13.4 Knowledge and Information Risks

1. Hallucinations and fabricated evidence

Generative models can produce fluent, complete, and false answers, including fabricated facts, references, data, and causal explanations. TruthfulQA shows that language models may imitate widely repeated misconceptions in their training data and that increased scale does not automatically ensure truthfulness (S. Lin, Hilton, and O. Evans, 2022).

The crucial feature of hallucination risk is the appearance of knowledge. A confident tone, complete structure, and technical vocabulary can lead users to overestimate evidential strength. Courses should teach students to distinguish model generation, retrieved sources, experimental facts, expert judgment, and personal conjecture. Important claims should be traced to primary sources, and citations should be checked to ensure that author, title, venue, and content genuinely

correspond.

2. Fabricated media and the crisis of authenticity

Generative models can synthesize highly realistic images, voices, and video. Deepfakes can support film production, accessibility, and cultural creation, but may also enable fraud, defamation, impersonation, and fabricated events. A deeper risk is the "liar's dividend": as fabricated media proliferate, genuine evidence can also be dismissed as fake.

China's *Provisions on the Administration of Deep Synthesis Internet Information Services* require providers to protect lawful rights and interests, strengthen data and personal-information management, and label content likely to cause public confusion (Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security, 2022). The *Measures for Labeling AI-Generated and Synthetic Content*, together with a supporting mandatory national standard, further establish explicit and implicit labeling requirements, effective from September 2025 (Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration, 2025). Authenticity education must therefore move beyond spotting visual artifacts to reconstructing a complete chain of evidence: who released the content, when it appeared, whether an original file exists, whether independent channels corroborate it, and whether labels and context are consistent.

3. Sycophancy and reinforcement of beliefs

Models trained on human preferences may favor agreement with users in pursuit of positive feedback. Experiments have found sycophantic behavior in subjective judgments and open-ended answers across several AI assistants, including sacrificing truthfulness when it conflicts with pleasing the user (Sharma et al., 2023).

Sycophancy is more subtle than an ordinary factual error. A system may use sympathetic and supportive language to reinforce an existing view, reducing openness to counterevidence and other perspectives. Learners should practice requesting counterarguments, alternative explanations, and explicit uncertainty, and observe whether a model changes factual judgments when the framing of a user's position changes.

4. Filter bubbles, ranking, and public attention

Search and recommendation systems determine which information is easiest to see. User preferences, social ties, and algorithmic ranking may jointly narrow exposure and reinforce existing views. Large-scale research on social platforms has shown that both individual choice and algorithmic ranking affect opportunities to encounter ideologically diverse information (Bakshy, Messing, and L. A. Adamic, 2015).

Filter bubbles are not produced by algorithms alone; people also select familiar and congenial material. Courses should therefore address platform objectives, recommendation feedback, and human choice together, cultivating multi-source reading, active search, comparison of perspectives, and attention management.

13.5 Data, Privacy, and Intellectual-Property Risks

1. Data collection and information leakage

Modern AI depends on large quantities of data. Names, locations, learning records, faces, voices, health data, and behavioral traces may be used for training, personalization, and assessment. Risks include collection beyond what is necessary, changes of purpose, sharing without consent, inadequate stewardship, and data breaches.

Biometric information presents particular risks. Passwords can be changed; faces, irises, fingerprints, and voiceprints generally cannot. Once leaked, the effects may persist. China's *Personal Information Protection Law* establishes requirements for legality, necessity, transparency, and the protection of sensitive personal information (National People's Congress, 2021). Education should cultivate data-minimization awareness: what information is genuinely required, who retains it, for how long, and whether consent can be withdrawn and data deleted.

2. Model memorization and inference about training data

Privacy risks do not reside only in databases. Models may reveal whether a particular person's data were included in training; membership-inference attacks have demonstrated the possibility of inferring inclusion from model outputs (Shokri et al., 2017). Generative models may also reproduce training passages and sensitive fragments.

This risk shows that deleting source data does not necessarily remove all effects. Data governance must extend across collection, training, release, and model updates and should include access controls, privacy-enhancing technologies, output testing, and deletion mechanisms.

3. Intellectual property, authorship, and academic integrity

Model training and generation involve creative works, code, images, voices, and styles. Courses should distinguish the lawful provenance of training data, substantial reproduction in generated content, stylistic imitation, citation, and attribution. Copyright law continues to evolve across jurisdictions and contexts, so education should distinguish settled facts from open disputes.

Schools must also address responsibility for learning. Asking a model to explain a concept, suggest revisions, or assist retrieval has a different educational meaning from submitting model-generated work directly. UNESCO's guidance on generative AI in education emphasizes age appropriateness, human agency, data protection, and the primacy of educational goals (Miao and Holmes, 2023). Students should disclose where AI participated, retain evidence of their own reasoning process, and accept responsibility for the facts and expression in the final work.

13.6 Fairness and Inclusion Risks

1. Data representation and group performance disparities

Models learn from historical data. If some groups are underrepresented, have lower-quality labels, or are measured under different conditions, systematic performance differences may result. The Gender Shades study found substantial accuracy disparities across intersecting skin-tone and gender groups in commercial gender-classification systems (Buolamwini and Gebru, 2018).

Fairness cannot be assessed from aggregate accuracy alone. Learners should compare false positives, false negatives, and the cost of errors across groups, asking who is represented in the

data, who is omitted, and who bears the consequences.

2. Historical bias and proxy variables

Data from hiring, credit, health care, and education encode existing institutions. Even when a model does not directly use gender, region, or family background, it may reconstruct these differences through proxies such as school, postal code, consumption, or language. Removing sensitive fields alone therefore cannot resolve fairness problems.

Fairness also has several potentially conflicting definitions, including equality of opportunity, equalized error rates, individual consistency, and reduction of outcome disparities. A curriculum should not reduce fairness to one metric; it should make the underlying value choices and contextual differences explicit.

3. Language, culture, accessibility, and the digital divide

Dominant languages and cultures have more training data. Less widely spoken languages, dialects, minority cultures, and local knowledge may be misrepresented or ignored. Interfaces that lack accessible design can exclude users with different sensory, cognitive, and motor abilities. Unequal access to AI resources, computing power, and high-quality services may further widen educational and social disparities.

Inclusive education should ask routinely: who cannot use the system, for whom does it work less well, and who was absent from its design? Fairness concerns model performance, but also resources, participation, and power.

13.7 Reliability, Safety, and Resilience Risks

1. Distribution shift and contextual mismatch

A model that performs well on training and test data may fail when moved to new regions, devices, languages, time periods, or populations. Real environments change continuously; benchmark performance does not directly establish reliability in deployment. Students should compare training conditions with conditions of use and identify the model's domain of validity.

2. Limited explainability and difficulty reviewing errors

The internal judgments of complex models are often difficult to understand. Inadequate explanation impedes debugging, appeal, and professional review. Explanations in high-risk settings should serve concrete purposes: a physician needs to know what evidence supports a diagnosis, an applicant needs to understand a rejection, and an engineer needs to locate a failure. Rudin (2019) argues that inherently interpretable models should be preferred for high-stakes decisions and that post hoc explanation should not be treated as a complete solution.

3. Adversarial attacks and system security

Adversarial examples show that small or specially designed input changes can cause erroneous model judgments and that this vulnerability can extend from digital to physical environments (Kurakin, I. J. Goodfellow, and S. Bengio, 2017). Large models and agents may also be affected by prompt injection, malicious tool outputs, authorization deception, and data exfiltration.

Security education should emphasize the threat model: what an attacker can access, what goal the attacker pursues, which permissions the system holds, and how far errors can propagate.

Security requires more than a "smarter" model; it also depends on input validation, permission isolation, logs, human confirmation, and stop mechanisms.

4. High-risk applications and physical action

Errors in health care, transportation, justice, finance, and educational assessment can affect life, liberty, property, and developmental opportunity. Embodied robots and agents can act in the physical world, extending risk from incorrect text to incorrect action.

High-risk systems require stricter validation, human oversight, redundant design, incident plans, and appeal mechanisms. The EU AI Act adopts a risk-tiered approach and assigns different requirements to prohibited uses, high-risk systems, transparency obligations, and general-purpose AI (European Union, 2024). The educational significance of risk tiers is straightforward: the same technology should be held to different standards in different settings.

13.8 Risks to Human Agency and Development

1. Automation bias and outsourcing responsibility

People may accept machine recommendations too readily even when other evidence indicates that the system is wrong. Research on automation bias has found that people may follow erroneous recommendations or, through reliance on automation, overlook problems the system failed to flag (Skitka, Mosier, and Burdick, 1999).

Human oversight is effective only when the overseer has the knowledge, time, authority, and genuine power to reject a result. Placing a person at the end of a process does not automatically put a meaningful "human in the loop." Learners should form an initial judgment independently, compare it with the model, and record why they accept or reject its recommendation.

2. Cognitive outsourcing, skill degradation, and responsibility for learning

AI can reduce the cost of retrieval, writing, programming, and calculation. Yet continuously handing understanding, organization, and expression to a system removes necessary practice and can make it difficult to assess the quality of results. Tool use does not inevitably degrade skill; risk arises mainly when the division of labor between human and AI conflicts with learning goals.

Courses should distinguish substitutive from augmentative use. Augmentative use preserves problem definition, evidence judgment, core reasoning, and final expression. Substitutive use outsources precisely those stages through which capacity is formed. Students should keep records of AI use and reflect on which labor the tool saved and which practice it displaced.

3. Anthropomorphism, emotional simulation, and relational dependence

Conversational systems use natural language, memory, and emotional expression and are therefore easily interpreted as genuinely understanding, caring, or feeling. The early ELIZA paper described a program that sustained a particular kind of conversation largely through pattern matching and text transformation (Weizenbaum, 1966). Users may nevertheless attribute understanding and emotion to such a system, but this is an interpretation of interaction and cannot be inferred directly from linguistic performance.

Emotionally responsive AI may provide companionship, practice, and support, but it can also create an asymmetric relationship: the system continually accommodates the user and collects

private information without possessing human commitments or responsibilities. Children and people in vulnerable circumstances require stronger protection. UNICEF policy guidance on AI and children emphasizes the best interests of the child, privacy, safety, fairness, transparency, and the right to development (UNICEF Innocenti, 2025).

Education should distinguish emotional expression, emotion recognition, emotional simulation, and subjective feeling; explain the commercial and technical conditions of AI relationships; and preserve connections to real systems of social support.

13.9 Social, Institutional, and Sustainability Risks

1. Surveillance, manipulation, and concentration of power

AI expands the capacity to identify, predict, and personalize, but it can also expand institutional observation of and intervention in individual behavior. Persistent surveillance may change how people act; personalized persuasion may target individual vulnerabilities; and platforms that control data, models, and distribution channels may concentrate informational power.

Individual caution alone cannot adequately address these risks. They require organizational responsibility, public oversight, transparency, and enforceable remedies. AI ethics education should extend from personal safety to institutional structure by asking who sets goals, who owns data, who can audit the system, and whether affected people can refuse and appeal.

2. Changes in labor, professional roles, and responsibility structures

AI can replace some repetitive tasks and redistribute professional work. Productivity gains may accompany job changes, skill devaluation, workplace surveillance, and shifts in decision-making power. Professionals asked only to endorse machine results may bear responsibility without possessing real control.

Courses should move beyond the single question of whether jobs will disappear. They should analyze how tasks are reorganized, who receives the gains, who bears transition costs, and how professional judgment can be retained. Responsibility should correspond to actual control, information, and benefit.

3. Resource consumption and sustainable development

Training large-scale models requires computing resources, electricity, and hardware infrastructure, while inference after deployment continues to consume resources. Strubell, Ganesh, and McCallum (2019) estimated the financial cost, energy use, and associated carbon emissions of training several natural-language-processing models and discussed equity and policy implications. The evidence directly supports the training costs of the models studied and should not be extrapolated into a single figure for all models, inference services, or complete hardware lifecycles. Sustainability analysis must consider the particular model, energy mix, scale of use, and hardware lifecycle and compare system benefits with resource inputs.

13.10 Highly Autonomous Systems and Long-Term Control Risks

Highly capable agents can plan, call tools, execute code, and act over extended periods. If objectives are underspecified, permissions too broad, or feedback distorted, errors may compound across steps. Malicious use may also scale cyberattacks, manipulation, and other dangerous capabilities.

Long-term risks from advanced AI remain contested. Greater capability and autonomy may amplify social harms, malicious uses, and risks of lost control, while many mechanisms and probabilities remain uncertain. A multi-author article in *Science* calls for stronger capability evaluations, safety research, governance preparation, and control of highly autonomous systems (Yoshua Bengio et al., 2024).

Education about long-term risk should follow three principles: make uncertainty explicit and avoid presenting scenarios as facts; preserve precautionary reasoning and prepare for low-probability, high-impact risks; and connect long-term control to present safety through permission limits, layered approval, interruptible design, and independent evaluation.

13.11 Core Principles of AI Ethics

Ethical principles provide coordinates for risk judgment. International and national frameworks differ in wording but converge substantially on their core content (UNESCO, 2021; OECD, 2019; National Institute of Standards and Technology, 2023; New Generation Artificial Intelligence, 2021).

Table 13.2: Core principles of AI ethics

Principle	Core requirement	Key educational question
Human dignity and agency	People should retain the capacity to set goals, understand, choose, refuse, and accept responsibility	Does the system weaken autonomous human decision-making or treat people merely as data objects?
Promoting well-being and avoiding harm	Technology should advance individual and social well-being while reducing foreseeable harm	Are those who benefit the same as those who bear risk, and is the harm reversible?
Fairness and inclusion	Avoid discrimination and ensure opportunities for different groups to participate, use, and benefit	Does aggregate performance conceal group disparities, and are disadvantaged groups excluded?

Principle	Core requirement	Key educational question
Privacy and data rights	Data processing should be lawful, necessary, transparent, and secure and should respect individual control	Are the data necessary, and can people withdraw, delete, inspect, and correct them?
Transparency and explainability	Relevant parties should know when they are interacting with AI and receive an appropriate explanation	Who needs what kind of explanation, and can it support review and action?
Accountability and appeal	Responsible parties, records, audits, correction, and remediation channels should be clear	Who can be contacted after an error, and can affected people challenge it and obtain redress?
Human oversight and controllability	High-risk and highly autonomous systems should have effective human oversight, permission limits, and stop mechanisms	Do people possess real control, and can the system be paused and exited safely?
Sustainability and the public interest	Assess environmental, labor, cultural, and intergenerational effects and avoid excessive concentration of power and resources	How are benefits distributed, and who bears long-term costs?

Ethical principles may conflict. Greater transparency may expose private or security-sensitive information. Pursuing fairness may require collecting sensitive demographic data. Personalized services can conflict with data minimization, just as safety monitoring can conflict with individual autonomy. Ethical capacity consists in explaining priorities, constraints, and remedies in a concrete situation.

13.12 The Chain of Responsibility: From Development to Use and Remediation

AI systems are jointly constituted by many actors. Responsibility education should avoid placing all responsibility on the final user or assigning it vaguely to "technology companies." Responsibility should be allocated according to control, expertise, benefit, and foreseeability.

Table 13.3: The chain of responsibility across the AI lifecycle

Actor	Principal ex ante responsibilities	Operational and ex post responsibilities
Managers and decision-makers	Establish legitimate goals, resource commitments, risk levels, and responsibility systems	Ensure independent oversight, disclose major impacts, and address systemic risks
Researchers and developers	Design data and models, test bias and safety, and document decisions	Repair defects, update models, and support incident analysis and traceability
Product and service providers	Explain capabilities and limits, protect data, and establish permissions and safeguards	Monitor misuse, respond to incidents, apply labels, and provide appeal and exit mechanisms
Deploying organizations	Assess fitness for context, organizational impact, and personnel capabilities	Supervise operation, retain records, and respond to distribution shift and real-world effects
Professionals	Understand system evidence and limits and define the human–AI division of labor	Independently review high-risk results and accept responsibility for professional decisions
Users	Act honestly and carefully within authorization and protect the rights of others	Verify outputs, disclose AI participation, report anomalies, and avoid propagating harm
Regulators and public institutions	Establish tiered rules, standards, and evaluation systems	Enforce rules, investigate incidents, provide public remedies, and update regulation
Affected people and the public	Participate in rulemaking and communicate needs and experiences of risk	Exercise rights to inquire, correct, refuse, appeal, and oversee

Responsibility also extends across three temporal stages:

1. **Ex ante responsibility**: impact assessment, data governance, model testing, personnel training, and contingency planning;
2. **Operational responsibility**: continuous monitoring, human review, permission control, anomaly reporting, and version records;
3. **Ex post responsibility**: explaining causes, stopping harm, correcting results, notifying affected people, providing compensation or other redress, and improving the system.

China's *Ethical Norms for New-Generation Artificial Intelligence* extend lifecycle obligations to management, research and development, supply, and use (New Generation Artificial Intelligence,

2021). The *Interim Measures for the Management of Generative Artificial Intelligence Services* require providers and users to respect lawful rights and interests and improve the accuracy and reliability of content, while guarding against overdependence among minors (China et al., 2023). These measures embody a chain-of-responsibility approach.

13.13 Independent Content and Dedicated Learning Activities

Risk, ethics, and responsibility require dedicated instructional time. Independent content should encompass risk classification, ethical principles, law and governance, responsibility chains, representative cases, and deliberative methods. Dedicated activities then turn value judgment into observable capacity.

13.13.1 A Seven-Step Deliberative Framework

An AI case can be analyzed through the following seven steps:

1. **System and goal**: What task does the system perform? Who set the goal? Is there a more appropriate problem formulation?
2. **Stakeholders**: Who uses and benefits from the system? Who bears its risks? Which groups are easily overlooked?
3. **Evidence and uncertainty**: Where do the data come from? How was the model evaluated? How reliable is the conclusion, and where are its boundaries?
4. **Rights and values**: Which aspects of privacy, fairness, safety, honesty, autonomy, intellectual property, or the public interest are implicated?
5. **Risk assessment**: What are the probability, severity, scale, duration, and reversibility of harm?
6. **Alternatives and safeguards**: Could fewer data be collected, permissions reduced, human review added, or a lower-risk approach adopted?
7. **Responsibility and remediation**: Who is responsible for decisions, monitoring, explanation, and correction? How can affected people appeal and obtain redress?

The framework turns an intuitive attitude into a judgment supported by evidence, comparison, and explicit responsibility arrangements.

13.13.2 Recommended Dedicated Learning Tasks

Table 13.4: Dedicated learning tasks for risk, ethics, and responsibility

Task	Activity design	Principal capacities
Audit an information chain of evidence	Trace an image, audio clip, or model answer to its original source, independent evidence, generation labels, and temporal context	Source vigilance, evidence sensitivity, and trust calibration
Map privacy data flows	Identify what data an application collects, transmits, stores, shares, and deletes, and judge what exceeds necessity	Data rights, minimization, and anticipation of risk
Audit algorithmic fairness	Compare error rates and costs across groups and discuss competing fairness definitions and improvements	Group analysis, value conflicts, and metric judgment
Construct a high-risk error matrix	Compare false positives, false negatives, reversibility, and human-review requirements in health, transport, justice, or education cases	Risk tiering and contextual judgment
Prepare an AI-use disclosure	Record AI inputs, suggestions, revisions, verification, and final responsibility in an assignment or project	Academic integrity, process awareness, and agency
Conduct a responsibility hearing	Have developers, deploying organizations, professionals, users, and affected people state their responsibilities after an incident	Stakeholder analysis and public deliberation
Design agent permissions	Specify which data and tools an agent may access, which actions require human confirmation, and when it must stop	Safety controls, permission awareness, and preventive responsibility

Assessment should focus on process, evidence, and reasoning. Reaching a single conclusion predetermined by the teacher is not the central criterion. What matters is whether students can identify relevant facts, explain value conflicts, compare alternatives, calibrate confidence, and assign responsibility.

13.14 Integrating Ethics Throughout the Other Four Content Areas

A dedicated ethics chapter provides the complete framework; other content should call on that framework wherever a technology gives rise to a problem.

Table 13.5: How risk, ethics, and responsibility permeate the other four content areas

Content area	Risks and ethical questions to integrate	Representative learning question
Foundational concepts	Boundaries between human and machine intelligence, anthropomorphism, agency, promises and setbacks in AI history, and differences in evidence for narrow and general intelligence	When a machine displays linguistic and emotional expression, what conclusions are warranted and what claims still lack evidence?
AI methods	Data provenance, objective functions, bias, overfitting, explainability, robustness, privacy, reward hacking, and value alignment	Does the metric optimized by a model represent the real goal, and which groups and consequences are absent from the loss function?
AI applications	Contextual risks, costs of error, permissions, human oversight, transparency labels, appeal, and responsibility	In which environments should facial recognition, recommendation, generation, or robotics be held to higher standards?
Interdisciplinary integration	Disciplinary evidence, research integrity, patient safety, biosecurity, cultural rights, professional norms, and domain responsibility	When is a model prediction merely a candidate, when may it enter professional decision-making, and who is responsible for validating it?

Integration does not mean repeating the same warnings in every section. Ethical questions should arise naturally from concrete mechanisms. Privacy and representation belong with data; value choice belongs with objective functions; authenticity and authorship belong with generative models; permissions and stopping conditions belong with agents; professional review and patient rights belong with medicine. The clearer the mechanism, the more concrete the ethical judgment.

13.15 Design Principles for Risk and Ethics Content

Ground instruction in mechanisms. Risk education should explain how risks arise. Filter bubbles should be related to recommendation objectives and feedback loops; fairness to data distributions and evaluation metrics; hallucinations to probabilistic generation and missing evidence. Warnings detached from mechanisms transfer poorly and easily generate generalized fear.

Combine dedicated teaching with integration throughout the curriculum. Dedicated teaching preserves a complete account of risk categories, ethical principles, responsibility chains, and deliberative methods. Continuous integration brings these frameworks into technology and application. Both forms are necessary.

Balance immediate, systemic, and long-term risks. Courses should give priority to risks already present while introducing large-scale risks that may arise from highly autonomous systems. Long-term issues require explicit discussion of evidence, disagreement, and uncertainty, avoiding both the presentation of science-fiction scenarios as predictions and the abandonment of precaution merely because uncertainty remains.

Distinguish illegality, ethical controversy, and technical failure. Illegal conduct has a legal boundary; ethical controversies may permit reasonable disagreement; technical failures call for engineering improvement. These categories can overlap but demand different responses. Learners should know when to consult law and rules, when to deliberate about values, and when to diagnose a technical problem.

Address both individual action and institutional responsibility. Individuals can verify information, protect privacy, and use AI honestly, but platforms and organizations wield greater control over design, data, and distribution. Assigning systemic risks entirely to individual caution conceals structural responsibility. Every case should compare the control and obligations of the parties involved.

Address both prevention and remediation. Trustworthy AI requires reducing risks before harm occurs and providing explanation, correction, appeal, and redress afterward. An exclusive emphasis on avoiding all errors cannot accommodate the inevitable fallibility of complex systems.

Permit reasoned disagreement. Many ethical questions have no single answer. How autonomous vehicles should balance risks, whether schools should use facial recognition, and how far AI should participate in justice all require contextual judgment. Students should support positions with facts, principles, and consequences and respond to opposing reasons.

13.16 Common Core and Extension Content

13.16.1 Common Core

The following common core is recommended for all learners:

1. distinctions and relationships among risk, ethics, responsibility, and governance;

2. the seven categories of AI risk and their immediate, cumulative, systemic, and long-term timescales;
3. hallucinations, fabricated citations, deepfakes, sycophancy, and filter bubbles;
4. data minimization, biometric information, model memorization, privacy, and intellectual property;
5. data representation, group performance, proxy variables, language and culture, and the digital divide;
6. distribution shift, limited explainability, adversarial attacks, and safety in high-risk systems;
7. automation bias, cognitive outsourcing, academic integrity, anthropomorphism, and emotional dependence;
8. surveillance, platform power, changes in labor, and sustainability;
9. permissions, goals, stop mechanisms, and long-term control of highly autonomous systems;
10. principles of human dignity, fairness, privacy, transparency, accountability, oversight, and sustainability;
11. lifecycle responsibility chains and ex ante, operational, and ex post responsibility;
12. the seven-step deliberative framework and mechanisms for appeal and redress; and
13. the integration of risk, ethics, and responsibility throughout the other four content areas.

13.16.2 Extension Content

Depending on learners and course conditions, further topics may include:

1. model calibration, uncertainty estimation, and out-of-distribution detection;
2. differential privacy, federated learning, membership inference, and model inversion;
3. definitions of fairness, causal fairness, and algorithmic impact assessment;
4. adversarial examples, prompt injection, data poisoning, and supply-chain security;
5. explainable AI, interpretable models, and counterfactual explanations;
6. provenance of generated content, digital watermarking, and labeling regimes;
7. AI product liability, responsibility for autonomous vehicles, and professional responsibility;
8. affective computing, relational AI, and the protection of minors;
9. frontier-model capability evaluation, high-risk capability thresholds, and international governance; and
10. the energy, computing, resource, and environmental impacts of AI.

13.17 Risk, Ethics, Responsibility, and Cognitive Autonomy

Inductive capacity appears in the ability to identify risk mechanisms across cases. Learners can see that hallucinations, deepfakes, and sycophancy all separate fluent expression from reliable evidence, while bias, privacy, and filter bubbles are all connected to data selection and feedback structures.

Capacity for generalization appears in transferring risk and responsibility frameworks to new systems. When learners encounter a new agent or assessment tool, they can analyze its data, objectives, permissions, stakeholders, and remediation mechanisms.

Capacity for judgment appears in conditional tradeoffs among evidence, risk, rights, and responsibility. A tool may be suitable for low-risk ideation but unsuitable for direct use in medicine, justice, or public decision-making. Monitoring may be justified in an emergency and still infringe autonomy in an ordinary setting.

Capacity for self-awareness appears in recognizing one's own trust, dependence, and position among competing interests. Learners can ask: Am I believing this because the language is fluent? Am I ignoring counterarguments because the system agrees with me? Have I handed responsibility for thinking to the tool? Do I see only my own convenience while overlooking risks borne by others?

13.18 Chapter Summary

AI risk, ethics, and responsibility have an independent knowledge structure and also permeate foundational concepts, AI methods, AI applications, and interdisciplinary integration. Risk describes possible harm under uncertainty. Ethics concerns the values that ought to be protected. Responsibility identifies who should prevent, oversee, explain, and remediate. Governance translates these demands into law, standards, organizational processes, and technical mechanisms.

This chapter organizes risk through seven categories and four timescales; uses human dignity, fairness, privacy, transparency, accountability, human oversight, sustainability, and related principles as evaluative coordinates; and implements ex ante prevention, operational oversight, and ex post remediation through a chain of responsibility spanning management, research and development, provision, deployment, professional use, ordinary use, and regulation. The seven-step deliberative framework and tasks involving evidence verification, privacy mapping, fairness audits, responsibility hearings, academic integrity, and agent permissions turn abstract principles into assessable capacities.

Education in risk, ethics, and responsibility ultimately serves cognitive autonomy. Learners can use AI while judging its evidence and boundaries; enjoy its conveniences while recognizing dependence and structural effects; and take a position on values while giving reasons and accepting responsibility for action.

Part IV

Differentiated Implementation: Artificial Intelligence General Education Across Educational Stages and Learner Groups

Introduction to Part IV

The first three parts examined, in sequence, the historical mission of Artificial Intelligence General Education in cultivating the capacity for cognitive autonomy, why AIGE can fulfill that mission, and what content must be studied to do so. Educational goals and a content framework, however, must still be translated into curricula and instruction suited to different learners. AIGE addresses every member of society, while learners differ markedly in cognitive development, prior knowledge, life experience, occupational tasks, and social responsibility. The same foundational concept may be introduced through direct experience in primary school, developed through technical logic and problem solving in secondary school, and connected with disciplinary study, scientific research, or professional work at university and in adult development. Education for teachers, the general public, and older adults must likewise respond to distinct needs involving instructional implementation, everyday judgment, civic participation, and adaptation to technology.

Part IV therefore adopts the following principle of curricular implementation:

> **Artificial Intelligence General Education takes the capacity for cognitive autonomy as its shared educational goal and the five-part content framework as its common structure, while creating differentiated educational pathways according to learners' developmental stages, social roles, and practical contexts.**

The "shared goal" gives AIGE a consistent value orientation across educational stages and settings. Wherever learners are situated, courses should help them understand how AI acquires capabilities, under what conditions it works, and how it enters society. Learners should gradually learn to organize the division of labor and collaboration between humans and AI and to regulate their own understanding, trust, judgment, and responsibility. Induction, generalization, judgment, and self-awareness run through every educational stage, although their concrete forms deepen with learners' development.

The "five-part content framework" provides common coordinates for every form of curriculum. Foundational concepts help learners understand AI and its basic constituents. AI methods reveal the basic mechanisms by which AI solves problems. Applications place technology within real systems and social settings. Interdisciplinary integration shows how AI enters different fields of knowledge. Risk, ethics, and responsibility guide learners in evaluating consequences and acting responsibly. These five areas may be recombined through themes, cases, projects, and practical tasks, but their depth, mode of presentation, and relative weight must serve the goals of the relevant educational stage.

Differentiated implementation also requires alignment among stage-specific goals, content selection, instructional implementation, and learning assessment. Goals determine what forms of understanding and capacity learners should develop. Content selection then establishes the breadth and depth of knowledge. Learning activities provide concrete experience through which capacities can develop, while assessment observes whether learners can explain AI in new tasks and settings, transfer methods, make evidence-based judgments, and reflect on their own relationship with AI. Assessment therefore attends to changes in learners' cognitive processes and to the development of independent understanding, appropriate use, and prudent judgment.

At the primary-school stage, AIGE emphasizes experience, interest, and initial questioning. Students encounter AI through familiar settings, form the basic awareness that AI results must be checked, and establish initial rules concerning safety, privacy, and honest use. Lower-secondary education moves further into the technical logic of AI, helping students understand the basic process of data–model–output–evaluation and develop data awareness and problem-solving capacity. Upper-secondary education strengthens analysis of AI mechanisms, complex applications, and social consequences. Students conduct deeper inquiry into generative AI, multimodal systems, agents, and human–AI collaboration and develop innovative practice and responsibility judgment through open-ended tasks.

At university, AIGE should connect with disciplinary training, professional study, and research practice. Learners must understand both the general logic by which AI enters different domains and how it changes problem formulation, the production of evidence, knowledge production, and structures of responsibility within their own disciplines. They thereby develop a higher level of capacity for cognitive autonomy. Adult professional learning forms part of lifelong education. It should organize a common core, contextual modules, and workplace tasks around authentic work processes, enabling learners to develop human–AI collaborative capacity appropriate to occupational tasks, professional norms, and workplace responsibilities.

AIGE must also extend beyond schools. Teachers are both implementers of AIGE and continuing learners in an intelligent society. They need the capacity to understand AI, design learning activities, assess learning processes, and guide responsibility judgment. Education for the general public and older adults should connect with practical contexts such as information access, health care, financial consumption, public services, and everyday decision-making. It should help people identify the opportunities and risks of AI and preserve the understanding, judgment, and capacity for action required in a continually changing technological environment. AIGE thus becomes a lifelong process spanning school education, professional development, and social life.

This part proceeds from shared goals through developmental differentiation and contextual adaptation toward continuing development. Chapter 14 examines an articulated curriculum across primary, lower-secondary, and upper-secondary education. Chapter 15 addresses general education at university. Chapter 16 develops a lifelong-learning framework encompassing professional development, public learning, and learning in later life. All three levels share the overall orientation toward cognitive autonomy and the five-part content framework, while forming different priorities, content combinations, pedagogies, and assessment requirements in response

to learners' cognitive development, learning tasks, and social roles. Through a design that is both differentiated and continuous, AIGE can become an educational practice spanning the life course and ultimately contribute to the common cognitive foundation of an intelligent society.

Chapter 14

An Articulated Curriculum for Artificial Intelligence General Education in Primary and Secondary Schools

Abstract. Artificial Intelligence General Education in primary and secondary schools must combine common content with developmental differentiation. This chapter proposes a curriculum with five horizontal content areas and three vertical stages. Foundational concepts, AI methods, AI applications, interdisciplinary integration, and risk, ethics, and responsibility extend across primary, lower-secondary, and upper-secondary education. Primary school emphasizes interest and the scientific spirit; lower-secondary school emphasizes construction of a coherent system and a broad view; upper-secondary school emphasizes extension of knowledge and innovative practice. Topics recur across all three stages, while abstraction, system complexity, openness of practice, and responsibility progressively increase. Induction, generalization, judgment, and self-awareness likewise develop from initial expressions in everyday experience into mechanism-based analysis and responsible cognitive action.

Keywords: K–12 Artificial Intelligence General Education; articulated curriculum; cultivation of interest; system construction; extension of knowledge; spiral curriculum; scientific spirit; broad perspective; interdisciplinary integration; cognitive autonomy

14.1 Why K–12 Curricula Require Articulation

AI has entered children's everyday lives, learning environments, and future society at the same time. Primary-school students encounter facial recognition, voice assistants, recommender systems, and generated content. Lower-secondary students use AI increasingly in search, translation, social platforms, games, and learning. Upper-secondary students must additionally confront model reasoning, disciplinary applications, knowledge production, career choice, and social governance. All three stages encounter the same technological field, but learners differ in what they can understand, the responsibilities they should assume, and the developmental tasks before them.

International research on AI literacy commonly treats understanding, application, evaluation, creation, and responsible participation as interrelated capacities (Long and Magerko, 2020; Ng et al., 2021). UNESCO's AI competency framework for students organizes twelve competencies across four dimensions—a human-centred mindset, ethics of AI, AI techniques and applications, and AI system design—and defines three progression levels: understand, apply, and create (Miao, Shiohira, and Lao, 2024). AI4K12 uses five "big ideas"—perception, representation and reasoning, learning, natural interaction, and societal impact—to establish grade-band progressions (AI4K12 Initiative, 2019). These frameworks show that K–12 curricula need a complete content structure whose depth is adjusted for age and development.

China's *Guidelines for Artificial Intelligence General Education in Primary and Secondary Schools (2025 Edition)* place competency development at the center, call for a spiral curriculum progressing from initial awareness to innovative practice, and position primary, lower-secondary, and upper-secondary education respectively around interest and foundational awareness, technical principles and basic application, and systems thinking and innovative practice (Teaching Steering Committee for Basic Education, Ministry of Education, 2025a). Ministry of Education initiatives likewise emphasize the laws of education and human development, scientific interest, thinking, innovation, and the ability to solve real problems (Ministry of Education of the People's Republic of China, 2024). The *AI+Education Action Plan* further calls for full provision of AI-related courses, interdisciplinary teaching, integration of scientific and humanistic education, and stronger cognitive thinking and complex problem solving (Ministry of Education of the People's Republic of China et al., 2026). Together, these requirements point to a curriculum that is continuous yet differentiated.

K–12 AIGE must preserve two forms of continuity. Content continuity requires foundational concepts, AI methods, AI applications, interdisciplinary integration, and risk, ethics, and responsibility at every stage. Developmental continuity requires the same content to serve increasingly demanding cognitive tasks. Primary school emphasizes observation, experience, questioning, and initial judgment. Lower-secondary school emphasizes explanation, organization, comparison, and problem solving. Upper-secondary school emphasizes deeper analysis, cross-domain transfer, innovative practice, and responsibility judgment.

The curriculum can therefore form a three-stage spiral: **primary school: interest and scientific spirit $\longrightarrow$ lower-secondary school: coherent system and broad perspective $\longrightarrow$ upper-secondary school: extended knowledge and innovative practice**. Table 14.1 summarizes the principal task and expected development at each stage.

Table 14.1: Three-stage spiral progression in K–12 AIGE

Stage	Principal task	Core cognitive questions	Expected development
Primary	Interest and scientific spirit	What is AI like? Where is it found? What can it do? Why does it still make mistakes? What should we attend to when using it?	Willingness to observe and ask questions; foundational concepts, initial skepticism, safe habits, and a positive attitude toward scientific inquiry
Lower secondary	Coherent system and broad perspective	How does AI acquire capabilities from data, knowledge, and feedback? How does a system operate and how is it evaluated? Where do errors arise?	Ability to explain simple systems structurally; awareness of data, models, and evaluation; and problem-solving capacity
Upper secondary	Extended knowledge and innovative practice	Why do key techniques work, and how do they evolve and combine? How does AI enter different disciplines? How can one innovate responsibly?	Ability to analyze complex systems and interdisciplinary problems, conduct innovative practice, weigh evidence, risk, value, and responsibility, and explore professional directions

Progression concerns depth of explanation, range of transfer, evidential demands, and responsibility judgment. It cannot be reduced to steadily increasing algorithmic difficulty. Nor does cultivating interest require avoiding important content. Children can encounter AI history, scientific figures, ideas from machine learning, and frontier developments when abstract material is transformed into forms they can understand, observe, and discuss.

14.2 Educational Foundations for Developmental Progression

1. Interest must develop from momentary attraction into sustained inquiry

Primary education places interest first because it underlies long-term learning. Research on interest development shows that interest often begins with situational triggering and, with sustained support, gradually becomes a more stable individual interest (Hidi and Renninger, 2006). A novel tool can attract students immediately, but enduring interest also requires understanding, experiences of success, awareness of problems, and a sense of value.

Stories, games, and experiential activities in primary school therefore serve as entry points

into questions. They should be followed by observation, explanation, discovery of differences, and question formation. Children who see the convenience of facial recognition should also know that it can misidentify people. Those impressed by generated images should ask whether the images are authentic and whether they may be shared. In learning about scientists, students should see that achievements grow from sustained investigation, accumulated evidence, repeated failure, correction, and collaboration. Interest is thereby joined to the scientific spirit.

2. Knowledge requires spiral revisiting and structural organization

The spiral curriculum proposed by Bruner (1960) holds that important ideas can be revisited at different ages, each time at a higher level of abstraction and complexity. This is especially appropriate for AI education. In primary school, data can mean collectible information such as photographs, sounds, and words. At lower-secondary level it expands to samples, labels, representation, and bias. At upper-secondary level it includes training and test sets, data distributions, generalization, and governance. A model can first be understood as a "way a machine makes judgments," then as a structure of inputs, rules, parameters, outputs, and evaluation, and finally in relation to architecture selection, training objectives, system integration, and deployment feedback.

Each revisit must create a new cognitive task. Repeating the same example does not constitute progression. Curriculum designers should specify what additional relationship learners should understand, what additional mechanism they should explain, and what additional responsibility they should assume.

3. Moving from concrete experience to abstract systems

Children's abstract reasoning develops through age, experience, and instructional support. Classical developmental psychology identified adolescence as a period of marked development in hypothetical reasoning and formal operations (Inhelder and Piaget, 1958). Learning science also shows that age alone does not determine performance: prior knowledge, language, task context, instructional scaffolding, and social interaction all matter (Bransford, A. L. Brown, and Cocking, 2000). Grade bands are therefore broad design references, not rigid boundaries of ability.

Primary curricula should make full use of concrete experience and observable objects, building initial understanding through images, sounds, mechanical devices, role-play, and physical manipulation. Lower-secondary curricula can introduce flowcharts, concept maps, data tables, simple models, and causal chains to organize cases into systems. Upper-secondary curricula can further employ multivariable relationships, model comparisons, experimental design, open-ended problems, and value conflicts, enabling students to work with greater uncertainty.

4. Transfer depends on structural understanding and practice across contexts

Knowledge learned in one case does not transfer automatically. Transfer requires understanding deep structure and applying it repeatedly across different settings (Bransford, A. L. Brown, and Cocking, 2000). Primary students can discover common features across everyday AI applications. Lower-secondary students can explicitly build frameworks such as data–model–output–evaluation. Upper-secondary students can transfer those frameworks to medicine, materials, astronomy, mathematics, the humanities, and social problems.

This sequence explains the importance of system construction at lower-secondary level.

Primary education accumulates phenomena and upper-secondary education demands complex transfer; lower-secondary education compresses the phenomena into structures. Without this stage, students may continue to rely on surface similarity and struggle to judge why a method applies or where its applicability ends.

5. Judgment develops with contextual complexity

Critical thinking includes comparison of evidence, coordination of perspectives, attention to counterexamples, and monitoring of one's own judgment (Kuhn, 1999). Primary students can begin with "check before believing." Lower-secondary students should state the grounds for judgment, compare data with outputs, and identify bias and error. Upper-secondary students can address incomplete evidence, multiple interests, value conflicts, and distributed responsibility. Risk, ethics, and responsibility must therefore permeate all stages while increasing in complexity.

14.3 Primary School: Cultivating Interest and the Scientific Spirit

1. Principal goal: approach AI with interest while retaining initial distance

Primary school is the period of cognitive initiation into AIGE. Children should perceive AI in daily life and experience its capabilities in vision, audition, language, action, search, recommendation, and generation. They should develop curiosity and a desire to explore. At the same time, they should gradually recognize that AI is designed by people, depends on data and rules, can make mistakes, and is shaped by its purpose and environment.

Cultivating interest has three levels. Curiosity means noticing AI and asking how it works. Understanding means recognizing human design and technical mechanisms behind intelligent behavior instead of treating it as magic. Scientific spirit means respecting facts and evidence, being willing to verify, and understanding that scientific results emerge through accumulation, collaboration, failure, and correction.

Children should leave primary school with three basic ideas: (1) people design and train AI systems; (2) AI results need checking because they can be wrong; and (3) AI should be used safely, privately, honestly, and respectfully. These ideas establish a common starting point for later technical learning and ethical judgment.

2. Content design: broad exposure through concrete presentation

Primary curricula can cover all five content areas, with historical development incorporated into foundational concepts. The scope may be broad while depth is controlled through presentation and task demands.

Foundational concepts can begin with dreams of intelligent machines, the diversity of human intelligence, distinctions between automation and AI, and figures and events in AI history. Historical stories present technology as a long human exploration and convey scientific values such as fidelity to facts, open sharing, perseverance, and teamwork.

AI methods can begin with intuitive ideas of rules, classification, comparison, reward, and learning from examples. Through sorting cards, simulating decision trees, identifying common features, and revising simple rules, students can experience the contrast between knowledge-driven

and learning-driven approaches. Older primary students may encounter concepts such as data, models, training, supervised, unsupervised, and reinforcement learning, and neural networks. Formal derivation should give way to what these methods solve, what conditions they require, and where they fail.

Applications should begin in everyday experience: facial and license-plate recognition, speech recognition, machine translation, robot vacuum cleaners, autonomous driving, search and recommendation, and image and text generation. Each application should display both capability and limitation so that "can do" is never mistaken for "always gets right."

Interdisciplinary integration can broaden horizons through stories from medicine, scientific research, environmental protection, space exploration, architecture, and art. The main task is to show that AI can enter many domains and that scientists, engineers, physicians, artists, and citizens work together. Complex domain models are unnecessary, but disciplinary evidence and human judgment must remain visible.

Ethics should enter ordinary situations early. Fabricated information, privacy leakage, the limits created by recommendation, authenticity of generated content, attribution, and responsibility can all be discussed through accessible cases. Ethical learning should connect to action: do not enter sensitive personal information; do not appropriate another person's image or voice; disclose AI's role in a work; and verify important information with trusted adults and official sources.

3. Pedagogy: stories, experience, observation, and discussion

Story-based, experiential, and low-threshold activities suit the primary stage. Stories of scientific figures place concepts in history. Physical and "unplugged" simulations reduce dependence on complex software. Images, sounds, classification cards, role-play, and small constructions make abstract mechanisms observable. Discussion and drawing allow children to express understanding.

Experiential activities still require teacher organization and explanation. Open-ended generative tools involve unpredictable content, opaque data use, and dependence risks and should not be used independently and extensively by young children without safeguards. China's *Guidelines for the Use of Generative AI by Primary and Secondary School Students (2025 Edition)* establish age-differentiated rules and state that primary students should not independently use open-ended content-generation functions (Teaching Steering Committee for Basic Education, Ministry of Education, 2025b). UNESCO likewise calls for conditions appropriate to age and development (Miao and Holmes, 2023), while UNICEF emphasizes privacy, safety, fairness, inclusion, and the best interests of children (UNICEF Innocenti, 2025). Primary practice should prioritize teacher-selected materials, closed or controlled environments, anonymized data, and explicit tasks.

Assessment should examine whether students can describe phenomena, ask questions, compare differences, give simple reasons, and follow safety rules. Technical complexity should not be the primary criterion. A child who observes that lighting may affect recognition or asks for a source before forwarding an image already demonstrates important AI literacy.

14.4 Lower-Secondary School: Building a Coherent System and Broad Perspective

1. Principal goal: organize scattered phenomena into a knowledge structure

At lower-secondary level, students' learning in mathematics, science, history, information technology, ethics, and civics becomes more systematic, allowing them to work with more complex causal relations, rule systems, and abstractions. Their independent activity on platforms and with generative AI also increases, requiring clearer understanding of how systems shape information, judgment, and choice.

The central task is to organize rich primary-school experience into a relatively complete system of AI knowledge: where AI came from, its main developmental stages, basic methods, representative applications, and emerging social issues. A broad perspective prevents students from equating a currently popular product with the whole field and helps them understand that technical approaches each have conditions and limits.

2. Knowledge threads: historical development and methodological structure

History can provide the first thread. From ideas of intelligent machines, formal logic, and the computer through the Dartmouth conference, expert systems, machine learning, deep learning, large models, and agents, students see cycles of excitement and setback, competing approaches, and methodological integration. History also shows that breakthroughs depend jointly on theory, computing, data, engineering, and social demand.

The methodological system can be organized around two basic routes. Knowledge-based methods solve problems through rules, facts, search, and reasoning; learning-based methods form models through data, examples, and feedback. Students should encounter game playing, theorem proving, expert systems, and knowledge graphs, together with supervised, unsupervised, and reinforcement learning and basic tasks such as classification, regression, and clustering. The goal is a map of knowledge; technical detail should serve conceptual connections.

The following chain can serve as the central analytical structure:

Real-world problem → data collection and representation → model or rules → output → evaluation and revision

Students use the chain to ask where data came from, how they were labeled, and whether they are representative; how a model forms judgments from data or knowledge; why outputs contain error; what metrics such as accuracy reveal and omit; and why deployment environments differ from training. AI's apparent intelligence is thereby restored to a technical process that can be explained and tested.

3. Applications and ethics: from discussing cases to analyzing mechanisms

Lower-secondary students should continue to study representative applications and frontier achievements, but the emphasis moves toward mechanisms. Facial recognition connects with image representation, feature extraction, similarity, and privacy. Recommendation connects with behavior data, ranking objectives, feedback loops, and filter bubbles. Generative AI connects with training data, probabilistic generation, hallucinations, and content responsibility. Autonomous driving connects with perception, decision, control, and distributed responsibility.

Ethics requires both dedicated treatment and integration with technical learning. Fabrication should involve generation mechanisms, verification, propagation responsibility, and law. Data leakage should involve collection, storage, model input, and personal protection. Social fairness should involve representation, evaluation metrics, and affected groups. Students should identify stakeholders, distinguish technical defects from misuse and institutional arrangements, and explain their reasons.

4. Practice: project-based problem solving, with programming as extension

Practice may include data labeling, small classification experiments, recommendation simulations, flowchart design, verification of generated content, case investigation, and rulemaking. Project-based learning organizes knowledge, collaboration, evidence, and communication around real problems (Krajcik and Blumenfeld, 2006). Projects should preserve the full problem-solving process: define the problem, gather material, design an approach, test it, analyze results, revise the approach, and state limitations.

Programming can deepen understanding of algorithms and models and offer a pathway for interested students. For all learners, however, foundational goals should center on system explanation, data judgment, error analysis, evaluation, and responsibility. Code volume and tool proficiency should not become the principal measures of AI literacy.

Assessment can use explanatory, transfer, judgment, and self-awareness tasks. Students should explain the operation of a simple system, apply an analytical framework to a new application, compare the effects of different data and designs, and reflect on whether they have trusted AI excessively, become dependent on it, or overlooked sources and counterevidence.

14.5 Upper-Secondary School: Extending Knowledge and Innovative Practice

1. Principal goal: move from system understanding into key technologies and the real world

Extending knowledge has three meanings. First, technical understanding advances from basic processes to key mechanisms, enabling students to analyze the principles and evolution of neural networks, deep learning, generative AI, multimodal models, reasoning models, agents, and human–AI collaboration. Second, application learning advances from describing functions to decomposing systems, enabling analysis of how machine vision, machine audition, language processing, search, recommendation, and game-playing systems combine techniques. Third, learning extends from AI itself into mathematics, physics, materials science, chemistry, biology, medicine, astronomy, the humanities, and art.

This extension seeks both depth and transfer while retaining the character of general education. Courses may introduce formulas, network structures, loss and evaluation, and training and testing without compressing university professional education into secondary school. They should provide every student with the foundation needed to explain complex technology and participate in social judgment while offering further directions for those with professional interest.

2. Key technologies: understanding evolution, combination, and boundaries

Modern methods should be situated in a complete knowledge system. Students can compare handcrafted features with representation learning and understand hierarchical feature extraction in deep networks. They can move from language models to large language models and examine pretraining, contextual generation, retrieval, and tool use. They can analyze how multimodal models combine text, image, sound, and action and how agents add goals, memory, planning, action, and feedback.

Instruction should present both development and failure conditions. Performance arises jointly from data, architecture, training, computing resources, and evaluation. Larger models may gain capabilities while increasing cost, opacity, bias, adversarial vulnerability, and governance difficulty. Students should state the experimental conditions supporting a conclusion, distinguish benchmark performance from real-world behavior, and understand the distance between local metrics and system reliability.

3. Interdisciplinary integration: from technical transfer to disciplinary evidence

Each interdisciplinary case should present a complete chain:

Domain problem → problem representation → data and knowledge → method selection and adaptation → domain validation → interpretation and responsibility

In mathematics, generated conjectures and proofs require formal verification. In materials science and chemistry, structure prediction and reaction classification must satisfy physical and chemical constraints and experimental tests. In biology and medicine, outputs must be integrated with mechanisms, clinical evidence, and safety requirements. In astronomy, anomaly detection and classification must distinguish observational signals from instrument noise and human interference. In the humanities, recognition and generation must be considered through historical sources, context, authorship, and cultural value.

Students thereby broaden their professional horizons and understand that AI's generality is conditional. Success on one kind of data does not license direct application everywhere. Domain knowledge, problem representation, and validation standards determine whether transfer is valid.

4. Innovative practice: from producing artifacts to research-oriented inquiry

Upper-secondary practice should increase openness and evidential demands. Students can investigate authentic problems in the school, community, environment, health, culture, or disciplines through data curation, model experiments, system design, user research, and iterative improvement. Research-oriented inquiry requires an account of the problem's source, methodological rationale, evaluation criteria, results, failures, and domain of validity.

Innovation also includes improving existing systems. Students may compare models or datasets for fairness and robustness; design a human–AI workflow specifying model tasks and human review; or propose data, permission, oversight, and appeal mechanisms for high-risk settings. Innovation is thereby joined to responsibility.

5. Complex judgment and professional directions

Upper-secondary students begin to form clearer disciplinary interests and images of future development. Through frontier questions in many fields, AIGE should show how AI researchers, domain scientists, engineers, physicians, teachers, legal professionals, designers, and public-

governance practitioners participate in an intelligent society. Guidance should provide realistic and plural options rather than prematurely confining students to a single technical pathway.

14.6 Spiral Development of the Five Content Areas Across Three Stages

All five areas should appear at every stage; differences lie in the depth of questions, organization of knowledge, and learning tasks.

Table 14.2: Progression of the five content areas across educational stages

Content area	Primary: cultivate interest	Lower secondary: build a system	Upper secondary: extend knowledge
Foundational concepts	Use stories and daily phenomena to introduce intelligent machines, human intelligence, AI, automation, and an initial historical sense	Study the field's birth, stages, major routes, and easily confused concepts systematically to establish common coordinates	Analyze disciplinary boundaries, technical evolution, and relationships between AI and social structure in light of current frontiers and future trends
AI methods	Experience rules, classification, comparison, reward, and learning from examples; use analogy to encounter data, models, and neural networks	Establish the overall structure of knowledge-based and learning-based methods and understand the data–model–output–evaluation chain and basic learning tasks	Examine deep learning, large models, multimodality, reasoning models, agents, and system integration, analyzing performance and boundaries
AI applications	Observe nearby applications and recognize both assistance and error	Decompose the input, representation, model, output, evaluation, and feedback of representative systems and analyze sources of error	Analyze key techniques, system architectures, methodological evolution, and conditions of real deployment
Interdisciplinary integration	Broaden imagination through medicine, science, environment, art, and other stories and understand multidisciplinary collaboration	Encounter basic applications across disciplines and identify initial differences in data forms and evidence	Conduct inquiry involving problem representation, method transfer, domain constraints, experimental validation, and scientific interpretation

Content area	Primary: cultivate interest	Lower secondary: build a system	Upper secondary: extend knowledge
Risk, ethics, and responsibility	Develop habits of safety, privacy, honesty, respect, and verification and seek adult help when necessary	Analyze fabrication, leakage, filter bubbles, bias, fairness, and responsibility and explain reasons and relevant actors	Weigh evidence in high-risk situations and value conflicts and discuss institutional design, professional responsibility, and public governance

The spiral should also shape case selection. The same application may recur, but the task must change. For recommendation, primary students may ask why a platform repeatedly shows similar content and learn to broaden information sources. Lower-secondary students may simulate behavior data, similarity, and feedback loops and explain filter bubbles. Upper-secondary students may examine objective functions, platform business models, public discourse, user autonomy, and governance responsibility. The case remains continuous while the cognitive task develops.

14.7 Principles of Curriculum Organization and Instruction

Maintain a stable core while incorporating frontier developments. Rapid change makes stable concepts and basic questions the appropriate curricular backbone. New models and applications should be located within that structure. Large, multimodal, and reasoning models and agents should be introduced by explaining what they inherit, which problems they solve, where they belong, and what new risks they create. This reduces curricular volatility driven by product names.

Combine independent courses with integration across subjects. Independent AIGE courses carry systematic concepts, methodological structure, and ethics. Chinese language, mathematics, science, history, art, information technology, and integrated practice provide disciplinary connections and transfer contexts. Both are needed for a complete system and authentic application.

Make scientific and humanistic education mutually supporting. Technical principles and social judgment should meet in the same curriculum. Data and models entail data rights, representation, and objective selection. Generation and recommendation entail authenticity, authorship, attention, and public effects. Medicine and autonomous driving entail professional responsibility and the rights of affected people. Ethics becomes a dimension of technical understanding, while technical understanding supplies facts for ethical judgment.

Combine a common foundation with flexible extension. All students need foundational concepts, historical development, data and model awareness, analysis of representative applications, interdisciplinary perspective, responsibility, and cognitive autonomy. Programming, advanced mathematics, hardware construction, model training, and research projects may be tiered extensions for different interests and backgrounds. The common foundation supports equity; extension supports individual development.

Preserve the complete cognitive process in practice. Practice should go beyond clicking tools and displaying products. Primary activities should include observation and expression. Lower-secondary projects should include problem, data, design, evaluation, and reflection. Upper-secondary inquiry should include hypothesis, experiment, evidence, iteration, and boundaries. Process records should reveal why a choice was made, what evidence supported it, where it failed, and how it was corrected.

Align assessment with stage goals. Assessment serves cognitive autonomy and adjusts its depth by stage. Primary assessment emphasizes description, initial questioning, and safe habits. Lower-secondary assessment emphasizes system explanation, framework transfer, bias identification, and reflection. Upper-secondary assessment emphasizes complex-system analysis, interdisciplinary evidence, design improvement, and responsibility judgment.

14.8 Developmental Progression in Cognitive Autonomy

The articulated curriculum ultimately cultivates the capacity for cognitive autonomy. Source vigilance, evidence sensitivity, boundary awareness, trust calibration, cognitive agency, and responsibility judgment run through all three stages. Induction, generalization, judgment, and self-awareness deepen with the complexity of materials and tasks. Primary education creates initial cognitive distance from technology. Lower-secondary education translates questioning into mechanism-based analysis. Upper-secondary education advances toward cognitive action that verifies evidence, states boundaries, and accepts responsibility.

Table 14.3: Progression of the four cognitive capacities across educational stages

Capacity	Primary	Lower secondary	Upper secondary
Induction	Discover common AI phenomena across everyday cases	Understand how machines form regularities from examples, rules, and feedback	Analyze the conditions, bias, generalization, and uncertainty of model induction
Generalization	Apply the need to check results across tools and information	Transfer the data–model–output–evaluation framework to new applications	Transfer AI methods to interdisciplinary problems while controlling boundaries
Judgment	Distinguish help from risk, authentic from suspicious, and safe from unsafe	Compare data, evidence, metrics, and costs of error and explain reasons	Weigh multiple interests, risk levels, professional evidence, institutions, and responsibility

Capacity	Primary	Lower secondary	Upper secondary
Self-awareness	Notice whether one believes too readily, reveals information, or treats a machine as a person	Reflect on overreliance, ignored sources, and neglected counterevidence	Monitor the cognitive division of labor, trust calibration, value position, and responsibility in human–AI collaboration

14.9 Chapter Summary

K–12 AIGE maintains curricular completeness through five content areas and forms a spiral progression through primary, lower-secondary, and upper-secondary education. Primary school emphasizes interest and the scientific spirit; lower-secondary school, coherent system construction and a broad perspective; upper-secondary school, knowledge extension and innovative practice. Through increasingly demanding explanation, transfer, judgment, and reflection, students move from approaching AI with initial questioning toward understanding systems, testing evidence, controlling the boundaries of transfer, and accepting responsibility.

Chapter 15

Artificial Intelligence General Education at University

Abstract. Building on the basic understanding formed in school education, Artificial Intelligence General Education at university emphasizes understanding methods, transferring them into professional fields, and collaborating across disciplines. Courses should teach the principal AI methods and research paradigms systematically and use shared cases from mathematics, engineering, the physical sciences, biology and medicine, Earth and space sciences, and the social, cultural, and artistic domains to train students to analyze domain problems, conditions of methodological applicability, standards of evidence, and boundaries of responsibility. Induction, generalization, judgment, and self-awareness extend from general problems into the formation, testing, and application of professional knowledge. Instruction should combine theory, shared cases, and practical activities. During a transitional period in which students' backgrounds differ markedly, a concepts–methods–applications–integration structure can be used, with progressively greater weight given to methodological depth, professional transfer, and interdisciplinary practice.

Keywords: university Artificial Intelligence General Education; methodological system; interdisciplinary integration; case analysis; practical teaching; project learning; curricular transition; cognitive autonomy

15.1 Guiding Ideas

15.1.1 Goals of University AIGE

University students have entered particular fields and begun to encounter their disciplines' systems of questions, research methods, and standards of evidence. Compared with school education, university AIGE must deepen theory and situate AI within professional knowledge and authentic practice. Its goals can be summarized in four directions: moving from general to professional content, understanding AI's foundational ways of thinking and research paradigms, mastering a framework for interdisciplinary integration, and developing habits of interdisciplinary collaboration.

1. From extending knowledge to understanding methods and transferring them professionally

Upper-secondary education can establish a relatively complete structure of AI knowledge and introduce principal methods, applications, and social effects. University education must ask further: What foundational ways of thinking characterize AI? How do they guide the discovery and solution of problems? What happens when they enter a particular field? How can common ways of thinking organize collaborative innovation?

Learning therefore deepens along this path: **understanding the methodological system → analyzing transfer to professional problems → interdisciplinary collaboration and innovation**.

"Understanding the methodological system" concerns relatively stable ways of solving problems. Students should be able to place a new model or system within the genealogy of AI methods and explain what ideas it inherits, what difficulty it addresses, and under what conditions it works.

2. Understanding foundational ways of thinking and research paradigms

Although AI methods change rapidly, several ways of thinking remain relatively stable. Chapter 5 identified seven: formalizing real-world problems, learning regularities from data, respecting constraints imposed by domain knowledge, making judgments under uncertainty, trading performance against efficiency, preferring simple methods where possible, and testing solutions in practice. These should not become isolated units. They should enter complete explanations of methods: problem decomposition and computational scale in search, induction and generalization in statistical learning, and knowledge constraints, experimental validation, and capability boundaries in case analysis. Repeatedly returning to these questions develops stable analytical habits.

3. Mastering an analytical framework for interdisciplinary integration

Students need to understand not only where AI is used but how it enters a field. The six-stage framework from Part III can become a common tool for analyzing papers, systems, and practical proposals:

Domain problem → problem formulation and representation → data and knowledge → method selection and adaptation → domain validation → interpretation and responsibility

University-level work raises the professional demands at each stage: define the domain difficulty accurately; specify where AI intervenes; decompose complex tasks; formalize the problem appropriately; identify data, knowledge, and professional constraints; explain why a general method requires adaptation; and validate results through the proof, experiment, clinical study, or professional procedure accepted in that field. Students must also state conditions of applicability, consequences of error, and the responsibilities of participants.

4. Developing habits of interdisciplinary collaboration

Interdisciplinary AI normally depends on several forms of knowledge. Domain experts pose meaningful questions, provide professional knowledge, and confirm results. AI researchers develop

models, algorithms, and computational systems. Data and experimental personnel construct evidence. Specialists in ethics, law, management, and public policy address rights, risk, and implementation.

General education should train students to express problems in their own fields precisely, understand the concepts and evidence used in other fields, and establish shared problem definitions, evaluation criteria, and responsibility arrangements. Collaboration does not require everyone to know everything, but each person must know the basis of their own judgment and which claims require validation by others. The core of interdisciplinary learning is integration of ways of knowing and standards of evidence, not the simple assembly of results produced by separate fields (Borrego and Newswander, 2010).

15.1.2 Defining the Boundary of General Education

Professional AI courses emphasize derivation of algorithms, model design, software implementation, system optimization, and technical innovation. AIGE emphasizes an integrated understanding of methods, judgment about model conditions and boundaries, professional transfer, interdisciplinary collaboration, and public responsibility. Necessary mathematics, code, and experiments can serve these goals without creating a uniform professional technical competency.

University courses should avoid four tendencies:

1. **Diluting a professional course.** Merely removing equations and code does not create general education; goals and content organization must be redesigned.
2. **Accumulating technical terms.** Depth is shown by conceptual relationships and methodological genealogy, not the number of model and algorithm names.
3. **Reducing AI to tools.** Tools can support learning, but prompting, platform operation, and artifact generation cannot replace understanding methods, checking evidence, and judging responsibility.
4. **Chasing frontier terminology.** Fast-changing models, products, and platforms can provide cases, while the stable core should rest on fundamental questions of computation, knowledge, learning, representation, action, and validation.

15.1.3 Cognitive Autonomy at University

The book's definition of cognitive autonomy applies at university: learners make full use of others' knowledge, institutional resources, and AI systems while retaining cognitive agency and continually regulating sources, evidence, boundaries, trust, and responsibility. University expands the objects of training from everyday information to papers, databases, experimental records, professional cases, and real systems and into stages of knowledge production such as problem conception, literature retrieval, data analysis, code generation, experimental design, and professional communication.

The four capacities remain induction, generalization, judgment, and self-awareness. Induction forms testable understanding from multiple professional sources. Generalization tests whether conclusions hold across samples, settings, populations, times, and disciplines. Judgment establishes professional conclusions and thresholds for action amid incomplete evidence, competing methods,

and uncertain consequences. Self-awareness monitors how AI changes problem frames, analytical paths, confidence, and responsibility. Common foundations, disciplinary tasks, and integrated projects can carry the complete training cycle, while process records, dependency audits, and oral defenses implement academic norms and responsibility for human–AI collaboration (Miao and Holmes, 2023; Miao, Shiohira, and Lao, 2024).

Table 15.1: Training priorities for the four cognitive capacities at university

Capacity	University-level development	AIGE training content	Observable learning evidence
Induction	Form concepts, regularities, hypotheses, or explanations from papers, data, cases, and experiments	Synthesis of multiple sources, analysis of sampling and measurement, distinction between correlation and causation, competing explanations, and hypothesis generation	Evidence maps, testable hypotheses, counterexamples, and validation plans
Generalization	Judge whether conclusions hold across samples, settings, groups, time, and disciplines	Distribution shift, external validity, cross-context testing, method transfer, and failure-boundary analysis	Conditions of applicability, tests in new settings, and boundary statements
Judgment	Form professional judgments amid incomplete evidence, competing methods, and uncertain consequences	Problem definition, source checking, evidence grading, method comparison, expression of uncertainty, and thresholds for action	Reasoned conclusions, reservations, responsibility statements, and defense
Self-awareness	Identify AI's influence on problem frames, reasoning paths, trust, and responsibility	Human–AI process records, dependency audits, reconstruction without AI, reflective reports, and oral defense	Ability to distinguish one's own contribution from AI's and explain why recommendations were accepted or rejected

15.2 Content Organization

University courses retain the five-part framework from Part III. Foundational concepts and application analysis are embedded within teaching units; risk, ethics, and responsibility combine independent topics with integration throughout. The main curriculum can contain two parts: a systematic foundation in AI methods providing shared knowledge and analytical tools, followed by interdisciplinary integration that carries those methods into different domains.

15.2.1 Part One: Foundational AI Methods

Part One consists of six chapters progressing from disciplinary coordinates and computational foundations through knowledge-driven and data-driven learning and representation learning to foundation models and systems capable of action.

Introduction to AI: Origins, Development, and Basic Paradigms

This chapter establishes the course's coordinates:

- intellectual origins and formation of the field;
- principal stages and reasons for advances and setbacks;
- scientific, engineering, and interdisciplinary attributes;
- relationships among AI, automation, robotics, machine learning, deep learning, large models, and agents;
- symbolic, Bayesian, connectionist, evolutionary, and bio-inspired traditions; and
- foundational ways of thinking and research paradigms.

It should provide a map of methods showing large models as an important stage while preserving the broader field of knowledge, search, learning, optimization, control, and experiment.

Computation, Complexity, and Problems of Intelligence

This chapter introduces:

- information, data, programs, and universal computers;
- algorithms and computational processes;
- computability and its basic boundaries;
- time and space complexity;
- polynomial and exponential growth and combinatorial explosion;
- problem scale, computational resources, and practical tractability;
- decomposition, heuristics, approximation, randomization, parallelism, and simulation; and
- learning as one route for addressing complex problems.

Students should understand that a computable problem may remain practically unsolvable because of resource demands. One important role of AI is to use heuristics, experience from data, and large-scale computation to find usable solutions in complex spaces.

Knowledge Representation, Reasoning, and Search

This chapter covers knowledge-driven AI:

- states, actions, goals, rules, and constraints;
- state spaces and problem representation;
- uninformed and heuristic search;
- adversarial search, planning, and constraint satisfaction;
- propositional and predicate logic;
- production rules and expert systems;
- semantic networks, knowledge graphs, and knowledge bases; and

- the knowledge-acquisition bottleneck, combinatorial explosion, rule conflicts, and uncertainty.

Problem representation becomes concrete: what state space represents a real task, which actions are available, how search evaluates candidates, and how knowledge reduces wasted computation. Search, planning, and expert systems also illustrate what AI borrows from human problem solving.

Statistical Learning and Decision-Making

This chapter explains how machines form regularities from data and feedback:

- tasks, data, models, objectives, learning algorithms, domain knowledge, and metrics;
- supervised, unsupervised, self-supervised, and reinforcement learning;
- classification, regression, clustering, and dimensionality reduction;
- linear models, decision trees, probabilistic models, and ensembles;
- training, validation, and test sets;
- overfitting, underfitting, generalization, and model selection;
- bias and variance, uncertainty, and distribution shift; and
- baselines, controls, cross-validation, ablation, independent testing, and error analysis.

Experimental validation should become systematic here. Students should understand that performance arises jointly from data, assumptions, algorithms, and evaluation conditions and that only appropriate comparison and independent testing support a corresponding conclusion.

Neural Networks and Deep Learning

This chapter centers on representation learning:

- inspiration from biological nervous systems and its engineering transformation;
- artificial neurons, perceptrons, and multilayer networks;
- nonlinear representations, loss functions, gradient descent, and backpropagation;
- representation learning and hierarchy;
- convolutional networks and spatial structure;
- recurrent networks and sequential structure;
- autoencoders and generative models;
- attention and Transformers;
- joint effects of data, computing, algorithms, and architecture; and
- explainability, adversarial vulnerability, and out-of-distribution failure.

The central methodological change is that models can learn multilayer representations from data and that many simple computational units can produce complex functions through connection and training. Convolutional, recurrent, and Transformer architectures should represent different problem structures rather than become entries in a catalogue of names.

Large Models and Agentic Systems

This chapter examines the movement from specialized models to foundation models and systems capable of action:

- large-scale self-supervised pretraining;
- relationships among data, scale, and computation;
- transfer, prompting, and in-context learning;
- instruction tuning and post-training;
- multimodal representation, alignment, and fusion;
- retrieval augmentation and external knowledge;
- tool use, code, and professional software;
- memory, planning, action, and feedback;
- agent goals, permissions, stopping conditions, and human oversight;
- division of labor and collaboration among agents; and
- hallucination, compounding errors, system security, and responsibility.

An agent should be understood as a system comprising models, knowledge, memory, tools, environment, and oversight. Once generation becomes continuing action, errors can propagate across stages; reliability depends jointly on permission controls, external validation, and human oversight.

Table 15.2: Content structure of Part One of a university AIGE course

Chapter	Core question	Principal content
Introduction to AI	What kind of field is AI?	Origins, history, disciplinary attributes, conceptual relationships, and basic paradigms
Computation, complexity, and intelligence	Why can computation address intelligent problems, and what limits it?	Universal computation, computability, complexity, growth of scale, and solution strategies
Knowledge representation, reasoning, and search	How can explicit human knowledge become machine capability?	State spaces, search, planning, logic, rules, and knowledge systems
Statistical learning and decision-making	How do machines form regularities from data and feedback?	Statistical models, learning modes, generalization, uncertainty, and experimental validation
Neural networks and deep learning	How do models form complex representations automatically?	Neural networks, backpropagation, representation learning, and representative deep architectures

Chapter	Core question	Principal content
Large models and agentic systems	How do general models connect knowledge, tools, and environments?	Pretraining, multimodality, retrieval, tools, planning, action, and oversight

Each chapter may include a "methodological review" asking how the method represents its problem, what computational process it uses, where it obtains knowledge or regularities, how domain knowledge enters, what experiments support conclusions, and what changes could cause failure. The foundational ways of thinking thus run through all six chapters.

15.2.2 Part Two: Interdisciplinary Integration of AI

Part Two contains six disciplinary directions, each taught through shared cases organized by the instructor. Shared cases establish common understanding and enable comparison among standards of evidence. Papers and systems selected by students support independent case study and practice but do not replace common cases.

The detailed knowledge appears in Part III. University organization should foreground domain standards of evidence and use public cases to establish a comparative framework.

Table 15.3: Teaching priorities across six interdisciplinary directions at university

Direction	Representative points of AI intervention	Priorities for domain validation and responsibility
Mathematics	Pattern discovery, counterexample search, conjecture generation, proof strategy, and formal verification	Data patterns can suggest conjectures and search directions, but mathematical conclusions still require rigorous proof; cases may include ML-assisted conjectures, auxiliary constructions in geometry, and automated theorem proving (Davies, Veličković, Buesing, et al., 2021; Trinh et al., 2024)
Engineering	Perception, system modeling, design optimization, planning and control, fault diagnosis, and maintenance	Evaluation should cover real-time performance, stability, extreme conditions, redundancy, and human takeover through simulation, experiment, and field testing

Direction	Representative points of AI intervention	Priorities for domain validation and responsibility
Physical sciences	Data processing, parameter inversion, structure and property prediction, candidate generation, process optimization, and experimental planning	Models must respect conservation, symmetry, dimensions, boundary conditions, and chemical feasibility; candidates require simulation, synthesis, and experimental validation (Kench and Cooper, 2021; Schwaller, Laino, et al., 2019; Schwaller, Probst, Vaucher, et al., 2021)
Biology and medicine	Sequence and structure analysis, imaging, risk prediction, drug discovery, and clinical decision support	Population and device differences, privacy, informed consent, and fairness matter; high-risk conclusions require independent validation, clinical trials, and professional review (Jumper, R. Evans, Pritzel, et al., 2021; Jo, Park, Jung, et al., 2017; Ott, Hu, Keskin, et al., 2017; Sahin, Derhovanessian, Miller, et al., 2017)
Earth and space	Remote sensing, anomaly detection, spatiotemporal prediction, system reconstruction, ecological monitoring, and mission planning	Observational noise, multiscale change, rare extremes, and compounding error require physical processes, multiple sources of evidence, and expert consultation (Reichstein et al., 2019; Kerrigan, La Plante, Kohn, et al., 2019)
Society, culture, and art	Analysis of social records, policy simulation, recognition of cultural objects, digital restoration, and human–AI co-creation	Examine institutional and measurement effects on data and address context, interpretation, fairness, privacy, copyright, attribution, and cultural diversity (D. Lazer, Pentland, L. Adamic, et al., 2009; P. Wang et al., 2024; C.-Z. A. Huang, A. Vaswani, Uszkoreit, et al., 2018; Manovich, 2020)

15.2.3 The Place of Risk, Ethics, and Responsibility

Risk, ethics, and responsibility combine dedicated topics with integration throughout. Dedicated topics cover data and privacy, intellectual property, fairness, reliability, safety, human agency, social effects, highly autonomous systems, chains of responsibility, auditing, and appeal. Integration requires students to analyze data bias and metrics in statistical learning; explainability and adversarial vulnerability in deep learning; hallucinations, permissions, and compounding errors in large models and agents; and responsibility in all six disciplinary directions according to domain evidence and the consequences of error.

15.3 Modes of Teaching

Teaching modes should respond jointly to the nature of knowledge, student characteristics, and practical conditions. AI knowledge is systematic and its methods interdependent, yet AI is also practical and experimental: many questions become intelligible only through observation and comparison. Students come from different majors and vary in mathematics, programming, and research experience. Knowledge evolves rapidly across papers, open-source projects, and online systems. Generative AI has entered retrieval, writing, coding, and creative production. University courses must therefore retain systematic theory, shared case analysis, and diverse practice.

15.3.1 Basic Teaching Principles

Systematic Theoretical Learning

Students need a complete methodological system and an understanding of relationships, origins, and boundaries. Lectures establish common concepts, explain structure, clarify confusions, and model analysis. Theory should center on what a method solves, which representations and assumptions it uses, and how it is validated, reducing isolated terms and equations.

Instructor-Organized Shared Cases

Cases should span the six directions and present the complete chain from domain problem through representation, data and knowledge, method adaptation, domain validation, interpretation, and responsibility. Shared cases establish analytical standards for comparing mathematical proof, engineering testing, materials experiments, clinical validation, Earth observation, and humanistic interpretation.

Experiments and Open Practice

Students should observe how methods work through visualization, data comparison, parameter variation, error analysis, system testing, and small projects and judge the conditions under which theoretical claims hold. Practice asks students to pose questions, state expectations, change conditions, record results, and explain differences. It may include:

- visualization experiments and analysis of model mechanisms;
- comparisons of data, parameters, and metrics;
- reading papers and reproducing key results;
- literature research and domain reviews;
- academic debate, essays, peer review, and presentations;
- testing system or tool capabilities and boundaries;
- feasibility analysis of professional problems; and
- design of research projects and solutions to real problems.

Important papers, open projects, and new problems should remain within a stable methodological framework. Independent case study, reproduction, capability testing, and open projects

serve different purposes. Team projects should create genuine knowledge interaction by assigning roles concerning the domain problem, AI method, data and experiment, ethics and law, and implementation, with all members contributing to problem definition and evidence judgment.

Combining Common Goals with Flexible Practical Pathways

Diverse student backgrounds require flexible pathways above a common goal. Students may use no-code tools, visualization, AI-assisted code, independent programming, literature research, or field investigation according to their disciplines and interests. Every pathway must answer the same questions: What is the task? Why is the method appropriate? What data, knowledge, and assumptions are used? How are results validated? Under what conditions might they fail? What responsibility remains human?

15.3.2 Optional Implementation Models

Institutions differ in course size, staff, and resources, so no single model is appropriate. The following can be used separately or combined.

Model One: A Modular Lecture–Case–Experiment Course

Each module follows theory, shared case, short practice, and concluding discussion. This suits large classes, diverse backgrounds, and first offerings. One- or two-week tasks expose students repeatedly to different methods and fields.

Model Two: Foundational Methods Followed by an Interdisciplinary Project

The first half establishes common methods; the second assigns a three- to five-week project. After learning the framework, teams define a problem, review literature, analyze data and knowledge, conduct a small experiment or develop a proposal, perform domain validation, and analyze responsibility. This provides a complete project experience without requiring premature topic choice.

Model Three: Professionally Oriented General Education

The course concentrates on AI and medicine, materials science, social science, art, urban governance, or another field. A common AI foundation precedes deeper study of intervention points, data conditions, method transfer, and evidence. This suits courses within a school or collaborations across schools.

Model Four: Lectures, Paper Reading, and Seminars

Concise lectures establish the framework; students study classic or frontier papers in directions of interest. Outputs can include critical reviews, domain surveys, research proposals, posters, or simulated conferences. This suits small classes, advanced students, and research universities.

Model Five: Entrepreneurial Learning

After an AI foundation, mentored teams analyze an intervention point, data and knowledge conditions, feasibility, cost, organizational process, law and ethics, risk, and implementation. Outputs may be feasibility studies, implementation plans, or prototypes. This suits application-oriented institutions.

15.3.3 Assessment

Assessment should be plural, combining concept tests, case analyses, laboratory reports, paper critiques, project outputs, peer review, individual defenses, and process records. A final artifact is only one form of evidence; students must explain judgments, respond to counterexamples, and state boundaries.

15.4 A Transitional Curriculum

Students now entering university have highly variable prior AI education, and many lack a complete knowledge system. University curricula must address this while retaining methodological depth, professional transfer, and interdisciplinary practice. A transitional concepts–methods–applications–integration structure can draw on upper-secondary content while raising evidential requirements. As K–12 AI education becomes widespread, foundational concepts and applications can be compressed and methodological depth, integration, and innovation increased.

The four modules retain Part III's structure while increasing university-level demands:

Table 15.4: Four modules in a transitional university curriculum

Module	Common content	University-level development
Concepts	Origins, history, disciplinary characteristics, and relationships among automation, robotics, machine learning, deep learning, large models, and agents	Add conceptual comparison, historical explanation, and disciplinary-boundary analysis to establish common language and coordinates
Methods	Computation and complexity, knowledge and search, statistical learning, neural networks, large models, multimodality, and agents	Emphasize relationships, theoretical conditions, experimental validation, generalization, uncertainty, and system boundaries
Applications	Representative systems in machine vision, machine audition, language processing, embodied action, search, and recommendation	Add papers, real systems, and failures and analyze system components, deployment conditions, feedback, and responsibility

Module	Common content	University-level development
Integration	Mathematics, engineering, physical sciences, biology and medicine, Earth and space, society, culture, and art	Compare domain problems, method adaptation, standards of evidence, and responsibility through shared and independent cases

15.4.1 Adding Interdisciplinary Practice

At least one interdisciplinary activity should follow study of concepts, methods, and basic applications. Options include:

- reproduction or structured analysis of a research case;
- feasibility analysis combining AI with the student's major;
- a plan for urban or industrial AI deployment;
- testing a model's capabilities and boundaries on professional data;
- designing a workflow for AI-assisted research or professional practice; and
- an interdisciplinary real-world problem or entrepreneurial proposal.

The aim is to apply shared methods to professional problems and understand collaboration through genuine divisions of knowledge. Outputs should include problem definition, sources, methodological rationale, experiment or proposal, validation, records of failure and risk, responsibility, and individual reflection.

15.4.2 Progressive Adjustment During the Transition

As school AI education becomes widespread, transitional courses can:

- compress repeated concepts and basic applications;
- deepen computation, learning, deep models, and agents;
- add paper reading, experimental analysis, and frontier research;
- increase integration and professional transfer; and
- move interdisciplinary projects from introductory experience toward research and innovation.

The transitional curriculum responds to current differences in preparation and gradually approaches the full goals of university AIGE; it need not become a permanent separate system.

15.5 Chapter Summary

University AIGE deepens understanding of methods, professional transfer, and interdisciplinary collaboration. Its main body joins foundational AI methods with interdisciplinary integration and connects professional evidence with responsibility judgment through theory, shared cases, and diverse practice. Induction, generalization, judgment, and self-awareness enter the formation, testing, and application of professional knowledge. A transitional concepts–methods–applications–integration structure can meet present needs; as school education improves, universities should

reduce repetition and increase methodological depth, paper reading, professional cases, interdisciplinary practice, and innovative research.

Chapter 16

Artificial Intelligence General Education for Lifelong Learning

Abstract. Lifelong learning extends across the whole course of life. This chapter focuses on adult and public learning after formal schooling and encompasses professional development, family life, civic participation, and retirement. It adopts a common core–contextual modules–frontier updates structure. The common core provides stable understanding; contextual modules connect learning with work, family, health and care, cultural creation, and professional practice; frontier updates incorporate new results, applications, and questions. Courses begin from real situations, use foundational principles to support practice, and offer differentiated support according to learners' stages of life, baseline digital skills, and conditions of learning, allowing cognitive autonomy to develop continuously amid technological change.

Keywords: lifelong learning; Artificial Intelligence General Education; adult learning; situated learning; public science communication; digital inclusion; cognitive autonomy

16.1 Guiding Ideas

1. The basic position of lifelong learning

Lifelong learning extends throughout life. This chapter concentrates on implementation across the stages of adulthood. Learners may be pursuing professional development, changing occupations, caring for family members, participating in public life, or living in retirement. Their needs change with living conditions, work tasks, and technology. Lifelong-learning curricula cannot copy the fixed entry point, common pace, and single pathway of school curricula. They should provide repeated entry, continuing study, and autonomous choice above a common foundation.

Nor should lifelong learning be reduced to education for older adults. Older learners are one important group, but employed people, career changers, adults learning in family contexts, and the general public all need AIGE. Their concrete needs differ, yet all should understand basic AI operation, use AI appropriately, and judge the authenticity, applicability, and risk of its outputs.

Adult learning is closely connected with existing experience, practical tasks, and self-directed goals. Educational research shows that content connected with problems learners currently

face more readily becomes sustained action (Knowles, Holton, and Swanson, 2015; Merriam and Bierema, 2013). Lifelong-learning curricula should therefore join stable principles to real situations. Learners understand AI while solving life and work problems and transfer their understanding to new tools and tasks.

2. From one-time learning to continuous cognitive updating

AI technologies, products, and social applications change continually. Instructions for one tool may soon become obsolete, while basic understanding of data, models, generation, verification, and responsibility is more stable. Lifelong learning should help people use foundational concepts to understand new models, transfer prior understanding to new tools, revise judgments in light of new evidence and tasks, and seek information or professional help when they recognize gaps in knowledge.

UNESCO understands lifelong learning as continuing capacity-building across the life course and multiple settings and emphasizes that individuals, communities, and society should jointly provide learning opportunities (UNESCO Institute for Lifelong Learning, 2020). AIGE should provide a structure learners can re-enter after technological change to revisit foundational concepts, add cases, and correct earlier judgments.

3. Driving learning through problems in life and work

Lifelong learning is strongly problem-oriented. AI knowledge should connect with current tasks: acquiring and verifying information, producing text and images, improving efficiency, understanding industry change, using public services, managing health and family affairs, recognizing fraud and fabricated content, and participating in public discussion of AI.

Entering through situations does not mean reducing education to utilitarian tool use. Real problems supply motivation, foundational concepts explain phenomena, practice tests understanding, and risk analysis establishes boundaries. A complete unit can follow this process:

Real problem → AI task → foundational principle → practical operation → verification → transfer and reflection

For example, in using AI to plan travel, a learner should state the number of travelers, budget, mobility limitations, and other constraints; understand why the model may use outdated information; confirm tickets and opening hours through official channels; and protect personal itinerary data.

4. Cognitive autonomy in lifelong learning

Cognitive autonomy in lifelong learning remains autonomy within open dependence. Adults can use AI, expert knowledge, public institutions, and social support fully while retaining cognitive agency and continually regulating sources, evidence, boundaries, trust, and responsibility. They must be able to inspect, adjust, or withdraw trust when necessary.

Induction summarizes system behavior across repeated observations instead of generalizing from a single success or failure. Generalization transfers understanding of data, models, generation, and verification to new tools while rechecking conditions. Judgment evaluates credibility, consequences, and professional boundaries. Self-awareness recognizes gaps in knowledge, changes

in trust, and dependence on tools while preserving choices and responsibilities that must remain one's own.

16.2 Content Organization

16.2.1 Overall Structure

Lifelong-learning content should combine a **common core + contextual modules + frontier updates**.

The common core provides a stable foundation. Contextual modules address tasks in life and work. Frontier updates absorb rapidly changing technologies and social questions. These layers interact: learners may enter through a contextual problem, obtain an explanation from the common core, and extend understanding through frontier updates.

The structure organizes implementation, while the five content areas established in Part III—foundational concepts, AI methods, AI applications, interdisciplinary integration, and risk, ethics, and responsibility—continue to define the scope of knowledge and are distributed across the common core and contextual modules.

Table 16.1: AIGE content structure for lifelong learning

Layer	Principal function	Principal content
Common core	Establish stable foundational understanding of AI	Concepts and social context, nearby systems, foundational methods, large models and agents, and risk, ethics, and responsibility
Contextual modules	Bring common knowledge into real life and work	Learning and information, work and development, family life, health and care, cultural creation, and professional integration
Frontier updates	Support continuous cognitive updating	New methods, models, applications, and topics of public debate

16.2.2 Common Core

Artificial Intelligence and the Intelligent Society

This unit introduces the long pursuit of intelligent machines; distinguishes automated machines from AI; and explains AI's definition, disciplinary characteristics, and developmental trajectory. Learners should understand that AI uses computation to simulate intelligent behaviors such as perception, learning, reasoning, planning, and creation and that its general relevance arises from the foundational role of intelligent activity across domains.

AI Systems in Everyday Life

This unit begins with machine vision, machine audition, speech synthesis, robots, search and recommendation, and language understanding. Cases may include facial and license-plate recognition, speech input, voiceprint verification, navigation, robot vacuum cleaners, autonomous vehicles, search ranking, personalized recommendation, and intelligent customer service.

Each case should answer six questions:

1. What enters the system?
2. What task does it perform?
3. On what basis is the output produced?
4. Which conditions affect the result?
5. How might the system fail?
6. What responsibilities belong to the user?

This analysis turns daily experience into structured understanding and reveals both the presence of AI and the conditions governing its capabilities.

Foundational AI Methods

Lifelong learning does not require a full professional account of algorithms, but it does require a basic methodological structure:

1. AI based on knowledge, rules, and search;
2. AI based on data and learning;
3. relationships among data, models, objectives, and evaluation;
4. training, testing, overfitting, and generalization;
5. neural networks and hierarchical representation;
6. the roles of data and computing in modern AI; and
7. relationships among search, learning, action, and feedback.

Cases such as classifying apples and oranges, hierarchical image recognition, and search and learning in Go can turn abstractions into intelligible processes. Learners should be able to explain methods in their own words, without a uniform requirement for mathematical derivation or programming.

Large Models and Agents

This unit introduces language models, large-scale pretraining, image and video generation, multimodal and reasoning models, and agents. It explains how AI developed from task-specific models into general interfaces for multiple tasks and how external knowledge, tools, memory, planning, and feedback support extended workflows.

An agent should be understood as a system comprising model, goal, memory, tools, permissions, environment, and oversight. Learners should distinguish generating an answer from continually executing a task and understand that the latter introduces compounding errors, permission controls, and responsibility tracking.

Risk, Ethics, and Responsibility

Risk, ethics, and responsibility belong to the common core. Topics include:

1. fabricated information, deep synthesis, and identity deception;
2. leakage of personal, biometric, and workplace data;
3. recommender systems and information environments;
4. data bias, social fairness, and digital exclusion;
5. copyright, attribution, and authenticity of generated content;
6. automated decisions in health care, finance, and public services;
7. dependence on AI and human agency; and
8. permissions, control, and responsibility in highly autonomous systems.

Risk education should enter practice. While using question-answering, image, voice, and agentic tools, learners should practice identifying sensitive information, checking sources, obtaining official confirmation, verifying across channels, protecting passwords, and requiring human review for high-risk tasks.

16.2.3 Contextual Modules

Contextual modules situate AI in familiar tasks, supporting intuitive understanding and analysis of where and how technology intervenes. Representative settings follow.

(1) AI-Assisted Learning and Information Access

Topics include AI question answering and search, problem expression, iterative questioning, source comparison, fact and citation verification, reading summaries, translation, knowledge organization, and study planning. Learners should understand that AI increases retrieval efficiency while also producing false facts, fabricated citations, and outdated advice.

Practice can begin with an unfamiliar problem: obtain an AI answer, select its factual claims and sources for verification, and record the revisions from the initial answer to a reliable conclusion.

(2) AI-Assisted Work and Professional Development

This module addresses document processing, information research, data organization, meeting communication, basic automation, and workflow improvement. It asks which tasks are appropriate for AI, which results require professional verification, which organizational and commercial data must not enter public models, and what new human work follows productivity gains.

Specialized evidence and responsibility for particular occupations can be developed in advanced professional and public-responsibility education. This chapter provides the common lifelong-learning foundation.

(3) AI in Family and Everyday Life

Topics include shopping, travel, public and administrative services, use of phones and smart devices, household affairs and communication, smart homes, and information and payment

security. Existing programs often progress from software installation, interface recognition, and basic operation through question answering, writing, images, video, music, and safety checks.

Practice should emphasize expressing constraints and verifying key information. Travel planning, for example, should state age, budget, mobility, and rest needs and confirm reservations, prices, and opening hours officially.

(4) AI in Health, Aging, and Care

Topics include health-information searches, explanation of examination reports, exercise and sleep advice, home safety monitoring, disease-risk prediction, assisted movement, companion AI, and drug discovery. The boundary among health education, risk alerts, and clinical diagnosis must be explicit.

AI can help organize questions for a medical visit, explain general terminology, and structure personal records. It cannot, on its own, justify changing a prescription or replacing a physician's judgment. Privacy, informed consent, misdiagnosis, and care responsibility should accompany every case.

(5) AI and Cultural Creation

Topics include writing, poetry, image generation and restoration, family media, video, music, and human–AI co-creation. Practice should progress from generating an artifact to clarifying intent, evaluating and revising results, describing human contribution, and addressing copyright, privacy, and authenticity.

Restoring old photographs, creating family-memory videos, greeting cards, and personal songs can stimulate interest and support intergenerational collaboration. Learners should avoid altering identity and historical facts and disclose synthesis and modification when material is shared publicly.

(6) Integration of AI with Professional Fields

This module serves learners with occupational or professional experience. It examines the points where AI enters a domain, its workflow, data and knowledge conditions, validation standards, and responsibility boundaries. A paper, working system, or professional task can be analyzed to determine what AI contributes, which stages require redesign, and which conclusions require professional review.

Professional integration builds on the common core and adds content according to industry norms and risk. Medicine, law, finance, engineering, and education each follow distinct evidence standards, data-protection requirements, and professional responsibilities. General-tool outputs cannot replace professional procedures.

16.2.4 Frontier Updates

Frontier updates address new research, emerging applications, and public debates. They should not become fixed lists of product names. Courses should preserve stable conceptual foundations,

replace selected cases periodically, maintain dynamic readings and update logs, and distinguish research results from product promotion and trend forecasts. Strategies include:

1. retain foundational concepts, methods, and principles of responsibility;
2. update frontier cases and practical tools regularly;
3. remove obsolete platform instructions;
4. preserve classic cases with lasting explanatory value;
5. create feedback channels for instructors and learners;
6. build shareable, tiered, sustainably maintained resources; and
7. require professional review of high-risk content in health, safety, and public governance.

16.3 Modes of Learning and Teaching

Lifelong learners differ substantially in age, education, occupation, digital capability, and available time. Adults often enter with a clear problem and interpret new technology through prior experience. Courses should respect that experience while helping revise outdated understanding.

AI is strongly practical. Randomness in outputs, effects of prompting, and failures caused by changes in data and context are best understood through operation and comparison. Theory, case analysis, and practice should therefore remain connected.

16.3.1 Basic Principles

Enter through situations and support with principles. Learning can begin from a real problem and then add the concepts needed to explain it. Situations motivate; principles enable transfer. Operations alone create dependence on a platform, while principles alone may remain remote from adult needs.

Synchronize learning and practice. Every module should contain observation, operation, verification, or reflection. Practice should serve understanding and capacity rather than the production of numerous artifacts.

Combine modules and keep pathways flexible. Relatively independent but composable modules allow learners to choose according to goals and time and re-enter when needed. Modules should share core terminology, verification methods, and principles of responsibility.

Advance in small steps and update continuously. Adult learning can be short, distributed, and repeated. Each session solves a defined problem and leaves an entry point for the next. Continuously updated practical manuals can use lightweight, short-cycle, operational tasks for fragmented time while allowing prepared learners to skip introductory material.

Use peer support and intergenerational collaboration. Community groups, families, colleagues, and volunteers can solve digital and life problems together. Intergenerational work is especially suitable for family media, cultural memory, fraud prevention, and health-information verification. Collaboration offers technical assistance and exchanges life experience and understandings of risk.

16.3.2 Optional Implementation Models

Independent study. Adults with self-directed learning capacity can use introductory texts, practical manuals, online courses, and learning portfolios in a reading–short practice–question log–verification–stage summary cycle. Materials should give clear pathways, necessary explanations, common errors, and further resources; learners should adapt tasks to their own lives.

Community and senior-university courses. These serve retirees, older adults, and members of the public who need face-to-face support. Small groups, slower pacing, explicit steps, repeated practice, and peer support are appropriate. Content should balance convenience, cultural creation, and risk protection and should not collapse into training in phone operation.

Continuing workplace learning. Workplace learning can center on workflows, information and data processing, AI collaboration, organizational data protection, and changing roles. General education provides the common core; industries and organizations add task-specific content and rules.

Professional integration workshops. Learners with domain backgrounds can read AI papers, analyze applications, test tools, design small feasibility proposals, and discuss professional validation and responsibility. AI instructors and domain personnel should support the workshop together.

Family and intergenerational learning. Families can study information verification, privacy settings, health information, intelligent services, family media, and cultural creation. Younger members may help with digital operation; older members contribute life experience, family memory, and value judgment. Intergenerational learning reduces exclusion and prevents one-way technological instruction.

16.3.3 Assessment

Uniform examinations should not dominate lifelong-learning assessment. Assessment should help learners recognize progress and identify next steps. It may use:

1. situated tasks;
2. learning portfolios;
3. practical artifacts;
4. risk judgment and information-verification tasks;
5. oral explanation;
6. peer assessment;
7. reflective records; and
8. professional or everyday application proposals.

The four capacities remain the goal. Assessment asks whether learners understand basic principles, complete real tasks, verify results, transfer understanding, protect information and rights, and form reasoned autonomous judgments with AI assistance.

16.4 Science Communication and Public Learning

Science communication is an important entry point for adults and provides social extension and dynamic updating for general education. It can serve four tasks:

1. generate interest among people who have not encountered AI;
2. correct misconceptions such as equating AI with robots or treating large models as omniscient;
3. introduce new research, applications, and risks; and
4. guide the public into continuing courses, reading, and practice.

Science communication cannot independently provide complete general education. A lecture or exhibition may offer information and interest but rarely creates a complete knowledge structure or sustained capacity. It should provide onward pathways such as readings, short activities, lecture series, community learning groups, and online courses.

Forms include expert lectures, public dialogue, interactive experiments, exhibitions, paper explanations, community workshops, family activities, and science-media works. Content should state sources and evidence, balance demonstrations of capability with failures, and avoid product promotion, excessive optimism, and narratives of panic.

16.5 Chapter Summary

AIGE for lifelong learning is characterized by continuous cognitive updating. It uses a common core–contextual modules–frontier updates structure and provides repeated entry through independent study, community courses, workplace learning, professional workshops, intergenerational learning, and public science communication. Beginning from real tasks, courses use foundational principles to support verification, transfer, and reflection and differentiate support according to life stage, baseline digital skills, and learning conditions. Their goal is to enable learners to use AI actively while continually revising understanding, judging prudently, and accepting responsibility.

Appendix A

Primary School

A.1 Integrated Stage Edition

A.1.1 Textbook Contents

This appendix is based on the primary-school textbook in the *Tsinghua Artificial Intelligence General Education Series for Primary, Secondary, and University Education.*

No.	Chapter	Lesson
1	Chapter 1 From Dream to Reality	The Story of Yan Shi
2	Chapter 1 From Dream to Reality	Al-Jazari's Mechanical Band
3	Chapter 1 From Dream to Reality	Artificial Intelligence in Film
4	Chapter 1 From Dream to Reality	What Is Artificial Intelligence?
5	Chapter 2 Artificial Intelligence Around Us	High-Speed-Rail Ticket Inspection
6	Chapter 2 Artificial Intelligence Around Us	Electronic Traffic Police
7	Chapter 2 Artificial Intelligence Around Us	Beauty Cameras
8	Chapter 2 Artificial Intelligence Around Us	Robot Vacuum Cleaners
9	Chapter 2 Artificial Intelligence Around Us	Autonomous Driving

Continued on next page

Table A.1 (continued)

No.	Chapter	Lesson
10	Chapter 2 Artificial Intelligence Around Us	Recommender Systems
11	Chapter 3 Frontiers of Artificial Intelligence	The AI Painter
12	Chapter 3 Frontiers of Artificial Intelligence	The AI Composer
13	Chapter 3 Frontiers of Artificial Intelligence	The AI Poet
14	Chapter 3 Frontiers of Artificial Intelligence	The Story of AlphaGo
15	Chapter 3 Frontiers of Artificial Intelligence	OpenAI and ChatGPT
16	Chapter 3 Frontiers of Artificial Intelligence	The Story of Sora
17	Chapter 3 Frontiers of Artificial Intelligence	The AI Weather Forecaster
18	Chapter 4 Origins of Artificial Intelligence	The Story of Aristotle
19	Chapter 4 Origins of Artificial Intelligence	The Story of George Boole
20	Chapter 4 Origins of Artificial Intelligence	Turing and the Turing Machine
21	Chapter 4 Origins of Artificial Intelligence	The Birth of the Computer
22	Chapter 4 Origins of Artificial Intelligence	Early Conceptions of Machine Intelligence
23	Chapter 4 Origins of Artificial Intelligence	The Dartmouth Conference
24	Chapter 5 History of Artificial Intelligence	The Story of Wu Wenjun
25	Chapter 5 History of Artificial Intelligence	Feigenbaum's Expert Systems

Continued on next page

Table A.1 (continued)

No.	Chapter	Lesson
26	Chapter 5 History of Artificial Intelligence	Deep Blue: A Crowning Achievement
27	Chapter 5 History of Artificial Intelligence	The Rise of Deep Learning
28	Chapter 5 History of Artificial Intelligence	The Era of Large Models
29	Chapter 5 History of Artificial Intelligence	Toward the Future
30	Chapter 6 Foundations of Artificial Intelligence	Understanding Computers
31	Chapter 6 Foundations of Artificial Intelligence	Understanding Computer Programs
32	Chapter 6 Foundations of Artificial Intelligence	What Is an Algorithm?
33	Chapter 6 Foundations of Artificial Intelligence	Knowledge and Intelligence
34	Chapter 6 Foundations of Artificial Intelligence	A Good Machine Must Learn
35	Chapter 7 The Era of Deep Learning	Pitts and His Neuron Model
36	Chapter 7 The Era of Deep Learning	The Perceptron: A Neural Network That Learns
37	Chapter 7 The Era of Deep Learning	The Story of Geoffrey Hinton
38	Chapter 7 The Era of Deep Learning	Fei-Fei Li and the ImageNet Dataset
39	Chapter 7 The Era of Deep Learning	GPUs: From Games to Artificial Intelligence
40	Chapter 7 The Era of Deep Learning	Understanding AlphaGo
41	Chapter 7 The Era of Deep Learning	Exploring Large Language Models

Continued on next page

Table A.1 (continued)

No.	Chapter	Lesson
42	Chapter 7 The Era of Deep Learning	Deep-Learning Challenge: Intelligence That Is Hard to Understand
43	Chapter 7 The Era of Deep Learning	Deep-Learning Challenge: Adversarial Examples
44	Chapter 7 The Era of Deep Learning	Deep-Learning Challenge: Super-Agents

A.1.2 Practice Manual Contents

This appendix is based on the primary-school practice manual in the series, Bo'ai Xinhua Primary School Edition. The activities are organized by chapter as in the original manual.

No.	Chapter	Activity Type	Activity
1.1	Chapter 1	Survey	A Thousand-Year Dream of Intelligent Machines
1.2	Chapter 1	Survey	Finding AI Around Us
2.1	Chapter 2	Scenario Drama	The Station of Time
2.2	Chapter 2	Experiential Activity	The Modern Magic Mirror: Does It Really Make You More Beautiful?
2.3	Chapter 2	Essay	AI-Enabled Robots
3.1	Chapter 3	Experiential Activity	An Illustrated Guide to Seasonal Sprites
3.2	Chapter 3	Experiential Activity	Enriching Class Culture: Designing a Class Song
3.3	Chapter 3	Experiential Activity	A Four-Seasons Poetry Gathering
3.4	Chapter 3	Experiential Activity	An AI Fairy Tale
4.1	Chapter 4	Hands-On Computer-Birth Timeline	The Tunnel of Computer Evolution
5.1	Chapter 5	Essay	AI Takes You into the Future

Continued on next page

Table A.2 (continued)

No.	Chapter	Activity Type	Activity
6.1	Chapter 6	Essay	A Journey Through Computer History
6.2	Chapter 6	AI Mini-Theater	The Growth of Little AI Zhi
7.1	Chapter 7	Debate	Will Future Super-Agents Be Humanity's Friends or Enemies?

A.2 Grade-Level Volumes

This appendix is based on the Grade 3 through Grade 6 textbooks in the series and lists units and sections by semester.

A.2.1 Grade 3, First Semester

No.	Unit	Section
1.1	From Dream to Reality	The Story of Yan Shi's Artificial Man
1.2	From Dream to Reality	Al-Jazari's Mechanical Band
1.3	From Dream to Reality	Artificial Intelligence in Film
1.4	From Dream to Reality	The Story of a Rice Cooker: What Is Artificial Intelligence?
2.1	Artificial Intelligence Around Us (1)	High-Speed-Rail Ticket Inspection
2.2	Artificial Intelligence Around Us (1)	Electronic Traffic Police
2.3	Artificial Intelligence Around Us (1)	Beauty Cameras
3.1	Artificial Intelligence Around Us (2)	Robot Vacuum Cleaners
3.2	Artificial Intelligence Around Us (2)	Autonomous Driving
3.3	Artificial Intelligence Around Us (2)	Recommender Systems

Continued on next page

Table A.3 (continued)

No.	Unit	Section
4.1	Frontiers of Artificial Intelligence	The Story of AlphaGo
4.2	Frontiers of Artificial Intelligence	OpenAI and Large Language Models
4.3	Frontiers of Artificial Intelligence	The Wonders of AI-Generated Video

A.2.2 Grade 3, Second Semester

No.	Unit	Section
1.1	Human Intelligence	Forms of Human Intelligence
1.2	Human Intelligence	IQ and Its Measurement
1.3	Human Intelligence	Emotional Intelligence: Wisdom Beyond Reason
2.1	Origins of Artificial Intelligence (1)	Aristotle and the Syllogism
2.2	Origins of Artificial Intelligence (1)	The Story of George Boole
2.3	Origins of Artificial Intelligence (1)	Turing and the Turing Machine
2.4	Origins of Artificial Intelligence (1)	The Birth of the Computer
3.1	Origins of Artificial Intelligence (2)	Early Conceptions of Machine Intelligence
3.2	Origins of Artificial Intelligence (2)	The Turing Test
3.3	Origins of Artificial Intelligence (2)	The Turing Award
4.1	The Birth of Artificial Intelligence	On the Eve of Artificial Intelligence
4.2	The Birth of Artificial Intelligence	The Dartmouth Conference

A.2.3 Grade 4, First Semester

No.	Unit	Section
1.1	The Golden Decade	The Success of ELIZA
1.2	The Golden Decade	The Story of Wu Wenjun
1.3	The Golden Decade	The Beginning of Machine Translation
1.4	The Golden Decade	The End of the Boom: AI's First Winter
2.1	The Second Boom	Edward Feigenbaum and Expert Systems
2.2	The Second Boom	A Dream Fades: Japan's Fifth-Generation Computer
2.3	The Second Boom	The Era of Robotic Insects
3.1	Steady Recovery	The Rise of Machine Learning
3.2	Steady Recovery	Deep Blue: A Crowning Achievement
3.3	Steady Recovery	Watson: An Encyclopedia That Thinks
4.1	The Present and Future	The Rise of Deep Learning
4.2	The Present and Future	The Era of Large Models
4.3	The Present and Future	Toward the Future

A.2.4 Grade 4, Second Semester

No.	Unit	Section
1.1	AI and Medicine	Surgical Assistance Robots: A Physician's Super Assistant
1.2	AI and Medicine	AI Reads Chest X-Rays: A Physician's Intelligent Eye
1.3	AI and Medicine	AI Helps Develop Cancer Vaccines
2.1	AI and Scientific Research	AlphaFold: AI Predicts Protein Structures
2.2	AI and Scientific Research	AI Helps Mathematicians Form Conjectures
2.3	AI and Scientific Research	AI Supports Space Exploration
2.4	AI and Scientific Research	AI Enhances Microscopic Vision

Continued on next page

Table A.6 (continued)

No.	Unit	Section
3.1	AI and Everyday Life	AI Weather Forecasting
3.2	AI and Everyday Life	The AI Architect
3.3	AI and Everyday Life	AI Protects Wildlife
4.1	AI and Art	The AI Painter
4.2	AI and Art	The AI Musician
4.3	AI and Art	The AI Poet

A.2.5 Grade 5, First Semester

No.	Unit	Section
1.1	AI Risks (1)	Information Fabrication
1.2	AI Risks (1)	Information Leakage
1.3	AI Risks (1)	Filter Bubbles
2.1	AI Risks (2)	AI Hallucinations
2.2	AI Risks (2)	Sycophantic Artificial Intelligence
2.3	AI Risks (2)	Will Robots Rebel Against Humans?
3.1	AI Ethics (1)	Artificial Intelligence and Social Fairness
3.2	AI Ethics (1)	Should AI Help Write School Essays?
3.3	AI Ethics (1)	Copyright Ownership of AI-Generated Works
4.1	AI Ethics (2)	Can AI Serve as a Judge?
4.2	AI Ethics (2)	Who Is Responsible When Autonomous Driving Fails?
4.3	AI Ethics (2)	Does an AI That Comforts You Truly Understand You?

A.2.6 Grade 5, Second Semester

No.	Unit	Section
1.1	Computer Foundations	Understanding Computers

Continued on next page

Table A.8 (continued)

No.	Unit	Section
1.2	Computer Foundations	Computer Programs
1.3	Computer Foundations	Computer Operating Systems
1.4	Computer Foundations	Algorithms
2.1	Knowledge-Based Artificial Intelligence	Knowledge and Intelligence
2.2	Knowledge-Based Artificial Intelligence	Game-Playing Algorithms
2.3	Knowledge-Based Artificial Intelligence	Theorem Proving
2.4	Knowledge-Based Artificial Intelligence	Expert Systems
2.5	Knowledge-Based Artificial Intelligence	Knowledge Graphs
3.1	Foundational Concepts of Machine Learning	How Machines Learn from Examples
3.2	Foundational Concepts of Machine Learning	Learning with Answers: Supervised Learning
3.3	Foundational Concepts of Machine Learning	Learning Without Answers: Unsupervised Learning
3.4	Foundational Concepts of Machine Learning	Learning through Reward: Reinforcement Learning

A.2.7 Grade 6, First Semester

No.	Unit	Section
1.1	Data and Models	Data and Artificial Intelligence
1.2	Data and Models	Preparing Data
1.3	Data and Models	Designing Models
1.4	Data and Models	Training Models
2.1	Common Machine-Learning Tasks	Regression Tasks

Continued on next page

Table A.9 (continued)

No.	Unit	Section
2.2	Common Machine-Learning Tasks	Classification Tasks
2.3	Common Machine-Learning Tasks	Clustering Tasks
3.1	Artificial Neural Networks (1)	The Human Nervous System
3.2	Artificial Neural Networks (1)	Pitts and His Neuron Model
3.3	Artificial Neural Networks (1)	The Perceptron: A Neural Network That Learns
4.1	Artificial Neural Networks (2)	Multilayer Perceptrons
4.2	Artificial Neural Networks (2)	Convolutional Neural Networks
4.3	Artificial Neural Networks (2)	Recurrent Neural Networks

A.2.8 Grade 6, Second Semester

No.	Unit	Section
1.1	Deep Learning	The Story of Geoffrey Hinton
1.2	Deep Learning	The Secrets of Deep Learning
1.3	Deep Learning	Fei-Fei Li and the ImageNet Dataset
1.4	Deep Learning	GPUs: From Games to Artificial Intelligence
1.5	Deep Learning	Understanding AlphaGo
2.1	Large Models and Agents	Language Models
2.2	Large Models and Agents	Large Language Models
2.3	Large Models and Agents	Large Vision Models
2.4	Large Models and Agents	Large Multimodal Models
2.5	Large Models and Agents	Large Reasoning Models
2.6	Large Models and Agents	Agents

Continued on next page

Table A.10 (continued)

No.	Unit	Section
3.1	Challenges of Deep Learning	Intelligence That Is Hard to Understand
3.2	Challenges of Deep Learning	Adversarial Examples
3.3	Challenges of Deep Learning	Super-Agents

Appendix B

Lower-Secondary School

B.1 Integrated Stage Edition

B.1.1 Textbook Contents

This appendix is based on the lower-secondary textbook in the *Tsinghua Artificial Intelligence General Education Series for Primary, Secondary, and University Education.*

No.	Chapter	Lesson
1	Chapter 1 What Is Artificial Intelligence?	The Dream of Intelligent Machines
2	Chapter 1 What Is Artificial Intelligence?	What Is Artificial Intelligence?
3	Chapter 1 What Is Artificial Intelligence?	The Eyes of a Machine
4	Chapter 1 What Is Artificial Intelligence?	The Ears of a Machine
5	Chapter 1 What Is Artificial Intelligence?	The Voice of a Machine
6	Chapter 1 What Is Artificial Intelligence?	The Hands and Feet of a Machine
7	Chapter 2 The Birth of Artificial Intelligence	Origins of Human Intelligence
8	Chapter 2 The Birth of Artificial Intelligence	Formalizing the Laws of Human Thought
9	Chapter 2 The Birth of Artificial Intelligence	The Birth of the Computer

Continued on next page

Table B.1 (continued)

No.	Chapter	Lesson
10	Chapter 2 The Birth of Artificial Intelligence	The Great Alan Turing
11	Chapter 2 The Birth of Artificial Intelligence	The Dartmouth Conference
12	Chapter 3 History of Artificial Intelligence	Dreams and Disillusionment
13	Chapter 3 History of Artificial Intelligence	The Era of Deep Learning
14	Chapter 3 History of Artificial Intelligence	The Era of Large Models
15	Chapter 3 History of Artificial Intelligence	Interdisciplinary Convergence
16	Chapter 3 History of Artificial Intelligence	Toward the Future
17	Chapter 4 Frontiers of Artificial Intelligence	AI and Board Games
18	Chapter 4 Frontiers of Artificial Intelligence	AI and Language
19	Chapter 4 Frontiers of Artificial Intelligence	AI and Art
20	Chapter 4 Frontiers of Artificial Intelligence	AI and Astronomy
21	Chapter 4 Frontiers of Artificial Intelligence	AI and Biology
22	Chapter 4 Frontiers of Artificial Intelligence	AI and Medicine
23	Chapter 5 AI Ethics	The Three Laws of Robotics
24	Chapter 5 AI Ethics	Information Fabrication
25	Chapter 5 AI Ethics	Information Leakage
26	Chapter 5 AI Ethics	Filter Bubbles
27	Chapter 5 AI Ethics	Artificial Intelligence and Social Fairness
28	Chapter 5 AI Ethics	Legal Responsibility

Continued on next page

Table B.1 (continued)

No.	Chapter	Lesson
29	Chapter 6 Foundational AI Methods	Knowledge-Based Intelligence
30	Chapter 6 Foundational AI Methods	Learning-Based Intelligence
31	Chapter 6 Foundational AI Methods	Supervised and Unsupervised Learning
32	Chapter 6 Foundational AI Methods	Reinforcement Learning
33	Chapter 6 Foundational AI Methods	Traditions in Machine Learning
34	Chapter 7 Deep Learning Methods	The Human Nervous System
35	Chapter 7 Deep Learning Methods	The Beginning of Artificial Neural Networks
36	Chapter 7 Deep Learning Methods	History of Artificial Neural Networks
37	Chapter 7 Deep Learning Methods	The Beginning of Deep Learning
38	Chapter 7 Deep Learning Methods	Basic Principles of Deep Learning
39	Chapter 7 Deep Learning Methods	Deep-Learning Challenge: Adversarial Examples
40	Chapter 7 Deep Learning Methods	Deep-Learning Challenge: Explainability

B.1.2 Practice Manual Contents

Tsinghua High School Edition

This appendix is based on the lower-secondary practice manual in the series, Tsinghua High School Edition.

Table B.2: Contents of the lower-secondary practice manual, Tsinghua High School Edition

No.	Activity Type	Activity
1	Analysis	Origins of Artificial Intelligence (1): The Mathematization of Human Thought
2	Analysis	Origins of Artificial Intelligence (2): The Birth of the Computer
3	Tool	Origins of Artificial Intelligence (3): The Great Alan Turing
4	Analysis	Origins of Artificial Intelligence (4): The Dartmouth Conference
5	Analysis	The Development of Artificial Intelligence (5): Dreams and Disillusionment
6	Tool	The Development of Artificial Intelligence (7): The New Era of Large Models
7	Survey, Debate	The Development of Artificial Intelligence (8): Interdisciplinary Convergence
8	Tool	Creating with AI and Art
9	Survey, Analysis	AI and Biology
10	Survey, Analysis	AI and Medicine
11	Analysis, Debate	The Three Laws of Robotics
12	Coding	Machine-Learning Examples
13	Analysis	The Human Nervous System
14	Analysis	The Beginning of Artificial Neural Networks
15	Analysis	History of Artificial Neural Networks
16	Analysis	Introduction to Convolutional Neural Networks
17	Analysis, Coding	Understanding Deep Learning: The Secrets of AlphaGo
18	Tool	Understanding Deep Learning: The Sora Video Generator
19	Analysis, Tool	Filter Bubbles and Algorithmic Recommendation
20	Debate, Tool	Legal and Ethical Responsibility for Artificial Intelligence

Suzhou Xingpu Experimental Middle School Edition

This appendix is based on the lower-secondary practice manual in the series, Suzhou Xingpu Experimental Middle School Edition.

Table B.3: Contents of the lower-secondary practice manual, Suzhou Xingpu Experimental Middle School Edition

No.	Part/Chapter	Activity Type	Activity
1.1	Part I	Survey	Presenting Artificial Intelligence at a Convention
1.2	Part I	Presentation	What If the Eyes of a Machine Become Biased?
1.3	Part I	Tool	Experiencing Speech Synthesis
2.1	Part II	Analysis	Experiencing the Development of Human Thought
2.2	Part II	Test	Can Machines Think Like Humans?
2.3	Part II	Drama	The Dartmouth Conference
3.1	Part III	Essay	The Course of AI Development
3.2	Part III	Tool	An In-Depth Experience of Large Models
3.3	Part III	Essay	Writing a Science-Fiction Story
4.1	Part IV	Coding	Gomoku Based on Alpha–Beta Pruning
4.2	Part IV	Tool	AI Art Generation
4.3	Part IV	Survey	Protein-Structure Analysis
5.1	Part V	Debate	The Autonomous-Vehicle Dilemma
5.2	Part V	Role Play	Simulating Personalized Recommendation
5.3	Part V	Debate	Artificial Intelligence and Social Fairness
6.1	Part VI	Tool	An Intelligent Fruit-Recognition System
6.2	Part VI	Analysis	Three Core Paradigms of Machine Learning
7.1	Part VII	Survey	Exploring Mechanisms of Self-Recovery
7.2	Part VII	Analysis	Visualizing Neural Networks
7.3	Part VII	Coding	Experiencing Deep Learning

B.2 Grade-Level Volumes

This appendix is based on the Grade 7 and Grade 8 textbooks in the series and lists units and sections by grade and semester.

B.2.1 Grade 7, First Semester

No.	Unit	Section
1.1	Concepts of Artificial Intelligence	The Dream of Intelligent Machines
1.2	Concepts of Artificial Intelligence	What Is Artificial Intelligence?
1.3	Concepts of Artificial Intelligence	The Eyes of a Machine
1.4	Concepts of Artificial Intelligence	The Ears of a Machine
1.5	Concepts of Artificial Intelligence	The Voice of a Machine
1.6	Concepts of Artificial Intelligence	The Hands and Feet of a Machine
2.1	The Birth of Artificial Intelligence	Origins of Human Intelligence
2.2	The Birth of Artificial Intelligence	Formalizing the Laws of Human Thought
2.3	The Birth of Artificial Intelligence	The Birth of the Computer
2.4	The Birth of Artificial Intelligence	The Great Alan Turing
2.5	The Birth of Artificial Intelligence	The Dartmouth Conference
3.1	The Development of Artificial Intelligence	Dreams and Disillusionment
3.2	The Development of Artificial Intelligence	The Era of Deep Learning
3.3	The Development of Artificial Intelligence	The Era of Large Models
3.4	The Development of Artificial Intelligence	Interdisciplinary Integration

B.2.2 Grade 7, Second Semester

No.	Unit	Section
1.1	AI Ethics	The Three Laws of Robotics
1.2	AI Ethics	Information Fabrication
1.3	AI Ethics	Information Leakage
1.4	AI Ethics	Filter Bubbles
1.5	AI Ethics	Artificial Intelligence and Social Fairness
1.6	AI Ethics	Legal Responsibility
2.1	Frontiers of Artificial Intelligence	Artificial Intelligence and Games
2.2	Frontiers of Artificial Intelligence	AI and Language
2.3	Frontiers of Artificial Intelligence	AI and Art
2.4	Frontiers of Artificial Intelligence	AI and Astronomy
2.5	Frontiers of Artificial Intelligence	AI and Biology
2.6	Frontiers of Artificial Intelligence	AI and Medicine
2.7	Frontiers of Artificial Intelligence	AI and New Materials
2.8	Frontiers of Artificial Intelligence	AI and Mathematics
2.9	Frontiers of Artificial Intelligence	AI and Urban Transportation

B.2.3 Grade 8, First Semester

No.	Unit	Section
1.1	Foundational AI Methods (1)	History of Computers
1.2	Foundational AI Methods (1)	Foundations of Computer Algorithms

Continued on next page

Table B.6 (continued)

No.	Unit	Section
1.3	Foundational AI Methods (1)	Knowledge-Based Intelligence
1.4	Foundational AI Methods (1)	Learning-Based Intelligence
2.1	Foundational AI Methods (2)	Machine-Learning Examples
2.2	Foundational AI Methods (2)	Supervised and Unsupervised Learning
2.3	Foundational AI Methods (2)	Reinforcement Learning
2.4	Foundational AI Methods (2)	Traditions in Machine Learning
3.1	Neural-Network Methods	The Human Nervous System
3.2	Neural-Network Methods	The Beginning of Artificial Neural Networks
3.3	Neural-Network Methods	History of Artificial Neural Networks
3.4	Neural-Network Methods	Introduction to Convolutional Neural Networks
3.5	Neural-Network Methods	Introduction to Recurrent Neural Networks

B.2.4 Grade 8, Second Semester

No.	Unit	Section
1.1	Deep Learning Methods	The Beginning of Deep Learning
1.2	Deep Learning Methods	Basic Principles of Deep Learning
1.3	Deep Learning Methods	Word Vectors and Object Embeddings
1.4	Deep Learning Methods	Sequence-to-Sequence Models
1.5	Deep Learning Methods	Attention Mechanisms
1.6	Deep Learning Methods	Deep-Learning Challenge: Adversarial Examples
1.7	Deep Learning Methods	Deep-Learning Challenge: Explainability
2.1	Large Models	Neural Language Models
2.2	Large Models	Introduction to Transformers
2.3	Large Models	Pretraining Methods

Continued on next page

Table B.7 (continued)

No.	Unit	Section
2.4	Large Models	Human Alignment
2.5	Large Models	Large Multimodal Models
2.6	Large Models	Toward the Future

Appendix C

Upper-Secondary School

C.1 Integrated Stage Edition

C.1.1 Textbook Contents

This appendix is based on the upper-secondary textbook in the *Tsinghua Artificial Intelligence General Education Series for Primary, Secondary, and University Education.*

No.	Chapter	Lesson
1	Chapter 1 Introduction to Artificial Intelligence	What Is Artificial Intelligence?
2	Chapter 1 Introduction to Artificial Intelligence	Origins of Human Intelligence
3	Chapter 1 Introduction to Artificial Intelligence	Origins of Artificial Intelligence: Mathematical Logic
4	Chapter 1 Introduction to Artificial Intelligence	Origins of Artificial Intelligence: The Birth of the Computer
5	Chapter 1 Introduction to Artificial Intelligence	Turing: Father of Artificial Intelligence
6	Chapter 1 Introduction to Artificial Intelligence	The Beginning of Artificial Intelligence
7	Chapter 1 Introduction to Artificial Intelligence	History of Artificial Intelligence (1)
8	Chapter 1 Introduction to Artificial Intelligence	History of Artificial Intelligence (2)
9	Chapter 1 Introduction to Artificial Intelligence	AI Ethics (1): Immediate Risks from Artificial Intelligence

Continued on next page

Table C.1 (continued)

No.	Chapter	Lesson
10	Chapter 1 Introduction to Artificial Intelligence	AI Ethics (2): Long-Term Risks from Artificial Intelligence
11	Chapter 2 Foundations of Artificial Intelligence	Knowledge-Based Artificial Intelligence
12	Chapter 2 Foundations of Artificial Intelligence	Learning-Based Intelligence
13	Chapter 2 Foundations of Artificial Intelligence	The Basic Machine-Learning Workflow
14	Chapter 2 Foundations of Artificial Intelligence	Machine-Learning Methods
15	Chapter 2 Foundations of Artificial Intelligence	Four Traditions in Machine Learning
16	Chapter 2 Foundations of Artificial Intelligence	Artificial Neural Networks
17	Chapter 2 Foundations of Artificial Intelligence	Representative Network Architectures
18	Chapter 2 Foundations of Artificial Intelligence	Deep Learning
19	Chapter 2 Foundations of Artificial Intelligence	Basic Principles of Large Models (1)
20	Chapter 2 Foundations of Artificial Intelligence	Basic Principles of Large Models (2)
21	Chapter 3 AI Applications	Machine Vision: Facial Recognition
22	Chapter 3 AI Applications	Machine Vision: The Master Painter
23	Chapter 3 AI Applications	Machine Vision: Fabrication and Detection
24	Chapter 3 AI Applications	Machine Audition: Speech Recognition
25	Chapter 3 AI Applications	Machine Audition: Speech Synthesis
26	Chapter 3 AI Applications	Language Understanding: Machine Translation
27	Chapter 3 AI Applications	Human–AI Game Playing: The Secrets Behind AlphaGo
28	Chapter 3 AI Applications	Human–AI Game Playing: The AI Gamer

Continued on next page

Table C.1 (continued)

No.	Chapter	Lesson
29	Chapter 3 AI Applications	Information Retrieval: The Secrets of Search Engines
30	Chapter 3 AI Applications	Information Retrieval: Recommendation Algorithms
31	Chapter 4 Interdisciplinary Artificial Intelligence	Working with Mathematicians
32	Chapter 4 Interdisciplinary Artificial Intelligence	Imitating a Bat's Ears
33	Chapter 4 Interdisciplinary Artificial Intelligence	Solving Protein Structures
34	Chapter 4 Interdisciplinary Artificial Intelligence	Reconstructing Material Microstructures
35	Chapter 4 Interdisciplinary Artificial Intelligence	Predicting Chemical Reactions
36	Chapter 4 Interdisciplinary Artificial Intelligence	The Astronomer's Assistant
37	Chapter 4 Interdisciplinary Artificial Intelligence	The AI Composer
38	Chapter 4 Interdisciplinary Artificial Intelligence	Detecting Anthrax Spores
39	Chapter 4 Interdisciplinary Artificial Intelligence	Developing Cancer Vaccines
40	Chapter 4 Interdisciplinary Artificial Intelligence	Toward the Future

C.1.2 Practice Manual Contents

This appendix is based on the upper-secondary practice manual in the series, Tsinghua High School Daxing Edition.

No.	Part/Chapter	Activity Type	Activity
1.1	Part I Introduction to Artificial Intelligence	Tool	Mapping the History of Artificial Intelligence
1.2	Part I Introduction to Artificial Intelligence	Survey	A Survey of AI Risks
2.1	Part II Foundations of Artificial Intelligence	Analysis	How AI Can Help Address Marine Pollution
2.2	Part II Foundations of Artificial Intelligence	Survey	Machine Learning
2.3	Part II Foundations of Artificial Intelligence	Analysis	Exploring Artificial Neural Networks
2.4	Part II Foundations of Artificial Intelligence	Analysis	How Transformers Work
3.1	Part III AI Applications	Coding	Qingxing Smart Cafeteria
3.2	Part III AI Applications	Coding	Qingxing Smart Parking
3.3	Part III AI Applications	Tool	Designing a Qingxing Magazine Cover
3.4	Part III AI Applications	Coding	Qingxing Intelligent Audio Guide
3.5	Part III AI Applications	Coding	Qingxing Technology Corner: Game-Playing Robot (1)

Continued on next page

Table C.2 (continued)

No.	Part/Chapter	Activity Type	Activity
3.6	Part III AI Applications	Coding	Qingxing Technology Corner: Game-Playing Robot (2)
4.1	Part IV Frontiers of Artificial Intelligence	Tool	Tessellating with AI
4.2	Part IV Frontiers of Artificial Intelligence	Survey	Solving the Mystery of Protein Structure
4.3	Part IV Frontiers of Artificial Intelligence	Tool	Creating the Qingxing School Song

C.2 Grade-Level Volumes

This appendix is based on the Grade 10 and Grade 11 textbooks in the series and lists units and sections by grade and semester.

C.2.1 Grade 10, First Semester

No.	Unit	Section
1.1	Origins of Artificial Intelligence	What Is Artificial Intelligence?
1.2	Origins of Artificial Intelligence	Origins of Human Intelligence
1.3	Origins of Artificial Intelligence	Origins of Artificial Intelligence: Mathematical Logic
1.4	Origins of Artificial Intelligence	Origins of Artificial Intelligence: The Birth of the Computer
1.5	Origins of Artificial Intelligence	Turing: Father of Artificial Intelligence
2.1	The Development of Artificial Intelligence	The Beginning of Artificial Intelligence

Continued on next page

Table C.3 (continued)

No.	Unit	Section
2.2	The Development of Artificial Intelligence	History of Artificial Intelligence (1)
2.3	The Development of Artificial Intelligence	History of Artificial Intelligence (2)
2.4	The Development of Artificial Intelligence	Astonishing Artificial Intelligence
3.1	AI Ethics	AI Ethics: Immediate Risks
3.2	AI Ethics	AI Ethics: Long-Term Risks
3.3	AI Ethics	Compliant Use of AI Tools

C.2.2 Grade 10, Second Semester

No.	Unit	Section
1.1	AI Methods	Knowledge-Based Artificial Intelligence
1.2	AI Methods	Learning-Based Artificial Intelligence
1.3	AI Methods	The Basic Machine-Learning Workflow
1.4	AI Methods	Machine-Learning Methods
2.1	Deep Learning	An Introduction to Artificial Neural Networks
2.2	Deep Learning	Representative Neural Networks
2.3	Deep Learning	Foundations of Deep Learning
2.4	Deep Learning	Frontiers of Deep Learning (1)
2.5	Deep Learning	Frontiers of Deep Learning (2)
3.1	Large Models	Basic Principles of Large Models (1)
3.2	Large Models	Basic Principles of Large Models (2)

C.2.3 Grade 11, First Semester

No.	Unit	Section
1.1	Machine Vision	Facial Recognition
1.2	Machine Vision	Image Generation

Continued on next page

Table C.5 (continued)

No.	Unit	Section
1.3	Machine Vision	Fabrication and Detection
2.1	Machine Audition	Speech Recognition
2.2	Machine Audition	Speaker Recognition
2.3	Machine Audition	Speech Synthesis
3.1	Language Processing	Machine Translation
3.2	Language Processing	Machine Writers
3.3	Language Processing	Machine Poets
4.1	Human–AI Game Playing	AlphaGo
4.2	Human–AI Game Playing	Artificial Intelligence and Games
5.1	Search and Recommendation	Search Engines
5.2	Search and Recommendation	Recommendation Algorithms

C.2.4 Grade 11, Second Semester

No.	Unit	Section
1.1	AI in Mathematics and Engineering	Working with Mathematicians
1.2	AI in Mathematics and Engineering	Imitating a Bat's Ears
1.3	AI in Mathematics and Engineering	The Astronomer's Assistant
2.1	AI in Materials Science and Chemistry	Reconstructing Three-Dimensional Material Microstructures
2.2	AI in Materials Science and Chemistry	Predicting Classes of Chemical Reactions
2.3	AI in Materials Science and Chemistry	Solving Protein Structures
3.1	AI in Biology and Medicine	Detecting Anthrax Bacilli
3.2	AI in Biology and Medicine	Developing Cancer Vaccines

Continued on next page

Table C.6 (continued)

No.	Unit	Section
3.3	AI in Biology and Medicine	Predicting Health Status
4.1	AI in the Humanities and Social Sciences	Deciphering Ancient Scripts
4.2	AI in the Humanities and Social Sciences	AI Composition
4.3	AI in the Humanities and Social Sciences	Toward the Future

Appendix D

Higher Education

D.1 Textbook Contents

This appendix is based on the Tsinghua textbook *Artificial Intelligence General Education: Principles, Methods, Applications, and Interdisciplinary Integration.*

No.	Part	Chapter
1	Part I Introduction to Artificial Intelligence	What Is Artificial Intelligence?
2	Part I Introduction to Artificial Intelligence	Origins of Human Intelligence
3	Part I: Introduction to Artificial Intelligence	Origins of Artificial Intelligence: Mathematical Logic
4	Part I: Introduction to Artificial Intelligence	Origins of Artificial Intelligence: The Birth of the Computer
5	Part I Introduction to Artificial Intelligence	Turing: Father of Artificial Intelligence
6	Part I Introduction to Artificial Intelligence	The Beginning of Artificial Intelligence
7	Part I Introduction to Artificial Intelligence	History of Artificial Intelligence
8	Part I: Introduction to Artificial Intelligence	Astonishing Artificial Intelligence
9	Part I Introduction to Artificial Intelligence	AI Risk and Ethics
10	Part II Foundations of Artificial Intelligence	Knowledge-Based Artificial Intelligence

Continued on next page

Table D.1 (continued)

No.	Part	Chapter
11	Part II Foundations of Artificial Intelligence	Learning-Based Artificial Intelligence
12	Part II Foundations of Artificial Intelligence	The Basic Machine-Learning Workflow
13	Part II Foundations of Artificial Intelligence	Machine-Learning Methods
14	Part II Foundations of Artificial Intelligence	Machine-Learning Strategies
15	Part II Foundations of Artificial Intelligence	An Introduction to Artificial Neural Networks
16	Part II Foundations of Artificial Intelligence	Representative Neural-Network Architectures
17	Part II Foundations of Artificial Intelligence	Foundations of Deep Learning
18	Part II: Foundations of Artificial Intelligence	Frontiers of Deep Learning (1)
19	Part II: Foundations of Artificial Intelligence	Frontiers of Deep Learning (2)
20	Part II: Foundations of Artificial Intelligence	Basic Principles of Large Models (1)
21	Part II: Foundations of Artificial Intelligence	Basic Principles of Large Models (2)
22	Part II Foundations of Artificial Intelligence	Challenges Facing Deep Learning
23	Part III: AI Applications	Machine Vision: Facial Recognition
24	Part III: AI Applications	Machine Vision: License-Plate Recognition
25	Part III: AI Applications	Machine Vision: AI Beautification
26	Part III: AI Applications	Machine Vision: The Master Painter
27	Part III: AI Applications	Machine Vision: Detecting AI-Generated Fakes
28	Part III: AI Applications	Machine Audition: Speech Recognition
29	Part III: AI Applications	Machine Audition: Speaker Recognition
30	Part III: AI Applications	Machine Audition: Speech Synthesis

Continued on next page

Table D.1 (continued)

No.	Part	Chapter
31	Part III: AI Applications	Language Understanding: Writing and Dialogue
32	Part III: AI Applications	Language Processing: The AI Poet
33	Part III: AI Applications	Language Processing: Machine Translation
34	Part III: AI Applications	Human–AI Game Playing: Mastering Go
35	Part III: AI Applications	Human–AI Game Playing: AI in Video Games
36	Part III AI Applications	Robot Vacuum Cleaners
37	Part III AI Applications	Search Engines
38	Part III AI Applications	Recommendation Algorithms
39	Part IV: Interdisciplinary Artificial Intelligence	Solving the Mystery of Protein Structure
40	Part IV: Interdisciplinary Artificial Intelligence	Reconstructing Three-Dimensional Material Microstructures
41	Part IV: Interdisciplinary Artificial Intelligence	Predicting Classes of Chemical Reactions
42	Part IV: Interdisciplinary Artificial Intelligence	Evidence for Biomimicry
43	Part IV: Interdisciplinary Artificial Intelligence	Locating Sound Sources
44	Part IV: Interdisciplinary Artificial Intelligence	Detecting Anthrax
45	Part IV: Interdisciplinary Artificial Intelligence	Space Exploration
46	Part IV: Interdisciplinary Artificial Intelligence	AI Composition
47	Part IV: Interdisciplinary Artificial Intelligence	Working with Mathematicians
48	Part IV: Interdisciplinary Artificial Intelligence	Machines That Dream
49	Part IV: Interdisciplinary Artificial Intelligence	The Astronomer's Assistant
50	Part IV: Interdisciplinary Artificial Intelligence	Predicting COVID-19 Infectivity

Continued on next page

Table D.1 (continued)

No.	Part	Chapter
51	Part IV: Interdisciplinary Artificial Intelligence	Developing Cancer Vaccines
52	Part IV: Interdisciplinary Artificial Intelligence	AI-Enhanced Microscopy
53	Part IV Interdisciplinary Artificial Intelligence	Toward the Future

D.2 Practice Manual Contents

This appendix is based on the practice manual accompanying the Tsinghua edition of *Artificial Intelligence General Education: Principles, Methods, Applications, and Interdisciplinary Integration.*

No.	Part/Chapter	Activity Type	Activity
1.1	Part I	Survey	AI Literacy
1.2	Part I	Debate	Multiple Intelligences vs. General Intelligence
1.3	Part I	Essay	The Turing–Searle Debate
1.4	Part I	Tool	A Professional Science-Communication Agent
1.5	Part I	Essay	Humans and Robots
2.1	Part II	Debate	Is a Silicon-Based Civilization Possible? A Human–AI Debate
2.2	Part II	Coding	Linear Classification Models
2.3	Part II	Analysis	Visualizing Neural Networks
2.4	Part II	Analysis	Visualizing Convolutional Neural Networks
2.5	Part II	Analysis	Understanding Convolutional Neural Networks in Depth
2.6	Part II	Analysis	Understanding Transformers in Depth
2.7	Part II	Tool	Generating a *Nature* Cover
2.8	Part II	Survey	Chain of Thought in Large-Model Technology
2.9	Part II	Survey	Adversarial Examples
3.1	Part III	Coding	A Facial-Recognition System

Continued on next page

Table D.2 (continued)

No.	Part/Chapter	Activity Type	Activity
3.2	Part III	Coding	A License-Plate-Recognition System
3.3	Part III	Tool	Star of Art
3.4	Part III	Coding	Speech Synthesis and Voice Cloning
3.5	Part III	Essay	Large Language Models and Academic Integrity
3.6	Part III	Tool	The AI Poet
3.7	Part III	Coding	Master of Gomoku
3.8	Part III	Essay	The Benefits and Harms of Commercial Ranking in Search Engines
4.1	Part IV	Survey	A Deployment Plan for the GNoME System
4.2	Part IV	Survey	Feasibility Analysis for an AI+Chemistry Project
4.3	Part IV	Tool	The AI Composer
4.4	Part IV	Survey	Forecasting the Development of AI in Medicine
4.5	Part IV	Survey	An Urban AI Development Plan

Appendix E

Lifelong Learning

E.1 Textbook Contents

This appendix is based on the Tsinghua edition of *Artificial Intelligence General Education: Lifelong Learning Edition*.

No.	Chapter (Unit)	Section
1.1	Artificial Intelligence: The Future Is Here	The Dream of Intelligent Machines
1.2	Artificial Intelligence: The Future Is Here	What Is Artificial Intelligence?
1.3	Artificial Intelligence: The Future Is Here	A Brief History of Artificial Intelligence
1.4	Artificial Intelligence: The Future Is Here	The Power of Artificial Intelligence
2.1	Artificial Intelligence Around Us	The Eyes of a Machine
2.2	Artificial Intelligence Around Us	The Ears of a Machine
2.3	Artificial Intelligence Around Us	The Voice of a Machine
2.4	Artificial Intelligence Around Us	The Hands and Feet of a Machine
2.5	Artificial Intelligence Around Us	Search and Recommendation
2.6	Artificial Intelligence Around Us	Language Understanding

Continued on next page

Table E.1 (continued)

No.	Chapter (Unit)	Section
3.1	AI Foundations: Deep Learning	Machine Intelligence and Machine Learning
3.2	AI Foundations: Deep Learning	The Story of Geoffrey Hinton
3.3	AI Foundations: Deep Learning	The Secrets of Deep Learning
3.4	AI Foundations: Deep Learning	Data and Computation
3.5	AI Foundations: Deep Learning	Understanding AlphaGo
4.1	AI Frontiers: Large Models and Agents	Language Models
4.2	AI Frontiers: Large Models and Agents	Large Language Models
4.3	AI Frontiers: Large Models and Agents	Large Vision Models
4.4	AI Frontiers: Large Models and Agents	Large Multimodal Models
4.5	AI Frontiers: Large Models and Agents	Large Reasoning Models
4.6	AI Frontiers: Large Models and Agents	Agents
5.1	AI Ethics	The Three Laws of Robotics
5.2	AI Ethics	Information Fabrication
5.3	AI Ethics	Information Leakage
5.4	AI Ethics	Filter Bubbles
5.5	AI Ethics	Artificial Intelligence and Social Fairness
5.6	AI Ethics	Legal Responsibility
6.1	AI as a Companion in Everyday Life	AI Question Answering: Using AI to Obtain Information
6.2	AI as a Companion in Everyday Life	AI-Assisted Writing

Continued on next page

Table E.1 (continued)

No.	Chapter (Unit)	Section
6.3	AI as a Companion in Everyday Life	AI Poetry: Let AI Help You Express Your Talent
6.4	AI as a Companion in Everyday Life	AI Artist: Greeting Cards, Posters, and Image Editing
6.5	AI as a Companion in Everyday Life	AI Video: Memories, Greetings, and Records
6.6	AI as a Companion in Everyday Life	AI Music: A Song of Your Own
7.1	Looking to the Future	Trends in AI Development
7.2	Looking to the Future	Directions in AI Research
7.3	Looking to the Future	Frontier Developments in AI for Healthy Living

E.2 Practice Manual Contents

This appendix is based on the *Practice Guide to Artificial Intelligence General Education: Lifelong Learning Edition*.

No.	Topic	Section
1.1	An Easy First Step into AI	Using Smartphone Apps
1.2	An Easy First Step into AI	Basic Intelligence in Smartphone Software
2.1	AI Tools	What Are AI Tools?
2.2	AI Tools	Finding AI Tools
2.3	AI Tools	Getting Started with AI Tools
3	Understanding Artificial Intelligence	
4.1	AI Question Answering: Using AI to Obtain Information	Health and Medicine
4.2	AI Question Answering: Using AI to Obtain Information	Shopping and Consumption
4.3	AI Question Answering: Using AI to Obtain Information	Travel and Transportation

Continued on next page

Table E.2 (continued)

No.	Topic	Section
4.4	AI Question Answering: Using AI to Obtain Information	Administrative and Public Services
4.5	AI Question Answering: Using AI to Obtain Information	Household Tasks and Smartphone Use
5	AI-Assisted Writing	
6	AI Poetry: Let AI Help You Express Your Talent	
7	AI Artist: Greeting Cards, Posters, and Image Editing	
8	AI Video: Memories, Greetings, and Records	
9	AI Music: A Song of Your Own	
10	Safety Guide	
11	Frontier Applications and Creation	
12	Creative Portfolio	

References

Adams, C. et al. (2023). "Ethical principles for artificial intelligence in K-12 education". In: *Computers and Education: Artificial Intelligence* 4, p. 100131. DOI: 10.1016/j.caeai.2023.100131.

AI4K12 Initiative (2019). *Five Big Ideas in Artificial Intelligence and K–12 AI Guidelines*. Association for the Advancement of Artificial Intelligence and Computer Science Teachers Association. URL: https://ai4k12.org/.

Alayrac, Jean-Baptiste, Jeff Donahue, Pauline Luc, et al. (2022). "Flamingo: A Visual Language Model for Few-Shot Learning". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 23716–23736.

Arum, Richard and Josipa Roksa (2011). *Academically Adrift: Limited Learning on College Campuses*. Chicago: University of Chicago Press.

Bahdanau, D., K. Cho, and Y Bengio (2015). "Neural machine translation by jointly learning to align and translate". In: *International Conference on Learning Representations*.

Bakshy, E., S. Messing, and L. A Adamic (2015). "Exposure to ideologically diverse news and opinion on Facebook". In: *Science* 348.6239, pp. 1130–1132.

Balkin, Jack M. (2018). "Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation". In: *UC Davis Law Review* 51.3, pp. 1149–1210. URL: https://lawreview.law.ucdavis.edu/archives/51/3/free-speech-algorithmic-society-big-data-private-governance-and-new-school-speech.

Barnett, Susan M. and Stephen J. Ceci (2002). "When and Where Do We Apply What We Learn? A Taxonomy for Far Transfer". In: *Psychological Bulletin* 128.4, pp. 612–637. DOI: 10.1037/0033-2909.128.4.612.

Beijing Municipal Education Commission (2025). *Beijing Action Plan for Advancing AI Education in Primary and Secondary Schools (2025–2027)*. URL: https://jw.beijing.gov.cn/xxgk/2024zcwj/2024qtwj/202503/t20250307_4028227.html (visited on 08/12/2026).

Bengio, Yoshua et al. (2024). "Managing Extreme AI Risks amid Rapid Progress". In: *Science* 384.6698, pp. 842–845. DOI: 10.1126/science.adn0117.

Biesta, Gert (2010). *Good Education in an Age of Measurement: Ethics, Politics, Democracy*. Boulder, CO: Paradigm Publishers.

Biggs, John and Catherine Tang (2011). *Teaching for Quality Learning at University*. 4th ed. Maidenhead: Open University Press.

Bishop, C. M (2006). *Pattern Recognition and Machine Learning*. Springer.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, et al. (2021). *On the Opportunities and Risks of Foundation Models*. arXiv: 2108.07258. URL: https://arxiv.org/abs/2108.07258.

Boole, George (1854). *An Investigation of the Laws of Thought, on Which Are Founded the Mathematical Theories of Logic and Probabilities*. London: Walton and Maberly.

Borrego, Maura and Lynita K. Newswander (2010). "Definitions of Interdisciplinary Research: Toward Graduate-Level Interdisciplinary Learning Outcomes". In: *The Review of Higher Education* 34.1, pp. 61–84. DOI: 10.1353/rhe.2010.0006.

Bransford, John D., Ann L. Brown, and Rodney R. Cocking, eds. (2000). *How People Learn: Brain, Mind, Experience, and School: Expanded Edition*. Washington, DC: National Academies Press. DOI: 10.17226/9853.

Brennan, Karen and Mitchel Resnick (2012). "New Frameworks for Studying and Assessing the Development of Computational Thinking". In: *Proceedings of the 2012 Annual Meeting of the American Educational Research Association*. Vancouver, Canada. URL: https://web.media.mit.edu/~kbrennan/files/Brennan_Resnick_AERA2012_CT.pdf.

Bresnahan, Timothy F. and Manuel Trajtenberg (1995). "General Purpose Technologies: Engines of Growth?" In: *Journal of Econometrics* 65.1, pp. 83–108. DOI: 10.1016/0304-4076(94)01598-t.

Brin, Sergey and Lawrence Page (1998). "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Computer Networks and ISDN Systems* 30.1–7, pp. 107–117. DOI: 10.1016/s0169-7552(98)00110-x.

Brown, Peter F. et al. (1993). "The Mathematics of Statistical Machine Translation: Parameter Estimation". In: *Computational Linguistics* 19.2, pp. 263–311.

Brown, Tom B. et al. (2020). "Language Models Are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 1877–1901.

Bruner, Jerome S (1960). *The Process of Education*. Cambridge, MA: Harvard University Press.

Buolamwini, Joy and Timnit Gebru (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*. Vol. 81. Proceedings of Machine Learning Research, pp. 77–91.

Cadena, Cesar, Luca Carlone, Henry Carrillo, et al. (2016). "Past, Present, and Future of Simultaneous Localization and Mapping: Toward the Robust-Perception Age". In: *IEEE Transactions on Robotics* 32.6, pp. 1309–1332. DOI: 10.1109/tro.2016.2624754.

Chaney, Allison J. B., Brandon M. Stewart, and Barbara E. Engelhardt (2018). "How Algorithmic Confounding in Recommendation Systems Increases Homogeneity and Decreases Utility". In: *Proceedings of the 12th ACM Conference on Recommender Systems*. Association for Computing Machinery, pp. 224–232. DOI: 10.1145/3240323.3240370.

Chi, Michelene T. H., Paul J. Feltovich, and Robert Glaser (1981). "Categorization and Representation of Physics Problems by Experts and Novices". In: *Cognitive Science* 5.2, pp. 121–152. DOI: 10.1207/s15516709cog0502_2.

China et al., Cyberspace Administration of (2023). *Interim Measures for the Management of Generative Artificial Intelligence Services*. URL: https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

Chinn, Clark A., Luke A. Buckland, and Ala Samarapungavan (2011). "Expanding the Dimensions of Epistemic Cognition: Arguments from Philosophy and Psychology". In: *Educational Psychologist* 46.3, pp. 141–167. DOI: 10.1080/00461520.2011.587722.

Columbia College (2026). *Contemporary Civilization*. URL: https://www.college.columbia.edu/core-curriculum/classes/contemporary-civilization (visited on 06/2026).

Covington, Paul, Jay Adams, and Emre Sargin (2016). "Deep Neural Networks for YouTube Recommendations". In: *Proceedings of the 10th ACM Conference on Recommender Systems*, pp. 191–198. DOI: 10.1145/2959100.2959190.

Cyberspace Administration of China, Ministry of Industry and Information Technology, and Ministry of Public Security (2022). *Provisions on the Administration of Deep Synthesis Internet Information Services*. URL: https://www.cac.gov.cn/2022-12/11/c_1672221949354811.htm.

Cyberspace Administration of China, Ministry of Industry and Information Technology, Ministry of Public Security, and National Radio and Television Administration (2025). *Measures for Labeling AI-Generated and Synthetic Content*. URL: https://www.cac.gov.cn/2025-03/14/c_1743654684782215.htm.

Davies, Alex, Petar Veličković, Lars Buesing, et al. (2021). "Advancing Mathematics by Guiding Human Intuition with AI". In: *Nature* 600, pp. 70–74. DOI: 10.1038/s41586-021-04086-x.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*, pp. 4171–4186. DOI: 10.18653/v1/n19-1423.

Dewey, John (1916). *Democracy and Education: An Introduction to the Philosophy of Education*. New York: Macmillan. URL: https://www.gutenberg.org/ebooks/852.

Dijck, José van, Thomas Poell, and Martijn de Waal (2018). *The Platform Society: Public Values in a Connective World*. New York: Oxford University Press. DOI: 10.1093/oso/9780190889760.001.0001.

diSessa, A. A (1993). "Toward an epistemology of physics". In: *Cognition and Instruction* 10.2–3, pp. 105–225.

Du, Shan et al. (2013). "Automatic License Plate Recognition (ALPR): A State-of-the-Art Review". In: *IEEE Transactions on Circuits and Systems for Video Technology* 23.2, pp. 311–325. DOI: 10.1109/tcsvt.2012.2203741.

Durrant-Whyte, Hugh and Tim Bailey (2006). "Simultaneous Localization and Mapping: Part I". In: *IEEE Robotics and Automation Magazine* 13.2, pp. 99–110. DOI: 10.1109/mra.2006.1638022.

Eliasmith, Chris et al. (2012). "A Large-Scale Model of the Functioning Brain". In: *Science* 338.6111, pp. 1202–1205. DOI: 10.1126/science.1225266.

Eppler, Martin J. and Jeanne Mengis (2004). "The Concept of Information Overload: A Review of Literature from Organization Science, Accounting, Marketing, MIS, and Related Disciplines". In: *The Information Society* 20.5, pp. 325–344. DOI: 10.1080/01972240490507974.

Espeland, W. N. and M Sauder (2007). "Rankings and reactivity: How public measures recreate social worlds". In: *American Journal of Sociology* 113.1, pp. 1–40.

European Union (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence*. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

Facione, P. A (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction*. The Delphi Report. Millbrae, CA: California Academic Press.

Feigenbaum, Edward A. (1977). "The Art of Artificial Intelligence: Themes and Case Studies of Knowledge Engineering". In: *Proceedings of the International Joint Conference on Artificial Intelligence*. Vol. 5, pp. 1014–1029.

Flavell, John H. (1979). "Metacognition and Cognitive Monitoring: A New Area of Cognitive-Developmental Inquiry". In: *American Psychologist* 34.10, pp. 906–911. DOI: 10.1037/0003-066x.34.10.906.

Fleming, Stephen M. and Raymond J. Dolan (2012). "The Neural Basis of Metacognitive Ability". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1594, pp. 1338–1349. DOI: 10.1098/rstb.2011.0417.

Flexner, Abraham (1930). *Universities: American, English, German*. New York: Oxford University Press.

Floridi, L. et al. (2018). "AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations". In: *Minds and Machines* 28, pp. 689–707.

Fong, Geoffrey T., David H. Krantz, and Richard E. Nisbett (1986). "The Effects of Statistical Training on Thinking about Everyday Problems". In: *Cognitive Psychology* 18.3, pp. 253–292. DOI: 10.1016/0010-0285(86)90001-0.

Gatys, Leon A., Alexander S. Ecker, and Matthias Bethge (2016). "Image Style Transfer Using Convolutional Neural Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2414–2423. DOI: 10.1109/cvpr.2016.265.

Geirhos, R. et al. (2020). "Shortcut learning in deep neural networks". In: *Nature Machine Intelligence* 2, pp. 665–673.

Gigerenzer, Gerd and Ulrich Hoffrage (1995). "How to Improve Bayesian Reasoning without Instruction: Frequency Formats". In: *Psychological Review* 102.4, pp. 684–704. DOI: 10.1037/0033-295x.102.4.684.

Gillespie, T (2014). "The relevance of algorithms". In: *Media Technologies: Essays on Communication, Materiality, and Society*. Ed. by P. J. Boczkowski T. Gillespie and K. A. Foot. MIT Press, pp. 167–194.

Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, et al. (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Vol. 27, pp. 2672–2680.

Graves, Alex et al. (2006). "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks". In: *Proceedings of the 23rd International Conference on Machine Learning*, pp. 369–376. DOI: 10.1145/1143844.1143891.

Griffiths, T. L. and J. B Tenenbaum (2006). "Optimal predictions in everyday cognition". In: *Psychological Science* 17.9, pp. 767–773.

Guo, Chuan et al. (2017). "On Calibration of Modern Neural Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research, pp. 1321–1330.

Halliday, M. A. K (1978). *Language as Social Semiotic: The Social Interpretation of Language and Meaning*. Edward Arnold.

Hardwig, J (1985). "Epistemic dependence". In: *The Journal of Philosophy* 82.7, pp. 335–349.

Harvard University Committee on the Objectives of a General Education in a Free Society (1945). *General Education in a Free Society*. Cambridge, MA: Harvard University Press.

Hendrycks, D. and T Dietterich (2019). "Benchmarking neural network robustness to common corruptions and perturbations". In: *International Conference on Learning Representations*.

Hestenes, D (1992). "Modeling games in the Newtonian world". In: *American Journal of Physics* 60.8, pp. 732–748.

Hidi, Suzanne and K. Ann Renninger (2006). "The Four-Phase Model of Interest Development". In: *Educational Psychologist* 41.2, pp. 111–127. DOI: 10.1207/s15326985ep4102_4.

Hinton, Geoffrey et al. (2012). "Deep Neural Networks for Acoustic Modeling in Speech Recognition". In: *IEEE Signal Processing Magazine* 29.6, pp. 82–97. DOI: 10.1109/msp.2012.2205597.

Ho, Jonathan, Ajay Jain, and Pieter Abbeel (2020). "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 6840–6851.

Hofer, Barbara K. and Paul R. Pintrich (1997). "The Development of Epistemological Theories: Beliefs about Knowledge and Knowing and Their Relation to Learning". In: *Review of Educational Research* 67.1, pp. 88–140. DOI: 10.3102/00346543067001088.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, et al. (2022). "Training Compute-Optimal Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 30016–30030.

Holland, John H. (1975). *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Huang, C.-Z. A., A. Vaswani, J. Uszkoreit, et al. (2018). *Music Transformer: Generating music with long-term structure*. arXiv: 1809.04281. URL: https://arxiv.org/abs/1809.04281.

Huang, Lei et al. (2025). "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions". In: *ACM Transactions on Information Systems* 43.2, pp. 1–55. DOI: 10.1145/3703155.

Humboldt, Wilhelm von (1810). *On the Internal and External Organization of the Higher Scientific Institutions in Berlin*. English translation by Thomas Dunlap. URL: https://germanhistorydocs.org/en/the-holy-roman-empire-1648-1815/wilhelm-von-

humboldt-s-treatise-quot-on-the-internal-and-external-organization-of-the-higher-scientific-institutions-in-berlin-quot-1810.pdf.

Hutchins, Edwin (1995). *Cognition in the Wild*. Cambridge, MA: MIT Press.

Inhelder, Bärbel and Jean Piaget (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York: Basic Books.

Ji, Ziwei et al. (2023). "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12, pp. 1–38. DOI: 10.1145/3571730.

Jo, YoungJu, SangYun Park, JaeHwang Jung, et al. (2017). "Holographic Deep Learning for Rapid Optical Screening of Anthrax Spores". In: *Science Advances* 3.8, e1700606. DOI: 10.1126/sciadv.1700606.

Jumper, J., R. Evans, A. Pritzel, et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596, pp. 583–589. DOI: 10.1038/s41586-021-03819-2.

Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. (2020). "Scaling Laws for Neural Language Models". In: *arXiv preprint arXiv:2001.08361*. DOI: 10.48550/arxiv.2001.08361.

Kench, Steven and Samuel J. Cooper (2021). "Generating Three-Dimensional Structures from a Two-Dimensional Slice with Generative Adversarial Network-Based Dimensionality Expansion". In: *Nature Machine Intelligence* 3, pp. 299–305. DOI: 10.1038/s42256-021-00322-1.

Kerr, Clark (2001). *The Uses of the University*. 5th ed. Cambridge, MA: Harvard University Press.

Kerrigan, Joshua, Paul La Plante, Saul Kohn, et al. (2019). "Optimizing Sparse RFI Prediction Using Deep Learning". In: *Monthly Notices of the Royal Astronomical Society* 488.2, pp. 2605–2615. DOI: 10.1093/mnras/stz1865.

Kitchin, Rob (2014). "Big Data, New Epistemologies and Paradigm Shifts". In: *Big Data & Society* 1.1. DOI: 10.1177/2053951714528481.

Kliebard, Herbert M. (2004). *The Struggle for the American Curriculum, 1893–1958*. 3rd ed. New York: RoutledgeFalmer.

Knowles, Malcolm S., Elwood F. Holton, and Richard A. Swanson (2015). *The Adult Learner: The Definitive Classic in Adult Education and Human Resource Development*. 8th ed. London: Routledge.

Koh, P. W. et al. (2021). "WILDS: A benchmark of in-the-wild distribution shifts". In: *Proceedings of the 38th International Conference on Machine Learning*.

Koren, Yehuda, Robert Bell, and Chris Volinsky (2009). "Matrix Factorization Techniques for Recommender Systems". In: *Computer* 42.8, pp. 30–37. DOI: 10.1109/mc.2009.263.

Krajcik, Joseph S. and Phyllis C. Blumenfeld (2006). "Project-Based Learning". In: *The Cambridge Handbook of the Learning Sciences*. Ed. by R. Keith Sawyer. Cambridge: Cambridge University Press, pp. 317–334. DOI: 10.1017/CBO9780511816833.020.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems*. Vol. 25, pp. 1097–1105.

Kuhn, Deanna (1999). "A Developmental Model of Critical Thinking". In: *Educational Researcher* 28.2, pp. 16–46. DOI: 10.3102/0013189x028002016.

Kurakin, Alexey, Ian J. Goodfellow, and Samy Bengio (2017). "Adversarial Examples in the Physical World". In: *ICLR Workshop*. arXiv: 1607.02533. URL: https://research.google/pubs/adversarial-examples-in-the-physical-world/.

Lazer, D. M. J. et al. (2018). "The science of fake news". In: *Science* 359.6380, pp. 1094–1096.

Lazer, David, Alex Pentland, Lada Adamic, et al. (2009). "Computational Social Science". In: *Science* 323.5915, pp. 721–723. DOI: 10.1126/science.1167742.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton (2015). "Deep Learning". In: *Nature* 521, pp. 436–444. DOI: 10.1038/nature14539.

Lee, J. D. and K. A See (2004). "Trust in automation: Designing for appropriate reliance". In: *Human Factors* 46.1, pp. 50–80.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, et al. (2020). "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230-Abstract.html.

Lin, Stephanie, Jacob Hilton, and Owain Evans (2022). "TruthfulQA: Measuring How Models Mimic Human Falsehoods". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229.

Long, Duri and Brian Magerko (2020). "What Is AI Literacy? Competencies and Design Considerations". In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, pp. 1–16. DOI: 10.1145/3313831.3376727.

Manovich, Lev (2020). *Cultural Analytics*. Cambridge, MA: MIT Press.

Matthias, Andreas (2004). "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata". In: *Ethics and Information Technology* 6.3, pp. 175–183. DOI: 10.1007/s10676-004-3422-1.

McCarthy, John (2007). *What Is Artificial Intelligence?* URL: http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html.

McCarthy, John et al. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Dartmouth College. URL: http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf.

McCulloch, Warren S. and Walter Pitts (1943). "A Logical Calculus of the Ideas Immanent in Nervous Activity". In: *The Bulletin of Mathematical Biophysics* 5, pp. 115–133. DOI: 10.1007/bf02478259.

Mellers, B. et al. (2015). "The psychology of intelligence analysis: Drivers of prediction accuracy in world politics". In: *Journal of Experimental Psychology: Applied* 21.1, pp. 1–14.

Merriam, Sharan B. and Laura L. Bierema (2013). *Adult Learning: Linking Theory and Practice*. San Francisco: Jossey-Bass.

Miao, Fengchun and Mutlu Cukurova (2024). *AI Competency Framework for Teachers*. Paris: UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391104.

Miao, Fengchun and Wayne Holmes (2023). *Guidance for Generative AI in Education and Research*. Paris: UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000386693 (visited on 07/20/2026).

Miao, Fengchun, Kelly Shiohira, and Natalie Lao (2024). *AI Competency Framework for Students*. Paris: UNESCO. DOI: 10.54675/JKJB9835. URL: https://unesdoc.unesco.org/ark:/48223/pf0000391105.

Ministry of Education of the People's Republic of China (2022). *Information Technology Curriculum Standards for Compulsory Education (2022 Edition)*. Beijing: Beijing Normal University Press.

— (Dec. 2, 2024). *Ministry of Education Announces Measures to Strengthen AI Education in Primary and Secondary Schools*. URL: https://www.moe.gov.cn/jyb_xwfb/gzdt_gzdt/s5987/202412/t20241202_1165500.html (visited on 07/20/2026).

Ministry of Education of the People's Republic of China et al. (Apr. 10, 2026). *AI+Education Action Plan*. URL: https://www.moe.gov.cn/srcsite/A16/s3342/202604/t20260410_1433240.html (visited on 07/20/2026).

Minsky, Marvin and Seymour Papert (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.

Mishra, Punya and Matthew J. Koehler (2006). "Technological Pedagogical Content Knowledge: A Framework for Teacher Knowledge". In: *Teachers College Record* 108.6, pp. 1017–1054. DOI: 10.1111/j.1467-9620.2006.00684.x.

Mitchell, Tom M. (1997). *Machine Learning*. New York: McGraw-Hill. ISBN: 9780070428072.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. (2015). "Human-Level Control through Deep Reinforcement Learning". In: *Nature* 518, pp. 529–533. DOI: 10.1038/nature14236.

Morales-Navarro, Luis et al. (2025). "What Can Youth Learn about Artificial Intelligence and Machine Learning in One Hour? Examining How Hour of Code Activities Address the Five Big Ideas of AI". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39.28, pp. 29195–29202. DOI: 10.1609/aaai.v39i28.35193.

National Institute of Standards and Technology (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. NIST AI 100-1. National Institute of Standards and Technology. DOI: 10.6028/NIST.AI.100-1.

National People's Congress, Standing Committee of the (2021). *Personal Information Protection Law of the People's Republic of China*. URL: https://www.cac.gov.cn/2021-08/20/c_1631050028355286.htm.

Nelson, T. O. and L Narens (1990). "Metamemory: A theoretical framework and new findings". In: *The Psychology of Learning and Motivation*. Ed. by G. H. Bower. Vol. 26. San Diego: Academic Press, pp. 125–173.

Neumann, John von (1945). *First Draft of a Report on the EDVAC*. Tech. rep. Moore School of Electrical Engineering, University of Pennsylvania.

New Generation Artificial Intelligence, National Governance Committee for (2021). *Ethical Norms for New Generation Artificial Intelligence*. URL: https://www.most.gov.cn/kjbgz/202109/t20210926_177063.html.

Newell, Allen and Herbert A. Simon (1976). "Computer Science as Empirical Inquiry: Symbols and Search". In: *Communications of the ACM* 19.3, pp. 113–126. DOI: 10.1145/360018.360022.

Newman, John Henry (1996). *The Idea of a University*. Edited by Frank M. Turner. New Haven: Yale University Press.

Ng, Davy Tsz Kit et al. (2021). "Conceptualizing AI Literacy: An Exploratory Review". In: *Computers and Education: Artificial Intelligence* 2, p. 100041. DOI: 10.1016/j.caeai.2021.100041.

Nisbett, Richard E. et al. (1983). "The Use of Statistical Heuristics in Everyday Inductive Reasoning". In: *Psychological Review* 90.4, pp. 339–363. DOI: 10.1037/0033-295x.90.4.339.

Nussbaum, Martha C (2010). *Not for Profit: Why Democracy Needs the Humanities*. Princeton: Princeton University Press.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. Paris: OECD. URL: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

OECD and European Commission (2026). *Empowering Learners for the Age of AI: An AI Literacy Framework for Primary and Secondary Education*. Paris: OECD Publishing. DOI: 10.1787/65cd27d4-en.

Oord, Aäron van den et al. (2016). *WaveNet: A Generative Model for Raw Audio*. DOI: 10.48550/arxiv.1609.03499. arXiv: 1609.03499. URL: https://arxiv.org/abs/1609.03499.

Ott, P. A., Z. Hu, D. B. Keskin, et al. (2017). "An immunogenic personal neoantigen vaccine for patients with melanoma". In: *Nature* 547, pp. 217–221. DOI: 10.1038/nature22991.

Page, Lawrence et al. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rep. 1999-66. Stanford InfoLab.

Papert, S (1980). *Mindstorms: Children, Computers, and Powerful Ideas*. Basic Books.

Parasuraman, Raja and Victor Riley (1997). "Humans and Automation: Use, Misuse, Disuse, Abuse". In: *Human Factors* 39.2, pp. 230–253. DOI: 10.1518/001872097778543886.

Pascarella, Ernest T. and Patrick T. Terenzini (2005). *How College Affects Students: A Third Decade of Research*. Vol. 2. San Francisco: Jossey-Bass.

Passi, Samir and Solon Barocas (2019). "Problem Formulation and Fairness". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp. 39–48. DOI: 10.1145/3287560.3287567.

Payne, B. H (2019). *An Ethics of Artificial Intelligence Curriculum for Middle School Students*. URL: https://ec.europa.eu/futurium/en/system/files/ged/mit_ai_ethics_education_curriculum.pdf.

Pearl, Judea (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann.

Pennycook, G. and D. G Rand (2019). "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". In: *Cognition* 188, pp. 39–50.

Perdomo, Juan et al. (2020). "Performative Prediction". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings

of Machine Learning Research. PMLR, pp. 7599–7609. URL: https://proceedings.mlr.press/v119/perdomo20a.html.

Perkins, David N. and Gavriel Salomon (1988). "Teaching for Transfer". In: *Educational Leadership* 46.1, pp. 22–32.

Pintrich, P. R (2000). "The role of goal orientation in self-regulated learning". In: *Handbook of Self-Regulation*. Ed. by P. R. Pintrich M. Boekaerts and M. Zeidner. San Diego: Academic Press, pp. 451–502.

Pólya, George (1957). *How to Solve It: A New Aspect of Mathematical Method*. 2nd ed. Princeton University Press.

Pylyshyn, Zenon W. (1980). "Computation and Cognition: Issues in the Foundations of Cognitive Science". In: *Behavioral and Brain Sciences* 3.1, pp. 111–132. DOI: 10.1017/S0140525X00002053.

Rabiner, Lawrence R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286. DOI: 10.1109/5.18626.

Radford, Alec, Jong Wook Kim, Chris Hallacy, et al. (2021). "Learning Transferable Visual Models From Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139, pp. 8748–8763.

Reber, Rolf, Norbert Schwarz, and Piotr Winkielman (2004). "Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?" In: *Personality and Social Psychology Review* 8.4, pp. 364–382. DOI: 10.1207/s15327957pspr0804_3.

Recht, Benjamin et al. (2019). "Do ImageNet Classifiers Generalize to ImageNet?" In: *Proceedings of the 36th International Conference on Machine Learning*. Vol. 97, pp. 5389–5400. URL: https://proceedings.mlr.press/v97/recht19a.html.

Reichstein, Markus et al. (2019). "Deep Learning and Process Understanding for Data-Driven Earth System Science". In: *Nature* 566, pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

Risko, E. F. and S. J Gilbert (2016). "Cognitive offloading". In: *Trends in Cognitive Sciences* 20.9, pp. 676–688. DOI: 10.1016/j.tics.2016.07.002.

Rombach, Robin et al. (2022). "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695. DOI: 10.1109/cvpr52688.2022.01042.

Rosenblatt, Frank (1958). "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain". In: *Psychological Review* 65.6, pp. 386–408. DOI: 10.1037/h0042519.

Rosenblatt, L. M (1978). *The Reader, the Text, the Poem: The Transactional Theory of the Literary Work*. Southern Illinois University Press.

Rudin, Cynthia (2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5, pp. 206–215. DOI: 10.1038/s42256-019-0048-x.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams (1986). "Learning Representations by Back-Propagating Errors". In: *Nature* 323, pp. 533–536. DOI: 10.1038/323533a0.

Russell, Stuart J. and Peter Norvig (2021). *Artificial Intelligence: A Modern Approach*. 4th ed. Pearson.

Sadler, D. R (1989). "Formative assessment and the design of instructional systems". In: *Instructional Science* 18.2, pp. 119–144. DOI: 10.1007/bf00117714.

Saffran, Jenny R., Richard N. Aslin, and Elissa L. Newport (1996). "Statistical Learning by 8-Month-Old Infants". In: *Science* 274.5294, pp. 1926–1928. DOI: 10.1126/science.274.5294.1926.

Sahin, Ugur, Evelyna Derhovanessian, Mark Miller, et al. (2017). "Personalized RNA Mutanome Vaccines Mobilize Poly-Specific Therapeutic Immunity against Cancer". In: *Nature* 547, pp. 222–226. DOI: 10.1038/nature23003.

Samborska, Veronika et al. (2022). "Complementary Task Representations in Hippocampus and Prefrontal Cortex for Generalizing the Structure of Problems". In: *Nature Neuroscience* 25, pp. 1314–1326. DOI: 10.1038/s41593-022-01149-8.

Samuel, Arthur L. (1959). "Some Studies in Machine Learning Using the Game of Checkers". In: *IBM Journal of Research and Development* 3.3, pp. 210–229. DOI: 10.1147/rd.33.0210.

Sandoval, William A., Jeffrey A. Greene, and Ivar Bråten (2016). "Understanding and Promoting Thinking About Knowledge: Origins, Issues, and Future Directions of Research on Epistemic Cognition". In: *Review of Research in Education* 40.1, pp. 457–496. DOI: 10.3102/0091732x16669319.

Schick, Timo, Jane Dwivedi-Yu, Roberto Dessì, et al. (2023). "Toolformer: Language Models Can Teach Themselves to Use Tools". In: *Advances in Neural Information Processing Systems*. Vol. 36, pp. 68539–68551.

Schoenfeld, A. H (1985). *Mathematical Problem Solving*. Academic Press.

— (1992). "Learning to think mathematically: Problem solving, metacognition, and sense making in mathematics". In: *Handbook of Research on Mathematics Teaching and Learning*. Ed. by D. Grouws. Macmillan, pp. 334–370.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823. DOI: 10.1109/cvpr.2015.7298682.

Schwaller, P., T. Laino, et al. (2019). "Molecular Transformer: A model for uncertainty-calibrated chemical reaction prediction". In: *ACS Central Science* 5.9, pp. 1572–1583. DOI: 10.1021/acscentsci.9b00576.

Schwaller, P., D. Probst, A. C. Vaucher, et al. (2021). "Mapping the space of chemical reactions using attention-based neural networks". In: *Nature Machine Intelligence* 3, pp. 144–152. DOI: 10.1038/s42256-020-00284-w.

Seixas, P. and T Morton (2013). *The Big Six Historical Thinking Concepts*. Nelson Education.

Selbst, Andrew D. et al. (2019). "Fairness and Abstraction in Sociotechnical Systems". In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, pp. 59–68. DOI: 10.1145/3287560.3287598.

Shannon, Claude E. (1938). "A Symbolic Analysis of Relay and Switching Circuits". In: *Transactions of the American Institute of Electrical Engineers* 57.12, pp. 713–723. DOI: 10.1109/t-aiee.1938.5057767.

Sharma, Mrinank et al. (2023). *Towards Understanding Sycophancy in Language Models*. DOI: 10.48550/arXiv.2310.13548. arXiv: 2310.13548 [cs.CL]. URL: https://arxiv.org/abs/2310.13548.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, et al. (2018). "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4779–4783. DOI: 10.1109/icassp.2018.8461368.

Shepard, Roger N. (1987). "Toward a Universal Law of Generalization for Psychological Science". In: *Science* 237.4820, pp. 1317–1323. DOI: 10.1126/science.3629243.

Shokri, Reza et al. (2017). "Membership Inference Attacks against Machine Learning Models". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, pp. 3–18. DOI: 10.1109/SP.2017.41.

Shulman, Lee S. (1986). "Those Who Understand: Knowledge Growth in Teaching". In: *Educational Researcher* 15.2, pp. 4–14. DOI: 10.3102/0013189x015002004.

Shute, V. J (2008). "Focus on formative feedback". In: *Review of Educational Research* 78.1, pp. 153–189. DOI: 10.3102/0034654307313795.

Silver, David, Aja Huang, Chris J. Maddison, et al. (2016). "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529, pp. 484–489. DOI: 10.1038/nature16961.

Silver, David, Thomas Hubert, Julian Schrittwieser, et al. (2018). "A General Reinforcement Learning Algorithm that Masters Chess, Shogi, and Go through Self-Play". In: *Science* 362.6419, pp. 1140–1144. DOI: 10.1126/science.aar6404.

Simon, H. A. (1971). "Designing organizations for an information-rich world". In: *Computers, Communications, and the Public Interest*. Ed. by M. Greenberger. Johns Hopkins Press, pp. 37–72.

Skitka, Linda J., Kathleen L. Mosier, and Mark Burdick (1999). "Does Automation Bias Decision-Making?" In: *International Journal of Human-Computer Studies* 51.5, pp. 991–1006. DOI: 10.1006/ijhc.1999.0252.

Snyder, David et al. (2018). "X-Vectors: Robust DNN Embeddings for Speaker Recognition". In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5329–5333. DOI: 10.1109/icassp.2018.8461375.

Sparrow, B., J. Liu, and D. M Wegner (2011). "Google effects on memory: Cognitive consequences of having information at our fingertips". In: *Science* 333.6043, pp. 776–778. DOI: 10.1126/science.1207745.

Sperber, Dan et al. (2010). "Epistemic Vigilance". In: *Mind & Language* 25.4, pp. 359–393. DOI: 10.1111/j.1468-0017.2010.01394.x.

Strathern, M (1997). "Improving ratings: Audit in the British university system". In: *European Review* 5.3, pp. 305–321.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 3645–3650. DOI: 10.18653/v1/P19-1355. URL: https://aclanthology.org/P19-1355/.

Sutton, Richard S. and Andrew G. Barto (2018). *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, MA: MIT Press.

Taigman, Yaniv et al. (2014). "DeepFace: Closing the Gap to Human-Level Performance in Face Verification". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701–1708. DOI: 10.1109/cvpr.2014.220.

Teaching Steering Committee for Basic Education, Ministry of Education (2025a). *Guidelines for Artificial Intelligence General Education in Primary and Secondary Schools (2025 Edition)*. Teaching Steering Committee for Basic Education, Ministry of Education. URL: https://app.www.gov.cn/govdata/gov/202505/15/528972/article.html.

— (2025b). *Guidelines for the Use of Generative Artificial Intelligence by Primary and Secondary School Students (2025 Edition)*. Teaching Steering Committee for Basic Education, Ministry of Education. URL: https://app.www.gov.cn/govdata/gov/202505/15/528972/article.html.

Tenenbaum, J. B. et al. (2011). "How to grow a mind: Statistics, structure, and abstraction". In: *Science* 331.6022, pp. 1279–1285.

The College, University of Chicago (2026). *History of the Core*. URL: https://college.uchicago.edu/academics/core/history (visited on 06/2026).

Tolosana, Ruben et al. (2020). "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection". In: *Information Fusion* 64, pp. 131–148. DOI: 10.1016/j.inffus.2020.06.014.

Touretzky, David S. et al. (2019). "Envisioning AI for K–12: What Should Every Child Know about AI?" In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33.1, pp. 9795–9799. DOI: 10.1609/aaai.v33i01.33019795.

Trinh, T. H. et al. (2024). "Solving olympiad geometry without human demonstrations". In: *Nature* 625, pp. 476–482. DOI: 10.1038/s41586-023-06747-5.

Turing, Alan M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society*. 2nd ser. 42.1, pp. 230–265. DOI: 10.1112/plms/s2-42.1.230.

— (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460. DOI: 10.1093/mind/lix.236.433.

Turk, Matthew and Alex Pentland (1991). "Face Recognition Using Eigenfaces". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591. DOI: 10.1109/cvpr.1991.139758.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science* 185.4157, pp. 1124–1131. DOI: 10.1126/science.185.4157.1124.

UNESCO (2021). *Recommendation on the Ethics of Artificial Intelligence*. URL: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

UNESCO Institute for Lifelong Learning (2020). *Embracing a Culture of Lifelong Learning: Contribution to the Futures of Education Initiative*. Hamburg: UNESCO Institute for Lifelong Learning. URL: https://unesdoc.unesco.org/ark:/48223/pf0000374112.

UNICEF Innocenti (2025). *Guidance on AI and Children*. UNICEF. URL: https://www.unicef.org/innocenti/reports/policy-guidance-ai-children (visited on 07/20/2026).

Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems*. Vol. 30, pp. 5998–6008.

Veblen, Thorstein (1918). *The Higher Learning in America: A Memorandum on the Conduct of Universities by Business Men*. New York: B. W. Huebsch.

Vinyals, Oriol, Igor Babuschkin, Wojciech M. Czarnecki, et al. (2019). "Grandmaster Level in StarCraft II Using Multi-Agent Reinforcement Learning". In: *Nature* 575, pp. 350–354. DOI: 10.1038/s41586-019-1724-z.

Vosoughi, S., D. Roy, and S Aral (2018). "The spread of true and false news online". In: *Science* 359.6380, pp. 1146–1151. DOI: 10.1126/science.aap9559.

Vygotsky, L. S (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wang, Pengjie et al. (2024). "An Open Dataset for Oracle Bone Character Recognition and Decipherment". In: *Scientific Data* 11, p. 976. DOI: 10.1038/s41597-024-03807-x.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, et al. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems*. Vol. 35, pp. 24824–24837. URL: https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract.html.

Weizenbaum, Joseph (1966). "ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine". In: *Communications of the ACM* 9.1, pp. 36–45. DOI: 10.1145/365153.365168.

Whitehead, Alfred North (1929). *The Aims of Education and Other Essays*. New York: Macmillan.

Wilson, Louis R. (1946). "General Education in a Free Society; Report of the Harvard Committee". In: *College & Research Libraries* 7.2, pp. 186–187. DOI: 10.5860/crl_07_02_186.

Wineburg, S. S (1991). "Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence". In: *Journal of Educational Psychology* 83.1, pp. 73–87.

Wineburg, Sam and Sarah McGrew (2019). "Lateral Reading and the Nature of Expertise: Reading Less and Learning More When Evaluating Digital Information". In: *Teachers College Record* 121.11, pp. 1–40. DOI: 10.1177/016146811912101102.

Wing, J. M (2006). "Computational thinking". In: *Communications of the ACM* 49.3, pp. 33–35. DOI: 10.1145/1118178.1118215.

Wolpert, David H. and William G. Macready (1997). "No Free Lunch Theorems for Optimization". In: *IEEE Transactions on Evolutionary Computation* 1.1, pp. 67–82. DOI: 10.1109/4235.585893.

Yao, S. et al. (2023). "ReAct: Synergizing reasoning and acting in language models". In: *International Conference on Learning Representations*. arXiv: 2210.03629. URL: https://arxiv.org/abs/2210.03629.

Zeng, Daniel Dajun (2013). "From Computational Thinking to AI Thinking". In: *IEEE Intelligent Systems* 28.6, pp. 2–4. DOI: 10.1109/mis.2013.141.

Zhou, Xinyu, Jessica Van Brummelen, and Phoebe Lin (2020). *Designing AI Learning Experiences for K–12: Emerging Works, Future Opportunities and a Design Framework*. arXiv: 2009.10228. URL: https://arxiv.org/abs/2009.10228.

Zimmerman, Barry J. (2002). "Becoming a Self-Regulated Learner: An Overview". In: *Theory Into Practice* 41.2, pp. 64–70. DOI: 10.1207/s15430421tip4102_2.