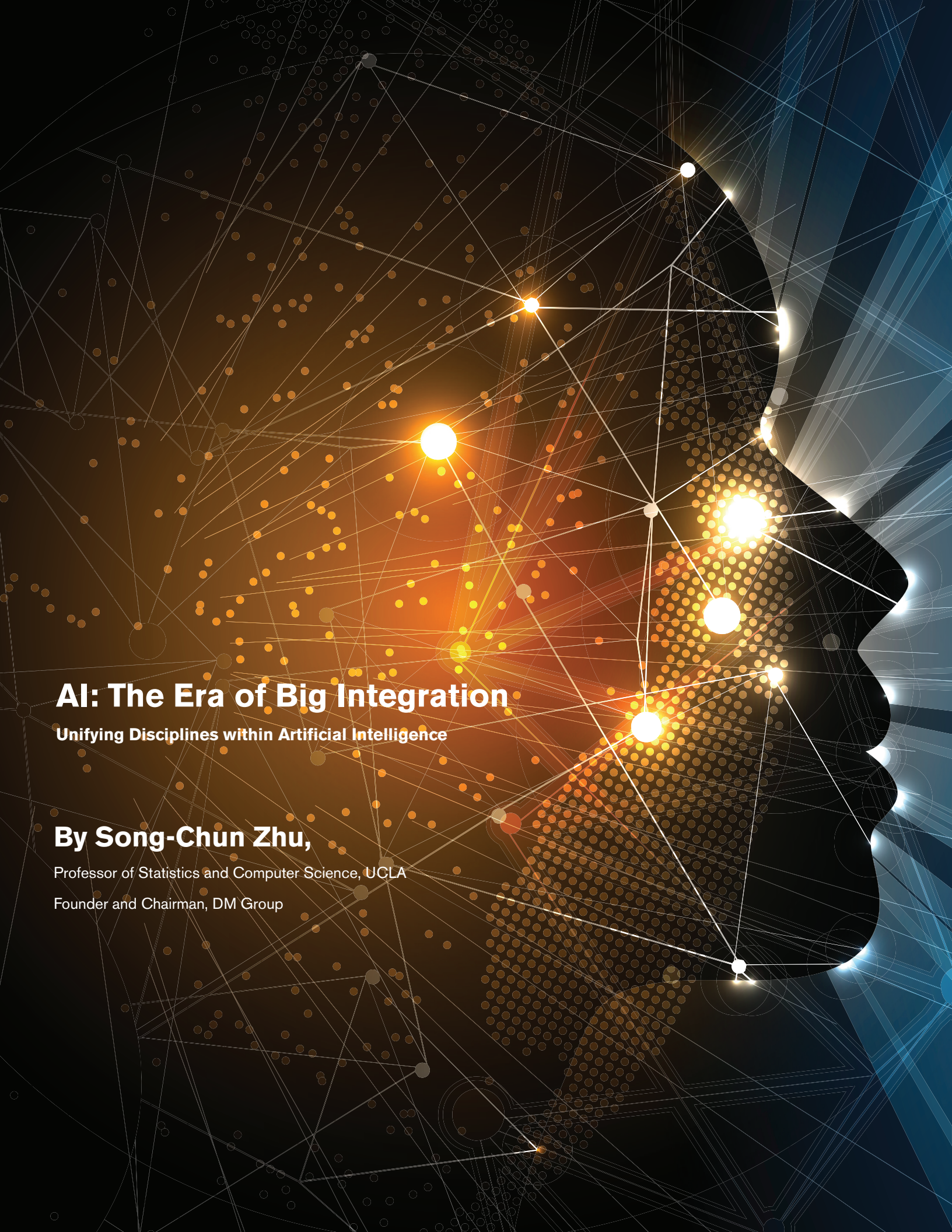


AI: The Era of Big Integration

Unifying Disciplines within Artificial Intelligence

By Song-Chun Zhu,

Professor of Statistics and Computer Science, UCLA

Founder and Chairman, DM Group

Table of Contents

Preface

The promise of artificial intelligence (AI) with human-level cognition has been beyond our grasp for the past sixty years. Over the past decade, advances in AI have created breakthroughs in many of the field’s most historically intractable challenges: speech-to-text, language translation, image and pattern recognition. With such advances, how far can we possibly be from surmounting all remaining obstacles and building machines that not only think like humans but machines that will improve our lives by reducing drudgery and by solving problems great and small?

In this study, Professor Song-Chun Zhu of UCLA offers his perspective on the history, current state, and path forward for next generation cognitive AI. He dispels the myth that human-level AI is already solved; in some regards, he argues, we have barely begun. He also imagines and illustrates a future in which humans and machines collaborate with a kind of mutual “understanding.” Written in plain language, this study seeks to draw AI newcomers into the field, while including enough technical material to keep an AI insider engaged.

In all, Professor Zhu explains a paradigm shift that moves away from big data for small tasks towards small data for big tasks. In the process, Professor Zhu challenges today’s “ABCD” conception of AI:

AI ≠ Big Data + Computing Power + Deep Learning

To appreciate the differences between human and artificial intelligences, let’s, for a moment, compare crows and parrots. Like deep learning networks and other algorithms that use big data, parrots mimic the sounds of the world around them.

But does mimicry, does imitation suggest authentic learning, particularly the mastery of concepts and the application of those concepts to new contexts? It is doubtful. Similar to humans, crows, on the other hand, armed with spatio-temporal-causal reasoning, with a sense of how things work, observe the world with singular intent.

In the case of human intelligence, we are capable of imagining the thoughts of others. This capacity gives us the power to reason not only about space, time, and the physics of cause and effect, but also about the intent and values of those around us. Such social reasoning is the basis of communication and the ultimate prize of language.

Human-level AI must be built with all these capabilities.

Raised in rural China, Song-Chun Zhu completed his Ph.D. at Harvard and Brown with Professor David Mumford, a Fields Medalist. Zhu’s unique Sino-American perspective on AI covers as much range culturally as it does scientifically and informs the body of work he has built in AI over twenty-five years, including the last fifteen at UCLA.

Zhu makes a case for integrating the disparate disciplines that comprise the AI research fields and sets a course that may yet bring AI to the next level on the ladder to true human-level cognition.

If you believe human-level AI is already here because your mobile phone answers when you speak, this study will clarify just how far we have to go.

If you believe human-level AI is impossible, this work may just change your mind.

Bio-sketch of Professor Song-Chun Zhu

Professor, jointly
Statistics and Computer Science
UCLA

Degrees

- 1996 Ph.D., Harvard University, Cambridge, MA
- 1994 M.S., Harvard University, Cambridge, MA
- 1991 B.S., University of Science and Technology of China, at Hefei, China

Appointments

- 2006 Professor, University of California at Los Angeles, Depts. of Statistics, Computer Science
- 2002 Associate Professor, University of California at Los Angeles, Depts. of Statistics, Computer Science
- 1998 Assistant Professor, Ohio State University, Depts. of Computer Science, Cognitive Science
- 1997 Lecturer, Stanford University, Dept. of Computer Science
- 1996 Postdoctoral Fellow, Brown University, Division of Applied Math.

Academic Honors

- 2017 Computational Modeling Prize, Cognitive Science Society, with Tianmin Shu et al.
- 2013 Helmholtz Test-of-Time Award*, 14th Int'l Conf. on Computer Vision at Sydney, Australia, for a region competition paper published in 1995.
- 2011 Fellow, IEEE Computer Society
- 2008 Aggarwal Prize**, the Int'l Association of Pattern Recognition. Citation
- 2007 Marr Prize honorary nomination, 11th Int'l Conf. on Computer Vision at Rio, Brazil, for Object Modeling with Y. Wu et al.
- 2006 Changjiang Scholar, Ministry of Education, China.
- 2003 Marr Prize, 9th Int'l Conf. on Computer Vision at Nice, France, for MCMC Inference for Image Parsing with Z. Tu et al.
- 2001 Young Investigator Award, Office of Navy Research.
- 2001 Sloan Fellow, Alfred P. Sloan Foundation.
- 2001 Career Award, National Science Foundation.
- 1999 Marr Prize honorary nomination***, 7th Int'l Conf. on Computer Vision at Corfu, Greece, for Texture Modeling with Y. Wu.
- 1995 Jury Prize, Harvard University.
- 1992 Harvard Fellowship, Harvard Graduate School of Art and Sciences.
- 1986-1991 Numerous undergraduate student awards in China

Professional Activities

- Int'l Association of Pattern Recognition, committee for the J.K.Aggarwal Prize (2008--2013) , Chair (2011-13).
- IEEE Computer Society, Fellow committee, Vice Chair, 2013.
- General Chair, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Providence, RI, 2012. Editing Board, International Journal of Computer Vision (2004 --)
- Editing Board, Foundations and Trends in Computer Graphics and Vision (2004 --)
- Editing Board, IEEE Transactions on Pattern Analysis and Machine Intelligence (2005--2009)
- General Chair, IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, 2019.
- Principal Investigator, MURI program on Scene Understanding 2010-2015.
- Principal Investigator, MURI program on Visual Commonsense Reasoning, 2016-2021

*This new award started in ICCV 2009. It is for a paper published more than ten years ago. This was Prof. Zhu's first ICCV paper.

** The J.K. Aggarwal Prize was first awarded in 2006 and is awarded biennially by the International Association of Pattern Recognition for one person under the age of 40.

***The Marr Prize used to be the highest honor in computer vision as it was the only award given biennially during the 1980s, 1990s, and early twenty-first century. In recent years, the community has created a few other categories. ECCV and CVPR also introduced their Best Paper prizes.

Introduction

After nearly thirty years of relative obscurity, the term “artificial intelligence” has made a comeback. What is AI, and how will it develop from here? These are questions of widespread interest – and of considerable dispute. There is, on the whole, a widespread misrepresentation of AI.

AI ≠ Big Data + Computing Power + Deep Learning

AI covers a wide range of disciplines and technologies, and gaining a comprehensive understanding of the entire breadth of AI research can prove daunting.

As a matter of practicality, AI principally consists of six core disciplines:

1) **Computer Vision** Object and pattern recognition, 3D scene reconstruction, scene understanding, image processing, and activity and behavior analysis.

2) **Cognitive Science** Functionality, intent, causality, theory of mind, and physical and social common sense.

3) **Natural Language Understanding and Communication** Natural language understanding, ontology, situated dialogue, voice recognition, and synthesis.

4) **Game Theory and Morality** Interaction, competition and cooperation of multi-agent systems, utility maximization, game theoretic equilibria, and social norms.

5) **Robotics** Task planning, motion planning, and dynamics and control.

6) **Machine Learning** Statistical modeling, stochastic computing, neural networks & deep learning, inductive and deductive learning, and predictive analytics.

Section 1
The Current State of AI

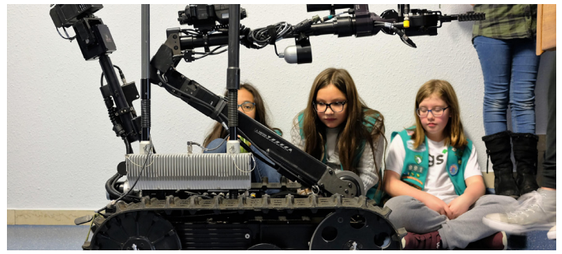
I define “artificial intelligence” as any technology that augments, through harmonious cooperation between humans and machines, human ability in performing tasks that change the physical and social world.

Intelligent computing machines can be virtual or physical robots. They differ from the tools and machines we have built over thousands of years: they can perceive, recognize, reason, make decisions, learn; they can perform tasks, collaborate and communicate within social settings; they can adapt to human preferences, emotions, and even our ethical principles.

Science fiction and fantasies aside, let’s acknowledge recent breakthrough applications of AI. Autonomous vehicles are already within reach. Robots can now be used in disaster relief, providing assistance in dangerous environments. iRobot PackBots, for example, are deployed in war zones to clear explosive devices and to detect biological, chemical, and radioactive threats. Intelligent prosthetic devices connect the human brain and body, restoring some control to people who have lost physical capabilities. Assistance robots are appearing in elder- and patient-care settings.

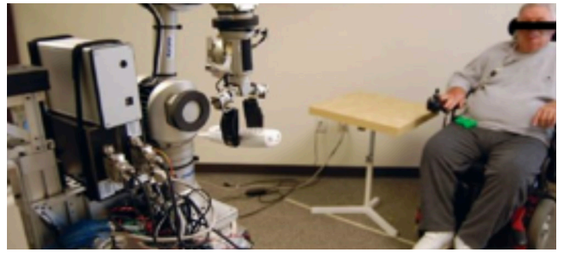
The vision for intelligent systems has often outpaced reality itself. Many public exhibitions of AI have been exposed as “hard-coded” robots that can only work in a specific setting, or only when operated directly by humans – controlled by “the man behind the curtain.”

Patrol

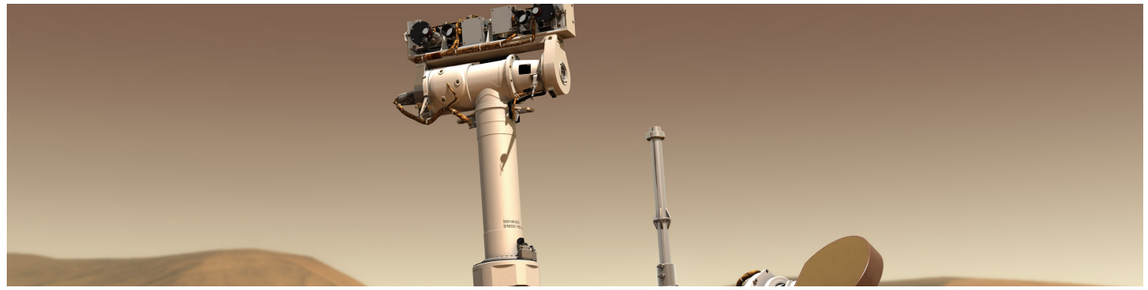


Disaster Relief

Smart Prosthetics



Assistive Robots



Level Setting

Real Progress Exposed

The Fukushima Daiichi nuclear disaster of March 2011 illustrated some of the limits inherent in robotic technologies. Robots designed for deployment into disaster relief areas encountered unexpected challenges. In one particularly memorable case, a robot entering the disaster area dragged a cable along with it, subsequently becoming entangled within the cable.

Upon observation, a robotics expert joked that with our current technology, a robot would need two miniature nuclear power plants, one for the computer and motors and the other for the cooling system, to escape this predicament.

This claim may seem like an exaggeration to some given the astonishing robotic performances we see on the internet. In video demonstrations from Boston Dynamics, pictured at right, robots rarely tip over when kicked, and they can even upright themselves after a fall. Some of these robots have been designed to move fast like animals; some models look like robotic donkeys or giant dogs, and they are built to carry heavy loads.

Boston Dynamics was initially tasked and supported by the US Department of Defense to develop a fully-featured robot. After its acquisition by Google, the company discontinued their defense projects; soon, however, Google sold the division to SoftBank.

The Hype

The media has been hyping AI for the last decade. But as we look around, do we see robots walking down the street?

No.

Do we see AI working seamlessly in our homes?

Not yet.

Our only direct experience with AI may be chatbots - text-based question-answering systems trained on big data with deep learning, and their voice-operated cousins: Alexa, Siri, Cortana, Bixby, and Google Assistant. While improving all the time, these bots are still unimpressive, as they lack the cognitive capabilities for communication with humans.

Level Setting

A Telling Observation

in 2015, to test the status quo of robotics technology, the US Department of Defense Advanced Research Project Agency (DARPA) sponsored the final DARPA Robotics Challenge (DRC). The first prize was $2 million. Many teams competed to develop the best robotic system for assisting humans in response to a simulated natural disaster.

The challenge took place across three arenas, each one replicating, with Hollywood ingenuity, a disaster relief scene. Within these scenes, each robot attempted the following tasks:

1. Drive a car to the disaster site.
2. Step out of the car.
3. Open a door and enter a building.
4. Acquire tools.
5. Locate and close the offending valve.
6. Use a tool to break a hole in a concrete wall.
7. Pass through a barrier made of bricks.
8. Walk up an industrial ladder.

The winning team (pictured above) was from the Korea Advanced Institute of Science and Technology. On the right is their robot providing “disaster relief” by opening a door.

As I sat in the audience watching, the robots appeared to be effective. I was shocked and impressed. But later I learned that all the robots’ actions were teleoperated. Every step and scene had an interface and was being controlled by a few students. Perception, autonomous action, and decision-making were all performed by humans. The robots themselves had little perception, cognitive reasoning, or planning ability.

When one robot tried to grasp the door handle, an error as small as one centimeter would prevent success. A small misstep on the stairs would cause the robot to lose balance. Because the off-stage human handlers had no “balance” signal, they could not react quickly enough to keep their robots upright. Think about it: we avoid falling if tripped because we respond as an integrated body, but the students watching from a distance were too far removed from the robots’ experience to respond in time.

And the robots staggered on.

Consider This

The whole scenario and corresponding tasks were programmed in advance, enabling the competing teams to drill and practice repeatedly. What if an unexpected decision becomes necessary in the middle of a crisis situation?

Also, the scenario included no other actors or robots. Adding just one – and the need for social activities such as language communication and division of labor – would more than double the situation’s complexity.

Level Setting

So How Far Along Are We?

The results of the DRC competition put a temporary damper on funding for major US robotics projects. To this day, there remains a gap between what people believe AI can achieve and what AI actually does.

The difficulties inherent in accounting for unexpected variables highlight a central challenge facing AI and robotics research: how do we capture and apply physical and social common sense?

But first, what is common sense? It is the basic knowledge that we rely upon every day to live in society and that we use to acquire further knowledge. Common sense allows us to see one example and to extend a principle from that example to a variety of contexts. At UCLA's Center for Vision, Cognition, Learning, and Autonomy (VCLA@UCLA), I have been leading an interdisciplinary team since 2010 to tackle the acquisition and application of reasoning by visual understanding.

So how far are we from building applicable and human-level AI?

In a word, we are close to our goal. The key is to identify the right problems and the right direction of solutions. As luck would have it, nature has provided us with an example that shows us where we need to go.

Above is a demonstration of a collaboration between my lab and IAI (www.i-a-i.com), in which a robot could be used for complex tasks, such as, unzipping and checking the inside of backpacks, determining bags hold any suspicious items, and defusing potential bombs using pliers.

If completely controlled by humans, surgical robots can perform surgery and save lives. Mechanical control within robotics has advanced quite well and is useful for some tasks. But even remote-controlled robots may not be practical; today's robots can walk on mountainous terrain, but the rumbling of motors would likely expose it as a target on the battlefield.

How can it be useful performing sentry or reconnaissance duties if it makes this much noise?

Section 2

Goals: A Revelation from a Crow

Compare a crow and a parrot, birds of similar size.

Parrots have a strong language imitation ability; they learn to mimic words by hearing them repeatedly. Repeat words and short sentences to a parrot enough times, and the bird will start talking back. This behavior resembles the data-driven chatbots of today; both can speak, but neither can understand semantics nor the context of speech. They cannot utter words that respond to the physical world, or to social objects, scenes, and characters. They cannot speak to logical cause and effect.

In contrast, consider the common crow. Crows can create tools, develop new physical skills, and engage in social activities that require common sense. In other words, crows are far more clever than parrots.

If you were to watch any number of YouTube videos of crows interacting with humans in urban environments, you would see why I believe that any AI research group would be wise to have a crow as its mascot. This is an animal that can teach us about the very nature of intelligence.

Observing the Crow

The following figures are taken from the study conducted by Nihei Yoshiaki (Tohoku University) and Higuchi Hiroyoshi (The University of Tokyo) in the publication, Tohoku Psychologica Folia. In Japan, carrion crows have been observed to use cars to crack nuts since the 1990s.

Figure (A) is a crow in the wild, untrained by man. It must rely on its faculties of observation, perception, cognition, learning, reasoning, and execution to make its way in life. If it were a robot, its explicit day-to-day goal would be to survive. In the city, this means managing the uncompromising urban environment.

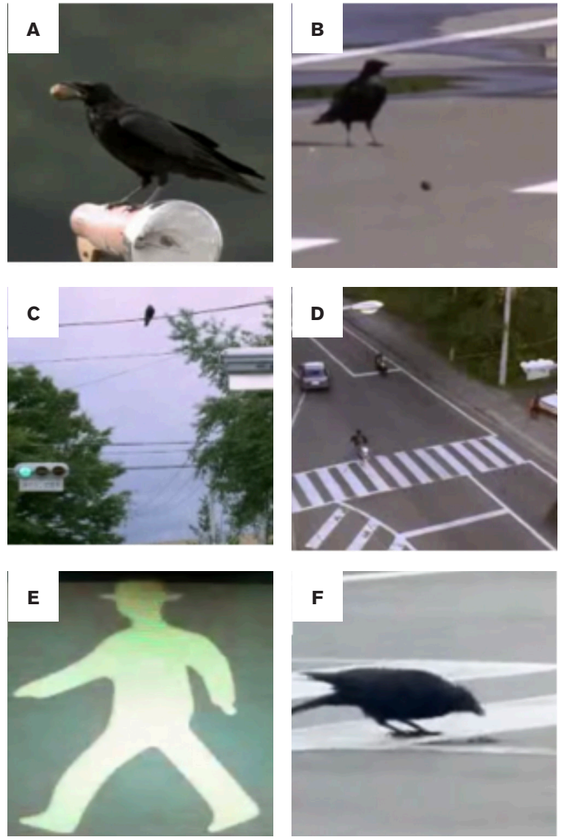
Our crow's first task is to acquire food. Luckily, it finds a nut; but to eat it, it must crack its shell, a task beyond the crow's physical ability. Other animals use tools: to crack a nut, a gorilla will knock it against a hard surface, such as a stone. Our crow tries to crack the nut by dropping it from elevation to the ground, but this strategy does not work, despite repeated attempts. But then a new idea emerges.

Figure (B) shows the crow using a new trick: it puts the nut in the middle of a road, where passing cars crush its shell – what we might call a "bird-machine interaction." But there's a new challenge: how to retrieve the cracked nut from the middle of a busy thoroughfare? After some gentle trial and error, the crow quickly realizes that attempting to retrieve its prize from traffic could result in its death.

Note how different the process of the crow learning to avoid traffic is from machine learning today. There is no dataset, no supervised learning, and no learning cycle; the crow does not run through the nut retrieval scenario multiple times to see exactly how it can go wrong. The crow has only one chance at life.

Figure (C) shows the crow beginning to observe how cars and people sometimes stop at the intersection near the red and green lights. To fully understand why, it must comprehend the complex causal relations between the traffic lights, crosswalk, pedestrian lights, and cars. It must understand how all of these impact cars' speed and stopping, and how this information is relevant to its goal of retrieving its cracked nut.

It must even understand, in some small way, how which lights in which direction will influence which object.



(Figure D) shows the crow choosing to observe from a wire just above a crosswalk at a different location from where it observed and learned about the relationship between all the variables of the intersection. It generalizes and transfers the causal relationships between these variables from one location to another. Current machine learning models cannot accomplish this; some reinforcement learning methods may teach robots how to engage with fixed objects, such as building blocks or toys, but the methods learned in this way can be "brittle" -- they fail once an object's position changes slightly.

The crow drops a nut onto the crosswalk and waits for a car to drive over it.

(Figure E) Once the nut is cracked, the pedestrian lights turn on, and the cars come to a stop, the crow can finally…

(Figure F) …fly down and walk leisurely over to eat the cracked nut off the ground while the cars are stopped at the red light.

"Speaking in metaphor, we are looking for a 'crow' mode of intelligence, rather than a 'parrot' mode of intelligence."

The Crow, the Nut, and the Intersection

You may be surprised by the crow's cleverness. But this is the level of intelligence I have come to expect from AI. Below is an example of intelligent behavior exhibited by a crow. The task is to safely crack a nut and eat the kernel inside its shell.

(A) *A common crow is identified by a researcher. It has no prior training.*

(B) *The crow discovers how to open nuts by placing them on a street to be driven over by a car.*

(C) *The crow finds that cars sometimes stop moving at intersections, creating periods of safety in crosswalks.*

(D) *The crow intentionally chooses a wire over the crosswalk to get the best view of the scene.*

(E) *The crow throws the nut on to the crosswalk and waits for a car to drive over it and the pedestrian lights to come on.*

(F) *The crow retrieves the nut safely.*

This metaphor shows that a solution exists for the kind of AI we want to build. At the same time, I certainly do not claim that this simple analogy clarifies every issue about intelligence. Meanwhile, not all aspects of intelligence need to be replicated in machines to be useful; we can appreciate that the parrot mode of intelligence is commercially viable for some vertical applications, and should by no means be discarded.

But the goal of scientific research is, of course, to look beyond immediate commercial value, to look deeper, to work towards advancing what is possible.

Three critical takeaways about crow intelligence:

1) It is entirely autonomous.

It has perception, cognition, reasoning, learning, and execution. We noted earlier that general cognition is a problem that can't be solved by the world's top scientists; the crow proves that a solution exists.

2) It does not need big data

A crow does not have millions of annotated inputs and outputs to analyze. It solves problems through a small amount of data and without any supervision from a teacher. The crow cannot try random solutions in dangerous situations to see what could go wrong; this is the crow's only life.

3) It is efficient.

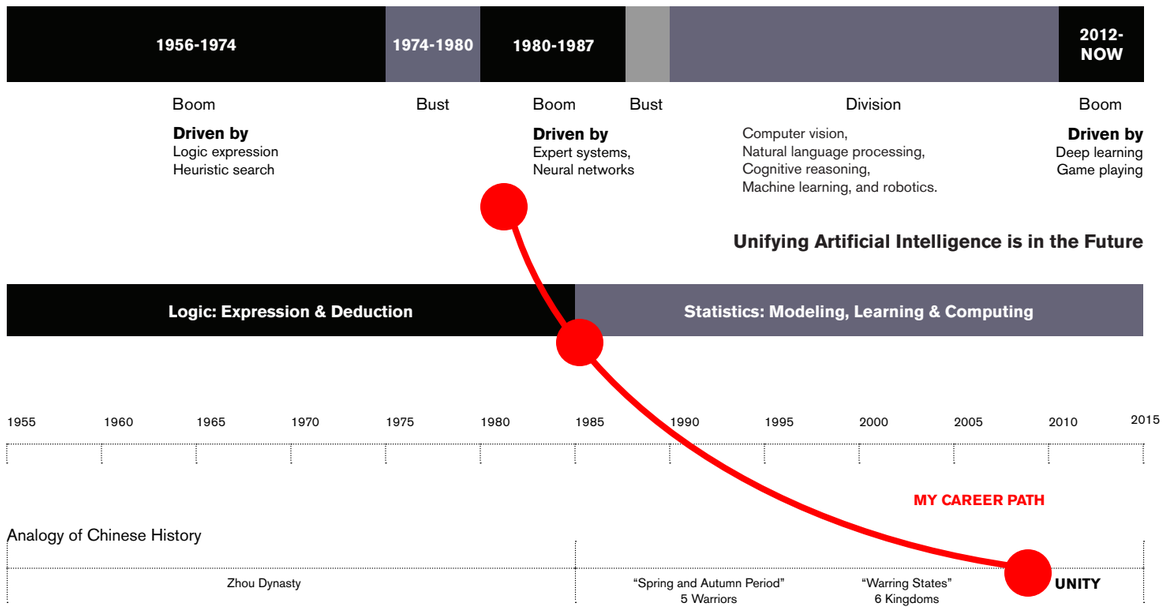
A crow's brain consumes only 0.1-0.2 Watts of power, while the human brain consumes between 10-25 Watts of power. The size efficiency of the human brain shows hardware chip designers what is possible.

Section 3
History: AI Grows and Divides

A Brief History of AI

To understand the future prospects of AI, we must first review its past. In this section, I will offer my perspective on AI's history based upon my own experience. While this view is not necessarily definitive or comprehensive, I hope to distill, as effectively as I can, what's important about the last sixty years of the field's story, and the brilliant people who have worked on it.

As has been described in both technology news and popular media, AI development has undergone several boom-bust cycles over the course of a half-century of advancement. Periods of rapid research synergy and progress were followed by periods where progress substantially slowed. The chart below illustrates how each boom was triggered by a specific technical development.



Booms and Busts

The First Boom | 1956-1974

The first AI boom was from 1956-1974. The technologies of the time included propositional logic, predicate logic, and other forms of knowledge expression, such as heuristic search algorithms. Studies of computer chess began during this time.

Then the first so-called "AI Winter" began, marked by the ALPAC (Automatic Language Processing Advisory Committee) report of 1966. This report concluded that machine translation had failed, after a $20 million effort, due to the "common sense" problem. Meanwhile, other pessimistic reports were published, concluding that nothing practical would emerge from AI research in any reasonable time frame.

The Next Boom | 1980-1987

The second boom began with a group of outspoken professors and researchers in the early 1980s. At the time, people in the field were particularly excited about expert systems, knowledge engineering, and medical diagnosis. Following the US, China sought to develop expert systems for traditional Chinese medicine. Although progress was made, the solutions sometimes lacked sound theoretical foundations.

In 1986, I went to the University of Science and Technology in China to pursue my undergraduate degree in computer science. I was not very interested in the computer itself, since it was such a well-prescribed tool with a well-defined set of accompanying skills required to operate it. But AI was a deep, dark territory ripe for long-term exploration, so I took a graduate-level course in AI.

I was disappointed that, contrary to my expectations, the class covered little more than symbolic reasoning, which seemed out of touch from the reality of intelligence. At that time, the AI community was quite pessimistic, and morale was low. So, I focused instead on the relevant areas of human intelligence: neurophysiology, psychology, cognitive science, and others. This path led me to become aware of the emerging discipline of computer vision.

A brief wave of interest in neural networks arrived in the late 1980s. At the time, my undergraduate program was in its fifth and final year, and my thesis was related to this exciting AI method. Once the wave subsided, AI fell into relative obscurity for nearly thirty years.

The Third Boom | 2012-Now

The third explosion of interest in AI arrived with the rise of deep learning, and continues to expand. At first, most researchers were cautious, pointing towards specific application areas, avoiding undue speculation about general-purpose artificial intelligence. Although Hollywood and some high-profile entrepreneurs and scientists discussed the concept of general AI (AGI, for Artificial General Intelligence), the attitude among researchers remains still cautious. I believe this trend may soon end.

As deep learning enabled breakthroughs in image recognition, speech transcription, and language translation, the term "AI" made a comeback in broader society. The corporate world seized this golden opportunity to generate positive PR; the term "AI" is used liberally, even by companies who do not fully understand it, to spark excitement and convey a sense of being on innovation's cutting edge. Surely, the thinking goes, AI will unlock vast transformations, and no one wants to be seen as missing the train.

Some believe that this boom will not bust – that this time around, winter will not come. But whether winter comes or not depends on how and what we do today.

For nearly thirty years from when I started university, the term "AI" disappeared from the public conversation. But AI did not disappear; it ramified into five distinct disciplines:

- **Computer vision**
- **Cognitive science**
- **Natural language understanding**
- **Robotics**
- **Machine learning**

Each discipline formed its own academic community, with its own international conferences and its own journals to match. Each discipline developed independently; a small community working on game playing and common sense reasoning were all that was left of what was once considered the field of "artificial intelligence."

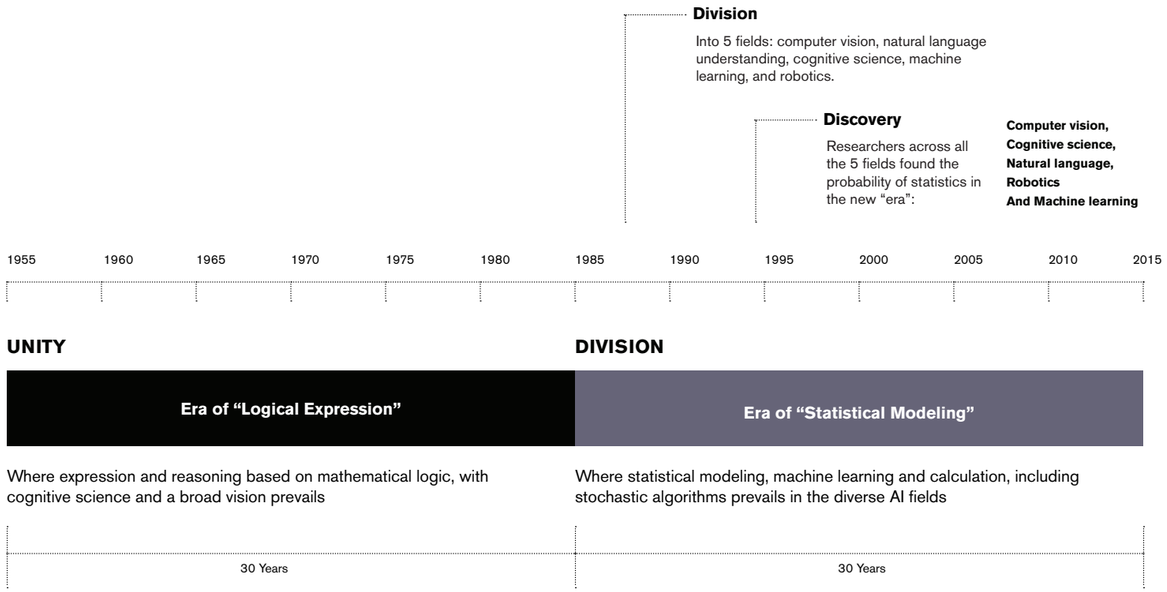
I call these thirty years the "divide-and-conquer period," a time in which the five disciplines of what is now AI grew and developed on their own.

Two Stages in the History of Artificial Intelligence

The First Stage:

The first thirty years of AI, through the first two boom-bust cycles, were dominated by reasoning based on mathematical logic. It was led by outstanding scientists such as John McCarthy, Marvin Minsky, and Herbert Simon. They had a bold vision, and they had mastered some aspects of cognitive science. They were the people I admired when I was in college.

Their creations were clean, beautiful, and certainly worthy of deep study. If you are interested, go find **The Handbook of Knowledge Representation** (2008), weighing in just north of a thousand pages. The study does not describe real-world systems; there are few charts or visual representations to draw from. The constraints of mathematical logic and reasoning limited what the approach taken by these scientists I admired could achieve in the real world.



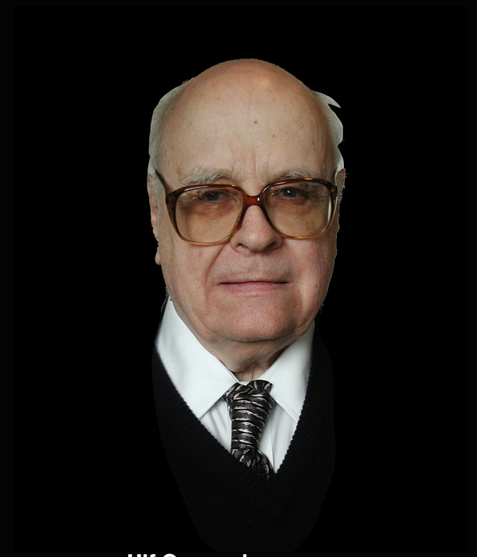
Artificial intelligence research was divided into five areas, without a governing authority, and progress was made independently across all the areas.

The Second Stage

The second stage of artificial intelligence covers the most recent thirty years of boom and bust, when statistical modeling and probability, machine learning, and calculation emerged as the prevailing AI methods. In the mid-1990s, after more than ten years of development, researchers across five AI disciplines began applying probability and statistics, ushering in a new era dominated by machine learning and stochastic algorithms.

Influential Researchers and Pioneers in the Paradigm Shift:

As the field of AI has transformed, several individuals have played central roles as pioneers in the paradigm shift from logic and symbolic reasoning to probabilistic modelling, inference and learning. It's not possible to list every person who has driven AI forward, but there are a few whose contributions cannot be underestimated.



Ulf Grenander
(1923-2016)
L. Herbert Ballou University
Professor Emeritus, Applied
Mathematics,
Brown University

Judea Pearl
(1936-)
Professor
Computer Science Department
Cognitive Systems Lab
UCLA

Leslie Valiant
(1949-)
T. Jefferson Coolidge Professor
Computer Science and Applied
Mathematics,
School of Engineering and
Applied Sciences,
Harvard University.

David Mumford
(1937-)
Professor Emeritus
Applied Mathematics
Brown University
Harvard University

Professor Ulf Grenander

Professor Grenander began his pioneering research on random processes and probability models in the 1960s. When other leading scholars were focused on logic and neural networks, he began working on probabilistic models and stochastic processes for computation. In an effort to build a unified mathematical model for various patterns in nature, he established generalized pattern theory.

Grenander is widely regarded as a trailblazer and thought leader by the academic community. I had the honor of introducing him and recognizing his contribution at the CVPR (Computer Vision & Pattern Recognition) 2012 conference by presenting him with a Pioneer Medal. Since then, the American Mathematical Society (AMS) has awarded the Grenander Prize to scholars who have contributed significantly to statistical modeling and computational fields.

Professor Judea Pearl

Professor Pearl is a colleague of mine at UCLA. While working on heuristic search algorithms in the 1980s, he proposed that Bayesian networks could model cognitive reasoning, through which it would be possible to use estimates of uncertainty in computational reasoning. In the late 1990s, he was once again ahead of his time, as he dove deeper into causal reasoning.

He was the first professor at UCLA to be jointly appointed in computer systems and statistics, followed by myself in late 2002. Pearl won the Turing Award in 2011. Now over 80 years of age, he publishes papers regularly and continues to make revolutionary contributions to interdisciplinary research in his two fields.

Professor Leslie Valiant

Professor Valiant has made enormous contributions to discrete mathematics, computer algorithms, and distributed architectures. In 1984, Valiant published the paper that pioneered computational learning theory. In it, he posed two important questions:

First, how many examples are required to learn a concept with a certain degree of confidence, aka, “probably approximately correct” (PAC) learning?

And second, can one improve classifier performance by combining two weaker classifiers?

These questions are the source of Boosting and AdaBoost, first introduced in an algorithmic implementation designed by a postdoctoral student. Short for “Adaptive Boosting,” AdaBoost is the first practical boosting algorithm and focuses on attempting to strengthen weak classifiers. Valiant received the Turing Award in 2010. In 1992, I took Valiant’s class in my first semester at Harvard. He assigned unsolved questions from his research to students, so no reference answers were available. As the fall semester unfolded, the number of students enrolled in his course dropped from forty to ten. It was that difficult, but for those of us who stayed, despite all our struggles and worries, we each received an “A” for our work. Valiant’s class was a true proving ground.

Professor David Mumford

Professor Mumford became interested in AI, along with the scientific modeling of thought and the brain, in the 1960s. Despite an exceptional career in mathematics, including receiving its top honor, the Fields Medal, he returned to AI in the mid-1980s, beginning with computer vision and computational neuroscience.

From the 1980s to the early 1990s, with geometric invariant theory as an area of significant interest in the computer vision community, Mumford chose instead to focus on probability. Eventually, he shifted his focus to generalized pattern theory.

Geoffrey Hinton
(1947-)
Professor
Department of Computer Science
The University of Toronto

Mr. Ray Kurzweil
(1948-)
American Inventor and Futurist

Professor Geoffrey Hinton

In 1986, Professor Hinton published a seminal paper with David E. Rumelhart and Ronald J. Williams that popularized an algorithm for training multi-layer neural networks called backpropagation. Since then, Hinton has become one of the icons of the deep learning community.

Indeed, Hinton's research on neural networks and deep learning has had a significant impact on the field of AI. A cognitive psychologist and computer scientist, he received the 2018 Turing Award, alongside Yoshua Bengio and Yann LeCun. Hinton currently works at Google Brain.

Mr. Ray Kurzweil

One of our leading innovators within the field of artificial intelligence, Ray Kurzweil has authored seven books to date. Five of these works have gone on to become national bestsellers. In recognition of his inventiveness, Kurzweil has been awarded the National Medal of Technology and Innovation.

As we conduct our research and seek to advance our understanding of the ways in which AI can empower people to live better lives, Kurzweil endures as a leading source of both information and inspiration.

Kurzweil's 2005 book The Singularity is Near has garnered high praise from scientists and philosophers alike, and he continues to make invaluable contributions to the field today.

History & Trend: from Perception to Cognition to Unify All Six Disciplines of AI

Over the course of the 1980s, the discourse around AI disintegrated, and in that disintegration, the notion of AI as a unified discipline dissipated. During this time, the term "artificial intelligence" was associated with little in the way of real-world systems. The study of AI as we understand it today was then divided, as you will recall, into five separate sub-disciplines: computer vision, natural language understanding, cognitive science, robotics, and machine learning.

All five disciplines developed independently of one another, yet through their development, each settled on probability modeling and random computing as optimal strategies. Despite some exploratory efforts at unifying the disciplines, they had little to do with one another during this period. As a result, we had relatively few breakthroughs during this time of independent growth and development.

To more fully round out our understanding of the field, we ought to consider two remaining major disciplines that fall under the umbrella of computer science: game theory and morality. As they are both similar in nature, I prefer to group them to group them together, leaving us with a total of six disciplines of AI.

In my work at UCLA, I usually advise graduate students against focusing on one or another discipline; it's better to explore multiple disciplines as well as the potential synergies among them. Inevitably, those who focus too much on a single area are out-competed by well-resourced private firms, or see their research area encroached upon by outsiders. Being open to possibility while learning and researching remains crucial to making discoveries.

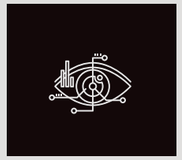
In our current cultural moment, we have entered a period where these erstwhile competitive disciplines are beginning to unify. Computer vision and machine learning were applied in conjunction relatively early. Today, vision and natural language, vision and cognitive reasoning, and vision and robotics have all begun to integrate. In recent years, I have organized several workshops with collaborators from across the six disciplines.

We have come into the time I term Big Integration, a period of large-scale integration that will bring about major shifts and many opportunities.

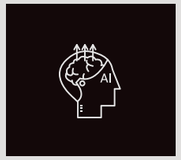
Before the fields can be fully integrated, we must understand each of them on its own terms -- their concepts, advances, semantics, and other core aspects. If we do not understand the terrain within each discipline, we will lack the ability to understand AI as a whole.

In the next part of this study, we will move beyond the historical and present trends in AI as we look towards the future of integration. What is the most effective means of integrating the many disparate ideas and problems across all the AI disciplines?

Towards this end, our efforts will focus upon raising important questions and exploring these ideas and examples for our collective consideration.

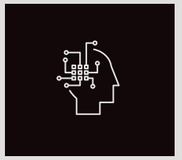


Computer Vision

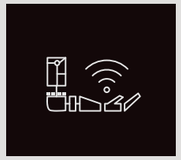


Cognitive Science

Natural Language



Game Theory/Morality

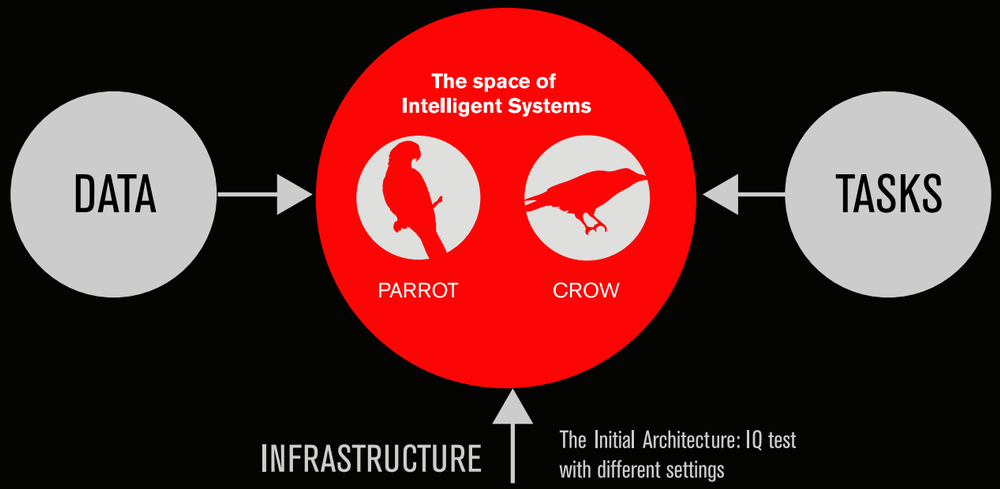


Robotics

Machine Learning

Section 4

Unified: The “Small Data, Big Task” Paradigm and Cognitive Frameworks



Intelligence is a phenomenon that manifests itself in the behavior of individuals and social groups. In the case of crows, intelligence is molded by two basic forces.

1 **The physical environment and chain of causal events**

The boundaries of a crow’s physical existence profoundly shape its intelligence; if it had evolved in different environmental conditions, its intelligent abilities would have developed differently. Any intelligent machine must understand and adapt to its physical world.

2 **The internal values shaped by tasks that must be performed by the species to survive and thrive**

Survival demands that an individual solve the problems of how to acquire food and ensure safety against various environmental threats. In addition, the survival of the species necessitates procreation through the finding of mates, which in many species requires social activity. Many tasks arise from these foundational problems; these tasks drive behavior.

Behavior aimed at completing these tasks is represented in a host of value and decision functions. These functions were formed in the evolutionary process and are modulated by chemical compounds in the brain such as dopamine (pleasure), serotonin (pain), acetylcholine (anxiety, uncertainty), norepinephrine (novelty, excitement) and many others.

Building an Autonomous Agent

Intelligence is the means by which a person or animal achieves physical goals while considering priorities and preferences, which we call “values,” while simultaneously constrained by a chain of causation in a physical environment. So, to construct an intelligent agent, we need to define two basic conditions:

1 **How an agent interacts with its environment through the motion of its body; and**

2 **A model space for its behavior, which must include a value function**

Biological genes provide these two basic conditions for intelligent animals. They give autonomous individuals the ability to understand, use, transform, and find their way through the world.

In our example, a model space is a mathematical concept expressed by the value function, decision function, perception, cognition, task planning, and other critical drivers of behavior. Our conception of a model space is constantly changing with our changing minds; it’s a quantitative expression of an individual’s values and perceptions of the world, and of itself. The complexity of this space determines an individual’s IQ and achievement. I will say more later about the expression of this model and what essential elements are required.

After defining these innate basic conditions, our next task involves the following question: what drives the model’s movement in the space; that is, what is the process of learning?

There are two elements that shape our answers. From the outside comes data. Signals from the world around us, coming from observation and experimentation, are perceived by our brain, shaping our model. Data from observation informs statistical models that can form a joint distribution about space and time. On the other hand, data from experimentation builds causal models by linking causes with effects. It’s important to note that statistics and causality are separate concepts. On the inside is the task, the intrinsic value function driving behaviors to achieve certain goals. Our value function is shaped in the process of evolution. As a result of the variety of tasks we must face, we tend to be sensitive to certain variables, while ignoring others. Such a difference forms various models.

The brain of a robot and the human brain can both be seen as a model. Both the data and the task create a given model.

If Not Big Data, Then What?

Founded on probability and statistics, many of the currently popular deep learning methods are under what I call a paradigm of "big data for small tasks." For certain tasks, such as face or object recognition, training a model requires a massive amount of data, often billions of distinct examples.

A model such as this can be effective, but such a model is simply a parrot, not a crow. It's not possible to use a "parrot" model to perform tasks other than the one for which it was trained. To make matters even worse, we are unable to explain why "parrot" models make the decisions they make. A parrot can say "Hello!" if it hears you answer the phone enough times, but it can't tell you what the word "Hello" means, and it can't say "goodbye," unless it hears you say "goodbye" hundreds of times too. For this reason I generally do not support approaches to building cognitive artificial intelligence through deep learning; while I was one of the first researchers to promote statistical modeling, I recognize fundamental limits with these methods.

Instead, I have, over the years, advocated for an alternative approach to advancing AI: "small data for big tasks." Under this approach, we use a large number of tasks to prompt learning by intelligent systems, as opposed to vast datasets. Philosophically speaking, this notion departs radically from current practice and requires a fundamental shift in how we think about AI.

Of course, we hypothesize that human intelligence has had its models designed and trained by millions of years of evolution. People's perceptions and behaviors are seemingly always task-driven. Isn't this just big data? Couldn't we bring human-level AI into being by replicating this process with just a lot of data?

The Three Stages

Taking all of time into account, we can break a person's biological learning process into three phases in history:

(1) Billions of years of evolution, driven by Darwinian natural selection, yielding phenotypes that have survived over time

(2) The ongoing formation and inheritance of human culture

(3) Decades of individual learning and personal adaptation

AI research typically considers only the third stage.

But challenging questions remain:

How do we define "a large number of tasks" needed by the approach I have advocated for? What is the number of tasks that drive human beings, and what is the task space?

Psychological and cognitive scientists have been unclear in their responses to these questions, and these questions illuminate key challenges in AI's continued development.

Guided by these questions, the following sections will explore the six AI disciplines as we seek to build a common, unified framework for artificial "crow" intelligence. For over two decades, we have been studying these six disciplines at the Center for Vision, Cognition, Learning, and Autonomy at UCLA with this goal in mind.

Section 5

Discipline 1: Computer Vision – From "Deep" To "Dark"

Img 19.a

Vision:

Vision is the most important source of information for the human brain and is the "entrance hall" of AI. It all begins with vision. As a vast and complex discipline, computer vision, with its layers of difficulties, is far from being solved.

Here's an example.

Above is a photograph of my daughter in our kitchen. Considered literally, the image is a two-dimensional matrix of pixels. But we know the image to be more than that, simply by seeing it. Rather than seeing only a matrix of pixels, we perceive a three-dimensional scene. We understand my daughter's actions and behavior, and the more we look at the photograph, the greater becomes our understanding of the scene.

How do we build such vision into a computer?

Here is a list of unsolved research questions critical for machines to see and understand everything in this photograph that we can see:

1
How do geometric common sense reasoning and three-dimensional scene construction work, just from a single photo taken by your smartphone?

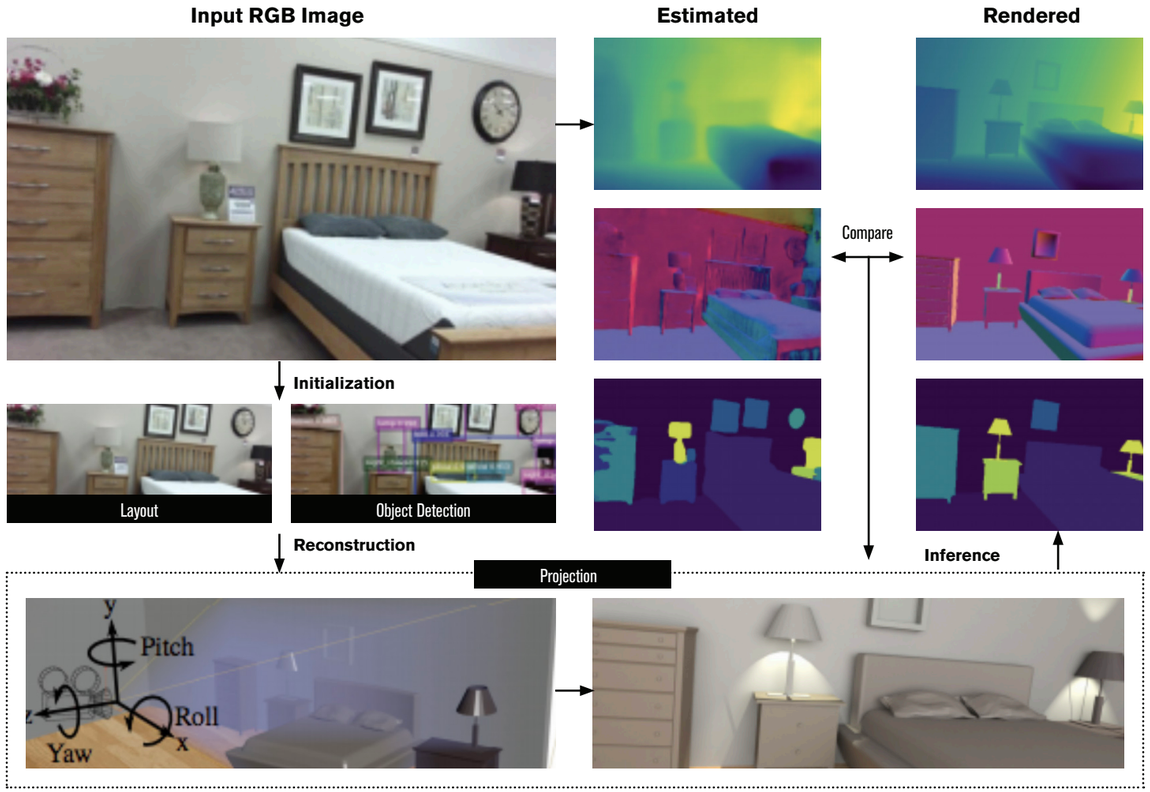
2
How do we replicate humans' functional reasoning in recognizing scenes?

3
How can we replicate the reasoning humans do about physical relations and stability?

4
How can we model intention, attention, and prediction?

5
How can we codify and replicate task-driven causal learning and reasoning?

Some of these questions have largely been ignored by the majority of the recent literature. We will not do that here.



1 Geometric Common Sense Reasoning and Three-Dimensional Scene Construction

Above is an illustration of the 3D indoor scene parsing and reconstruction in an analysis-by-synthesis fashion. A 3D representation is initialized by individual vision tasks (i.e., object detection, 2D layout estimation). A joint inference algorithm compares the differences between the rendered normal, depth, and segmentation map and the ones estimated directly from the input RGB image, and adjusts the 3D structure iteratively.

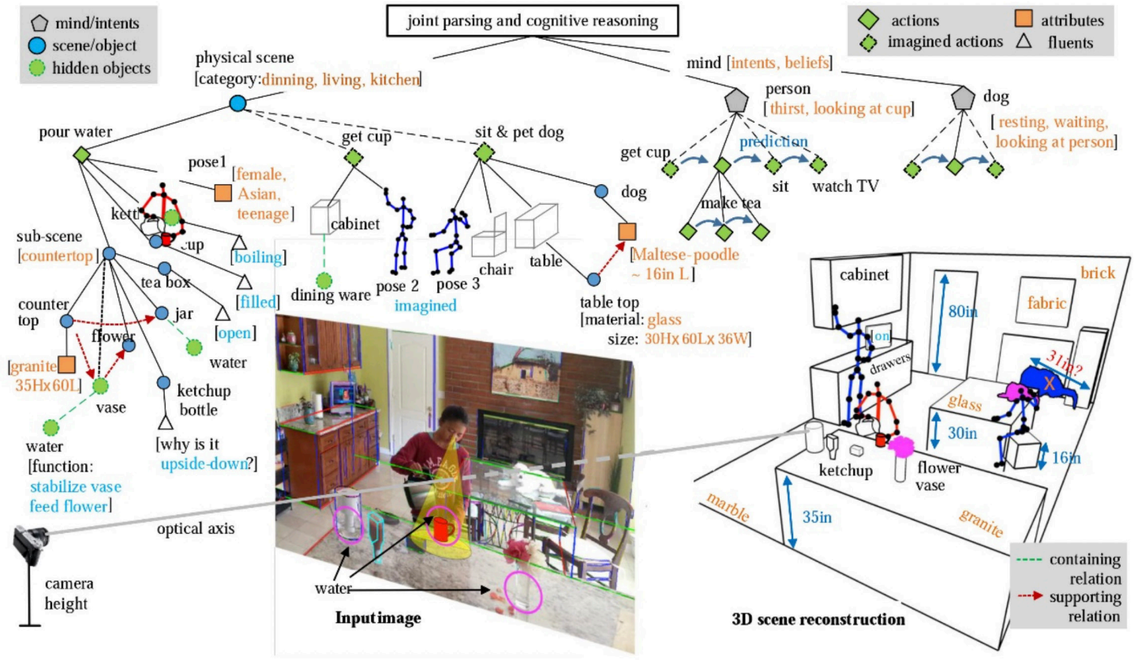
In the first wave of the study of computer vision (1970-1995), one would calculate the positions of points in a three-dimensional coordinate system (Structure from Motion-- SfM, Simultaneous Localization and Mapping -- SLAM) by analyzing multiple images (multi-view). Unlike machines, people only need to look at one image of a scene to estimate three-dimensional geometry.

In 2002, I published a paper with my student Han Feng on a statistical method for three-dimensional reconstruction from a single image. Proponents of the geometric school of thought balked at our idea. At the time, their view was that three-dimensional information could not mathematically be calculated from a single, two-dimensional image.

But in understanding three-dimensional geometry, humans draw upon common sense geometric rules that we don't think about consciously. Consider, for example, that a chair in the photograph is about as high off the ground as the length of my daughter's calf (sixteen inches), while a table is about thirty inches high, a counter top thirty-five, and a door about eighty. We can evaluate these heights by comparing the relative sizes of these objects, using our changing visual perspective as we move through a room.

Meanwhile, much of the world around us is built in common ways we know and use seemingly without thinking. Doors and windows are consistently similar sizes; architectural design and urban planning have standardized much of the physical landscape. Using geometric common sense, we can, taking camera position and optical axis into consideration, convert many points in our two-dimensional field of vision into three-dimensional understanding.

Vision as Joint Spatial, Temporal, and Causal Parsing



Img 21.a

In the above figure, our understanding of the three-dimensional scene is expressed as a hierarchical decomposition of the organization of time and space called a Spatial, Temporal, and Causal Parse Graph, referred to as an STC-PG. The STC-PG is a crucial concept that we will delve deeper into a bit later.

But for now, an interesting feature of our abilities at geometric reconstruction is that we often do not need very keen depth perception. For example, if you wanted to pick up a cup of coffee only a few feet away from you, you would only need a rough estimate of the cup's position to pick it up, and if you were unsuccessful on your first try, you could, with relative ease, focus more keenly on the cup's position on your next try. In other words, our perception of three dimensions can err significantly, and certain degrees of inaccuracy are perfectly acceptable for the performance of many tasks. When executing more difficult tasks, we improve accuracy as needed.

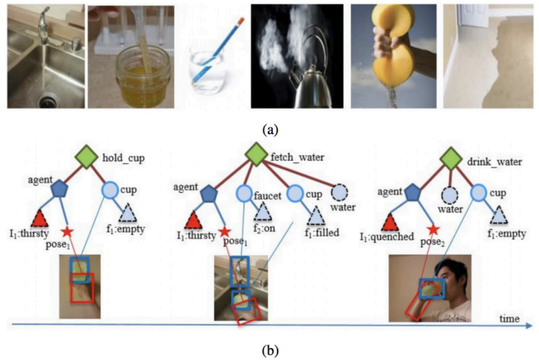
2 The Essence of Scene Recognition is Functional Reasoning

The human brain is highly efficient when calculating how much visual geometric accuracy is necessary to complete different tasks. This ability is a core aspect of the functional reasoning we perform when recognizing scenes.

Accurately describing a scene is essentially a matter of reasoning about the functions of things in it. Returning to this chapter's opening image, we can imagine water pouring from the kettle, we can imagine from where in the scene my daughter may have acquired a cup, and we can imagine where in the room one could sit and read a magazine. Moreover, modern design tends to feature mixed-use spaces and objects whose functions are not easily classifiable. For example, cooking, washing vegetables, having meals, and holding conversations can all happen in an open-plan kitchen. Sleeping, dressing, reading, and working can happen in a bedroom. The perception area inside our brains make connections between the tasks of identification and motion planning, and these tasks mutually influence how we perceive scenes.

To train AI systems that can perform reasoning about these functions, many scholars are now using image features from a large number of hand-labeled sample pictures to train a neural network to solve scene classification and segmentation. You might recognize this as a typical "parrot" approach to intelligence.

Meanwhile, a talented group of my former students, led by Xiaobai Liu, now an assistant professor at San Diego State University, has recently made remarkable progress on genuine functional machine reasoning, which can be seen in their publications. In addition, my former doctoral student Yibiao Zhao worked on similar issues as a postdoc at MIT in cognitive science and has now founded an AI company focused on autonomous driving.



Water and other fluids play important roles in our activities but are hardly detectable in images:

(a) Water causes minor changes.

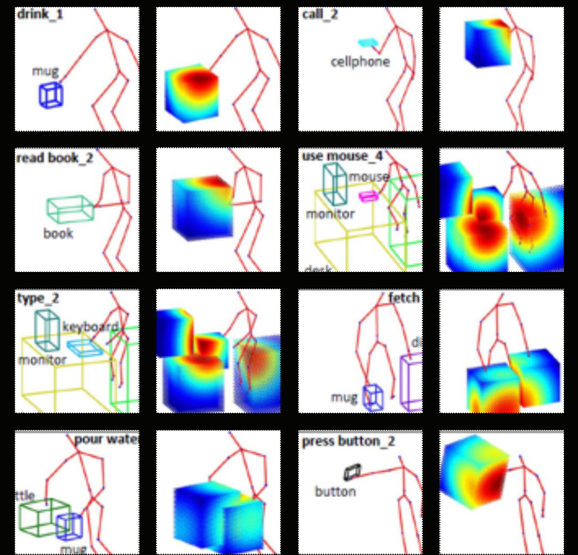
(b) The dark entities, water, fluents of the cup and faucet (triangle), and intent of humans are shown in dashed nodes. The actions (diamonds) involve agents (pentagon) and cups (object in circles)

Your brain's understanding of a scene incorporates actions that are underway in the image, or possible but yet to have happened. These potential actions are in your imagination, not in the picture itself, yet are critical to fully recognizing and defining the scene.

Two Categories of Visual Modeling

To "perceive" the functions of things in a scene, the human brain uses its very rich action model, whereby actions in the scene are divided into two categories:

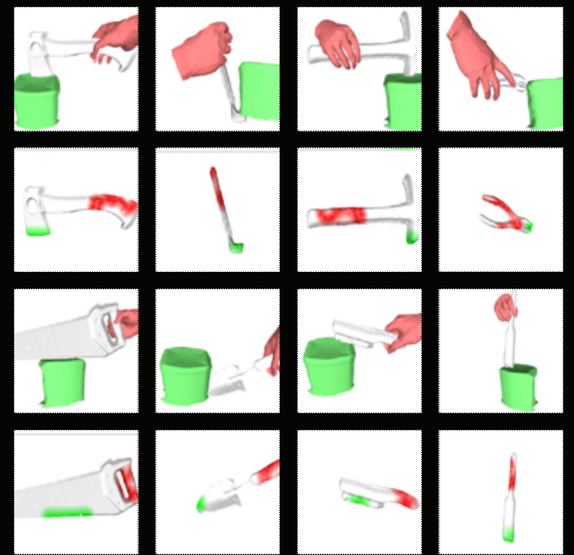
Modeling 4-D body interactions:



P. Wei et al ICCV 2013, PAMI 2017

The first category considers what is happening with the whole body, including such actions as sitting, standing, sleeping, and working.

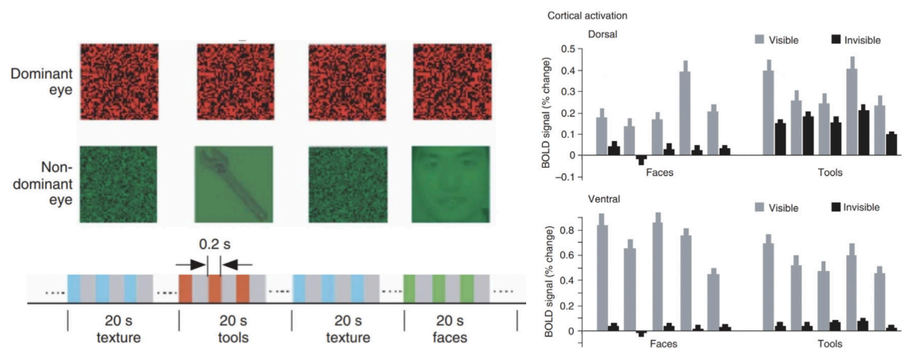
Modeling hand-object interactions:



Y. Zhu. Y.B. Zhao and S.C. Zhu, CVPR 2015

The second category considers actions of the hand, such as cracking, chipping, sawing, and prying.

These are four-dimensional models, taking into account both three-dimensional space and time. They can describe everyday activities, expressing the relationship between human actions and objects such as furniture, and the relationship between hands and objects as seen with tools. Psychological studies have backed up these categorizations; objects related to the body are stored in one area of the cortex, while objects that can be manipulated by hand are stored in another.



Cortical responses to invisible objects in the human dorsal and ventral pathways: Images adapted with permission from publisher. (a) stimuli(tools and faces) and experimental procedures. (b) both the dorsal and ventral areas responded to tools and faces. When stimuli were suppressed by high-contrast dynamic textures, the dorsal response remained strong to tools, not faces.

A Tale of Two Kitchens

This is why we all can understand that the following two pictures, despite their obvious differences in visual style and presentation, are functional equivalents, are the same type of scene. In simplest terms, both are kitchens. Human activities and behavior are essentially similar regardless of location and historical time period; this is the basis of our ability to generalize intelligently. If transported to a new environment, we don't need large amounts of new training to accomplish tasks.

Modern kitchen

Img 24.a

Ancient medieval kitchen

Img 24.b

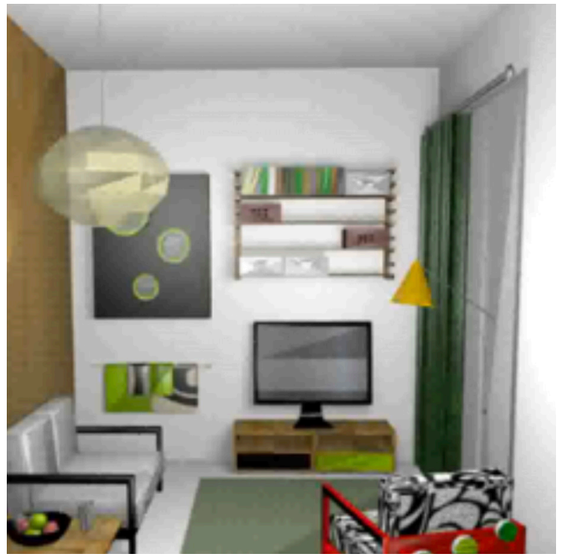
Looking back to the STC-PG (p. 23), we can see each scene broken down in a fundamental split between actions and functions (see the diamond node in STC-PG diagram in green).

With a blend of imagining and functional reasoning, our minds classify scenes, placing objects in a variety of positions and postures in the three-dimensional scene.

This classification method is completely different from current deep learning methods.

3 Physical Relations and Stability: A New Criterion for Scene Understanding

Our understanding of the images below incorporates the physical relationships between objects, including how each object is supported. For example, we immediately know that the chandelier and other items hanging on the wall in the image on the left would fall without adequate support, as they have in the image on the right. This understanding draws upon our innate comprehension of stability.



Img 25.a

Img 25.b

Professor Josh Tenenbaum of the MIT Department of Brain and Cognitive Sciences and I have been pursuing research in incorporating such abilities into machine vision for many years. As a part of my work, I have proposed a new criterion for scene understanding called "minimax": minimize instability and maximize functionality. This is more reliable for solving the basic problems of computer vision than image understanding using the MDL (minimum description length) standard. Function and physics, in my view, are the basic principles in scene understanding; geometric size is merely deduced from function or purpose.

In addition, perception of physical stability is a very powerful factor in our ability to reason about the relationship between bodies and furniture, between hands and objects. The height of a chair, as a result, is determined by its purpose of providing a comfortable seat; that's why a chair's height is the length of your calf.

Along with understanding how objects meet the various needs of human beings through functional tasks, we also understand how objects interact with the laws of physics. We can quickly perceive stability and instability. When we sense something is unstable, or about to fall, or both, we are quick to respond by moving out of the way. Professor Brian Scholl of Yale University recently found that the response time to physical instability was almost instantaneous, about 100 milliseconds.

4 Modeling Intention, Attention, and Prediction

Img 26.a

So how does the mind infer there's water in a container (refer to Img 26.a)? The water is not visible, yet the brain deduces there's water in both the vase and the kettle through a variety of reasoning steps. You may have also noticed the upside down ketchup bottle on the table, and you know exactly why. It's that way for the same reason you turn your shampoo bottle upside down when it's running low. You have internalized the physical and functional properties of viscous liquids, and the common sense solutions for when those dynamics make it difficult to get a liquid you want out of its bottle. Considering all of this, you can see how profoundly we understand scenes, incorporating far more than the object classification and detection currently possible for machines with deep learning.

Looking further, we see my daughter and our dog in the scene. We can see where my daughter's eyes are directed, which helps us identify her actions. From this and other factors, we can derive her motives and intentions. We can deduce what she is doing, and what she wishes to do. By accumulating this information, we can then reasonably determine what she might know. That is, if she looks at or notices certain things, then we can make predictions about what she wants to do next. Only through this level of scene understanding will it be possible for machines to better interact with people. Anticipation is key.

This type of work requires intention, attention, and prediction. How do these terms operate when it comes to understanding our work within computer vision?

"When we only see any picture or image, we add the dimension of time to our analysis through STC-PG. By considering the time before and after an image is captured, we improve analysis and prediction. When we reach the time when a robot can understand the intentions of others and predict their subsequent actions, it will possess the capacity to engage and cooperate with people. As we will discuss later, language can improve human-machine interaction and cooperation, but language communication also relies a good deal upon tacit understanding. Since what happens without words is also powerful in terms of communication, visual understanding is key."

The Visual Turing Test

Below is an example of machine visual understanding facilitated by an integrated scene taken by multiple cameras built in my lab. Understanding is demonstrated through an output in the form of a large integrated STC-PG. From this graph, we can generate a description in text form. We call it the Visual Turing Test.

As another DARPA project, this challenge involved taking a large set of video data and having a machine answer over a thousand questions about the data. The questions were based upon a three-dimensional model of human perception, including factors such as actions, attributes, relationships, and the like. Our system led the DARPA competition, besting many teams whose systems could not complete the task.

My lab had been working on a Visual Turing Test system for many years before the current work in VQA (Visual Q&A). Computer vision research groups investigating VQA today tend to train models with large numbers of image and text pairs, resulting in typical "parrot" systems that can't demonstrate real visual understanding. When considering VQA, systems built upon large datasets often provide answers that miss the contents of images, resulting in flawed logic. Unfortunately, the current reality of scientific research is that results must provide some kind of entertainment or buzz value to be well funded. Most attention goes toward technology that gives the appearance of intelligent behaviors while avoiding complex, real, and profound challenges.

On the Subject of Competition

Since it comes up often, we should consider the competitive atmosphere that surrounds computer vision and AI in general. From 2008 onwards, the annual CVPR (Computer Vision and Pattern Recognition) meeting has featured intense competition among algorithms trying to beat records on a variety of standard datasets covering image recognition, text detection, pose estimation, and other tasks. It has become a numbers game: newcomers submit algorithms that are simply small variations on prior algorithms, finely tuned only to beat incumbent winners, all without demonstrating substantial understanding.

One strategy for one-upping the competition is to download code that works, make small adjustments, and then build a larger module that includes the improvements. This is a fast and effective strategy for rising to the top of the leaderboard. While fast, it's hardly innovative.

I once visited a company (outside the computer vision realm) and listened to their head of R&D brag that his team always topped the list of their domain, beating teams from leading US universities. I reminded him that his opponents were usually made up of two students writing new systems, while his large team merely dug into other people's code, developing no new algorithms of their own. If others hadn't published code before them, they simply would not be able to compete. Many other institutions use this leaderboard strategy to gain bragging rights for having defeated the world's top AI labs – MIT, Google, Stanford, Berkeley, CMU, CalTech, UCLA, Harvard, and Brown.

"While winning competitions through incrementally building on others' ideas may be satisfying, it doesn't do much to drive the field of AI forward."

5 Task-Driven Causal Reasoning and Learning

Having discussed scene understanding, let us now turn to the recognition and understanding of objects, and why we need only to teach the ability to learn by analogy, rather than with large training datasets.

Humans are utilitarian social animals driven by tasks. When considering objects, we lean toward the teleological -- explaining objects in terms of the purpose they serve -- rather than toward the mechanical method of production. On the whole, when we see an object in a store, we tend to think of its potential uses -- its purpose -- rather than how it was manufactured or designed.

Of course, the use of an object is relative to one's impending tasks. When we can't see a potential use for an object, we do not notice it. But, once our upcoming task aligns with the use of an object, it becomes valuable, even fascinating. Objects around us are not neutral. Their value, their place in our lives, even their identities, are driven by the purpose they serve for accomplishing future tasks.

So if our knowledge of objects is task-driven, how do we determine what task requires which object? And how do we express this phenomenon mathematically?

One way is to consider the extent to which each task is, in fact, changing the state of certain objects in the scene. Newton invented a word for describing a change in state that suits our task here: fluent.

For example, when water boils, its temperature is a fluent. Ketchup and the spatial position of its bottle are fluents. Some conditions in physiology are fluents, such as hunger, tiredness, joy, and grief. Social relationships are fluents, too. Humans and social animals are busily changing fluents in various ways to improve our value function: from strangers and acquaintances, to friends, to close friends, and so on. Fluents impact our understanding of the world around us.

The Cognitive and Visual “Dark Matter” in every scene

Let’s consider three-dimensional scene understanding in terms of human actions. We apply causal reasoning to understand how actions and their potential interactions with objects lead to a change in fluents, or a change in the state of the scene.

In the world of literature, Detective Sherlock Holmes solves cases from small, seemingly out-of-the-way details that those who lack his training, his skills, and his experience often dismiss as insignificant. How does Holmes deduce so much information from such scant scatterings of clues that are so easy to overlook?

The example of Sherlock Holmes teaches us that great detective skills require us to be conversant with fluents. Amassing considerable knowledge about the world allows us to easily imagine what will happen next in a scene and why, allows us to know what the fluents are and how they will change, allows us to understand people and how we all typically behave, allows us to have an understanding of how objects respond to the laws of physics. Often, this knowledge is gained through experience and is deemed even more impressive when we are able to deduce a complex, nuanced understanding of a situation tacitly.

In physics, dark matter and dark energy are thought to comprise 95% of the total mass and energy in the universe, while observable mass and energy take up the remaining five. It’s helpful to think of visual understanding in a similar light: the observable, two-dimensional arrangement of pixels that make up an image informs only 5% of our understanding of a scene. We call the important aspects of a scene that are not explicitly present in an image “dark matter,” as the majority of our comprehension comes all that which is only expressed tacitly -- the function of objects, physics, causality, intent, motivation, and the like. Perceiving this “dark matter” relies on our minds’ abilities to imagine, to reason, to experience, and to recall implicit information learned through experience.

Like deep learning networks and other algorithms that use big data, the parrot learns to imitate the world. Crows, unlike parrots, observe the world with singular intent, and equipped with spatio-temporal-causal reasoning, crows develop a sense of how things work. When it comes to learning, crows are similar to people in that they are able to make inductive leaps when solving problems; they are capable of extrapolating general conclusions based upon specific observations.

In the case of human intelligence, we are capable of imagining the thoughts of others. This capacity gives us the power to reason not only about space, time, and the physics of cause and effect, but also about the intent and values of those around us. Such social reasoning is the basis of communication and the ultimate prize of language.

Learning from Observation Over Time

Task One: Cracking a Walnut

Let’s consider an example from our paper published in CVPR 2015 to see how observed knowledge from lived experience is critical for humans to solve problems. In this fascinating experiment, a UCLA student’s task is to choose from the variety of tools on a table and crack a walnut, or to change its fluent. Unsurprisingly, he chose a hammer. To a human, it’s the obvious choice; you would likely choose the same.

Learning from one example. Test: Generalization and Innovation. *Yixin Zhu et al. “Understanding Tools” CVPR 2015*

But when we think about it, this task is quite complex, requiring the student to use considerable information. Why choose the hammer? Why hold the hammer by the bottom of its handle? How to decide what amount of force to use ? There are millions of other possible combinations of choices that would have led to a different course of action, and the student is aware of the different choices he can make. The task is perceived as simple only when we overlook how complex these seemingly simple questions are.

In the scene on the right, the student must choose from a completely different set of tools to complete the same task. By observing the student’s pondering of the challenge and comparison of tools, we can understand his task and build an idea of how it will be solved. One can understand the importance of observed knowledge in considering the traditional relationship between a master and his apprentice. Rather than receiving direct lessons from his master, the apprentice, to develop his own skills, watches the master at work. This analogy is an important reminder of the importance of observation in human learning.

Using observed knowledge, the student on the right is not confounded by the array of new tools; he can understand that the wooden table leg is a close analog to the hammer, and while he may need to apply a slightly different amount of force and change other aspects of his action, as a general principle, cracking a nut with these tools is very similar to cracking a nut with the original tools. This example illustrates real intelligence: applying past knowledge to a new task, all without big new datasets, models, or training. This approach marks a departure from deep learning.

The algorithm for this observed phenomenon can be expressed as an STC-PG. The STC-PG maps out our understanding of physical space, including objects, three-dimensional shapes, and materials. It also describes our time action planning and causal reasoning. When deciding how to crack a nut, we are choosing what actions with what tools will cause the shell to break. We are able to plan this action by stitching together our understanding of cause and effect, of time, and of space into our analytical map. In this regard, visual understanding helps us achieve our goal of changing the nut’s physical fluent.

Expressing the STC-PG

First, the STC-PG is in place in our minds before we take an action. The nodes and the majority of the edges in the graph are not presented in the image, but in our heads. This is the "dark matter."

Second, a large number of calculations we make using dark matter are "top-down" calculations. That is, we apply the knowledge we already have in our cortex to the clues we see in an image or scene in order to formulate a solution. This top-down calculation process is missing from multi-layer neural networks. Currently, these networks are feed-forward only, propagating information layer-by-layer to the top. This backpropagation is a bottom-up process, building its conclusions about a scene up only from the patterns of light on a camera sensor, without applying any top-down knowledge. In his 2018 lecture at UCLA, Yann LeCun, director of AI research at Facebook, agreed that deep neural networks currently lack the top-down process that I have been advocating.

Third, learning such tasks requires only a few examples. If a person needed a large number of examples of an action to learn it, it would be difficult to survive in the world. She must learn to perform tasks using a relatively small amount of data acquired through observation.

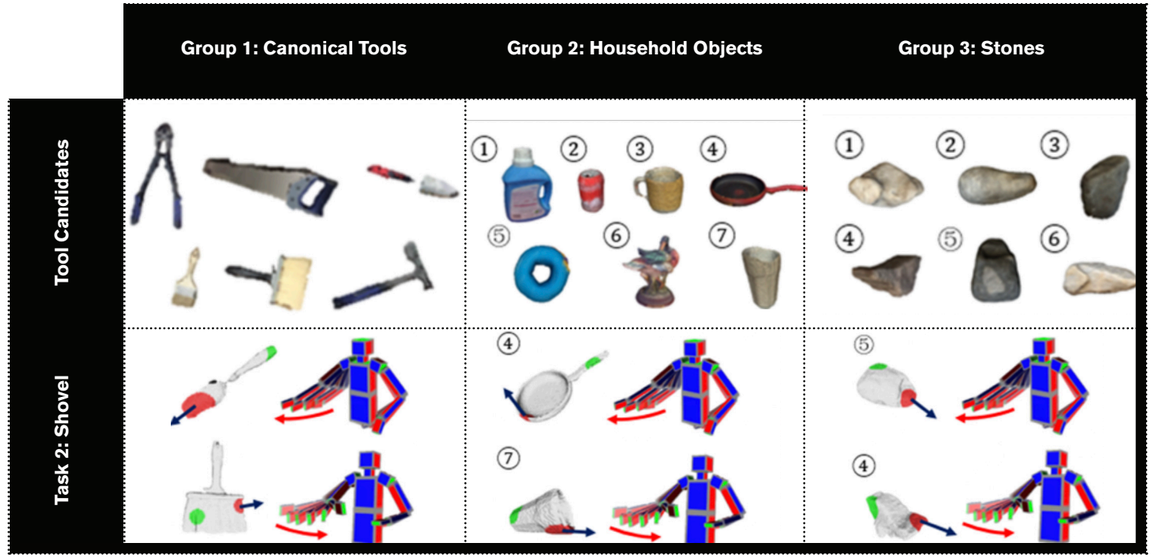
It takes more than a high IQ to be able to learn. At the end of every quarter at UCLA, students assess the quality of their instructors. I typically receive feedback that I give too few examples, but an ability secured only after analyzing an ocean of exercises is not genuine ability, nor does it conform to the nature of learning. The reality is that there is not nearly enough time in class to provide large numbers of examples while still covering the course's required theory and other aspects of the curriculum.

In the quote from Confucius above, "reasoning" could be substituted for "thinking." The nature of understanding social phenomena, behavior, and tasks is to form a self-consistent visual interpretation, which in my opinion can be captured in an STC-PG.

So how is STC-PG derived? It is based on the STC-AOG, where AOG is an And/Or Graph. An AOG is a complex probability grammar graph model, which can derive a large number of events with a probability that satisfies the constraints specified by the grammar model. In this case, each event is a STC-PG. This expression is consistent with language in the fields of cognition, robotics, and the sort. In my opinion, the STC-AOG is a unifying expression, as it can be connected with logic as well as with deep neural networks.

Task Two: Shoveling

To build on Yixin Zhu's nut-cracking experiment, let's consider a problem of greater complexity: shoveling. We can use the action of shoveling to examine an algorithm's generalization ability.



The first group of experiments

In this experiment, the robot is given an assortment of tools and tasked with shoveling soil. Fortunately for the robot, its first choice of tool for shoveling is the shovel. This is not just pattern recognition, as it recognizes the shovel by its use, and may even be familiar with its efficiency. Its second choice is a brush.

The green area on each tool above indicates where the robot wants to hold the object, while the red indicates the part it intends to pick up the soil. Note how these focal areas and the motion of the robot's arm will change depending on its choice of tools.

The second group of experiments

In the second phase, the task remains the same, but we replaced the original set of tools with common household objects that would not normally be used to shovel soil.

The robot's first choice is the pot, and its second choice is the cup. Indeed, these two objects were the best choices for the task. These choices were made automatically with computer vision.

The focal areas and arm motion again depend on the choice of tool. Even though they are all used to shovel, there are subtle differences in how the tools are used.

The third group of experiments

If we found ourselves in the Stone Age, could we shovel with a pile of rocks? Maybe not. Accustomed as we are to tools designed for specific tasks, we might not know how to choose from among objects that are not tailored to particular tasks. In this case, the visual cognition necessary to choose the most effective tool for shoveling may be reduced to pattern recognition – simply relying on the configuration of color and brightness in the objects we see, rather than understanding what each object is for.

In other words, modern living may be reducing our cognition from that of the crow to little more than that of the parrot.

Computer Vision Summary

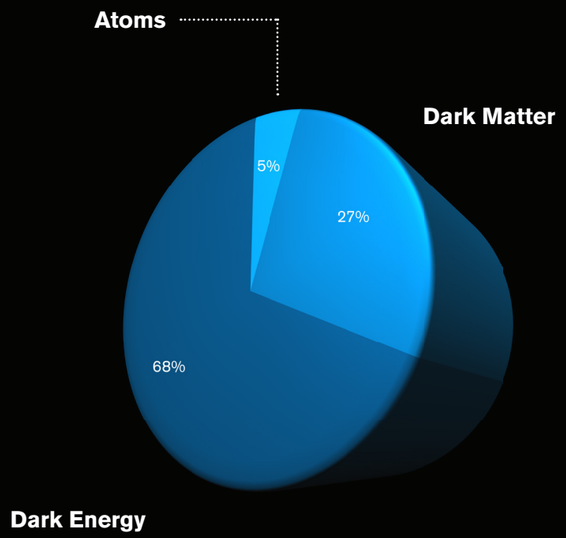
OBJECT CENTRIC	VIEW CENTRIC	TASK CENTRIC
Geometry-Based, 3D	Appearance Based, 2D	**Dark Matter** Functionality, Physics, Intentionality, Causality
1970-1995	**1990-2015**	**2015-?**

Functionality, Physics, Intentionality, Causality, Utility
2015-?

The first twenty-five years of computer vision research was geometric-based and object-centered. A fact of appearance-based and view-centered computer vision is that geometric shapes dictate appearance. But what reason undergirds the reliance upon geometry? The geometry of an object is shaped by the tasks. With a task orientation, we take into account the function, physics, causality, and design of objects to produce images, which is the core challenge.

The last twenty-five years, however, have been based upon extracting rich image features, relying upon the appearance of objects in the image itself to perform recognition, much like the human brain. A necessary part of this recognition involves, quite literally, reading the dark because, as we know, that which can be physically seen and used by deep learning algorithms is only a small part of what we perceive when looking upon an image.

Our perception of dark matter best represents our real “crow” intelligence and brings us closer to truly cognitive artificial intelligence. For computer vision to continue to evolve, dark matter is our new frontier.



Section 6

Discipline 2: Cognitive Science - Into the Inner World

The dark matter discussed in the previous section combines perception and cognition, bringing us into the inner world of humans and animals. This inner world is affected and distorted by necessary tasks and the motives necessary to achieve these tasks, thus reflecting the external world.

Research questions on the theory of mind include the following:

What did it see? What did it know? When did it know it?
These make up the historic accumulation of visual information over time.

What is it focusing on now?
This is the current task that is being executed.

What is its intention? What does it want later?
This is predicting future goals and motives.

What does it like? What is its value function?
These questions, with concrete examples, shall be considered in Section 8.

Psychological research on these questions is called “theory of mind.” In 2006, Rebecca Saxe and Nancy Kanwisher (a collaborator of mine from the MIT Department of Brain & Cognitive Science) found that the human cortex has a dedicated area for feeling and reasoning about someone else’s mind. This capacity helps us consider how others might behave, what they may be thinking, and what they may want to do. Theory of mind has been essential to AI since the dawn of the field. In his study The Society of Mind, Marvin Minsky explored these questions extensively.

The Chess Master

Consider the chess master. At a given time, a chess master may be maintaining several games, each with a different player.

This is extraordinarily difficult, as each of the games is different and depends on the chess master’s understanding of both what his opponent will do next, and what he believes his opponent thinks he himself will do. This is necessary to be able to deceive an opponent with unexpected moves.

The chess master’s cortex would be significantly more developed than those of his opponents, if they were each only playing against him. It’s similar to what happens to a spy in a film who receives training in “anti-reconnaissance,” or trying not to let other people find out what is on his mind. Minds expand when they must consider the thoughts, intent, and inner lives of others.

The Awareness of Other Minds

R. Saxe and N Kanmisher (2006) reported special cortical areas for the Theory of Minds in neuroimaging eperiments

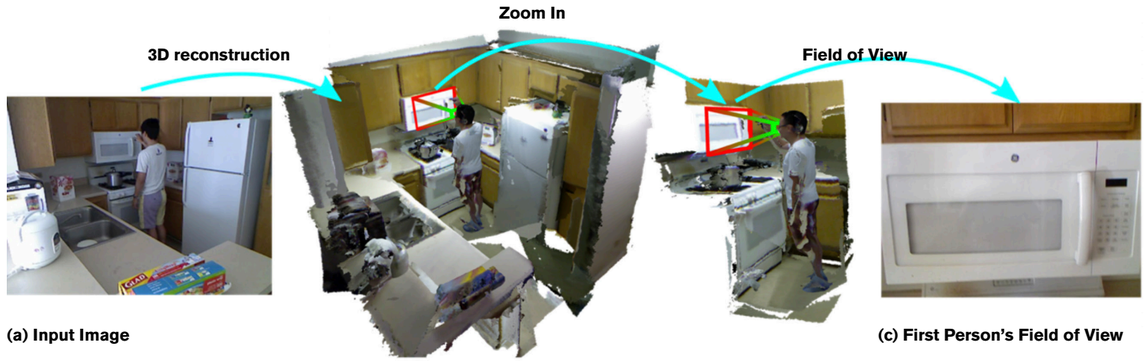
Human adults aren't the only creatures who can imagine what's on another's mind, and they aren't the only ones who work to conceal from others what's on their mind. Consider a bird looking to hide a nut from competitors. The bird might first look around to see whether there are other birds or animals around. If it's not alone, it may choose to wait for prying eyes to leave before finding a hiding place for his dinner. Likewise, an otter who has recently caught a fish might see a fox craftily watching him and understand that the fox might want to steal its lunch. The otter may then take the fish underwater, keeping it out of the fox's sight. These examples suggest that even animals can consider what other minds are thinking.

Infants have these skills too; humans first develop this awareness a little over a year into life. Conducting psychological experiments when he was a doctoral student, Felix Warneken, now a professor of social sciences at Harvard, showed that one-year-old children actively seek to help others with acts such as opening a cabinet door when they sensed adults trying to put items inside the cabinet. According to generations upon generations of parents, young children have, long been known to understand and to cooperate with adults, even if only in small, yet important, ways.

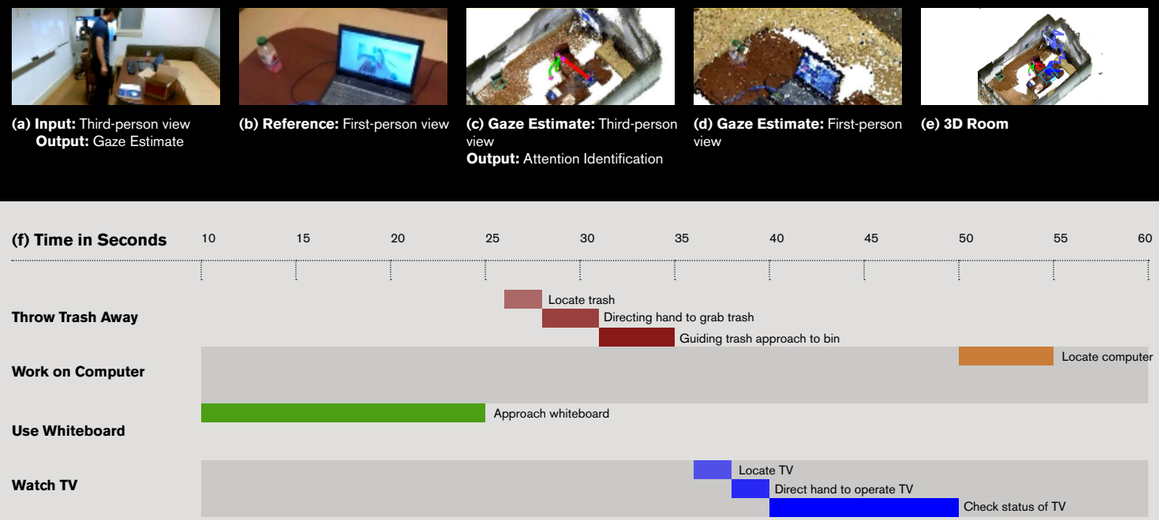
This is the human-machine interaction to which we aspire. A robot should be able to understand what its user wants to achieve, and this capacity will be a core aspect of human-level AI. So far, we have only theorized about toy examples.

I believe real-world research on this matter starts with computer vision. But currently within the computer vision community, most researchers are busy winning competitions. Too few are focused on the problem of building awareness of others' minds. Actively pursuing real-world understanding, my lab is on the leading edge of this fascinating area of research.

Task Three: Using the Microwave



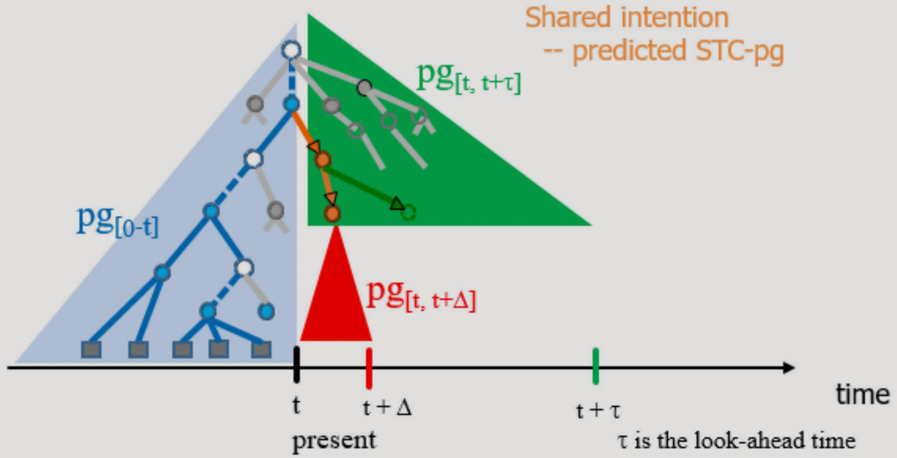
Let's consider a simple experiment, pictured above. The person is in a kitchen using a microwave. There is a camera pointed at him (left); think of it as the robot's eye. First, the robot's eye needs to figure out what the person is looking at (middle); then, it must use its perspective on what the person is looking at the calculate the person's field of view.



Above is a series of screenshots from a different experiment. To give context, imagine the images are from a nursing home or hospital ward. Assume that the robot is already familiar with the three-dimensional room (pictured in e) in which a person performs activities. The robot must try to figure out what the person is doing or seeing (c). Its only input is a two-dimensional video (a). By tracking the trajectory of the person's eyes (illustrated in f), the robot estimates what the person is looking at (d). In other words, the robot imagines the person's perception. Compare the result in (d) with the actual visual field of the person in (b). These results were provided by Dr. Wei Ping, a junior faculty member at Xi'an Jiaotong University in Professor Nan-Ning Zheng's research group. He visited my lab as a doctoral student and returned later for further training.

A machine performing this task must calibrate the spatial and temporal interaction between actions and objects where actions change over time and must understand hand-eye coordination. From these fluents, the system must infer what the person intends to do next.

Representing perspective with an STC-AOG and STC-PG



The perspective of the person can also be expressed with an STC-AOG and STC-PG, as shown above. The figure can be roughly broken down into four parts, shown at right:

- **Knowledge:** In a joint probabilistic STC-AOG with definitions of task space, utility table, etc.
- **Situation:** In a partial parse graph pg[0-t] for time interval [0,t]
- **Intent and Plan:** In a partial parse graph pg[t,t+τ] predicting the actions of an agent (or self)
- **Attention** In a sub-parse graph pg[t,t+Δ] focusing on current steps in a time interval Δ

First, STC-AOG, a probabilistic grammar model spanning across the space, time, and causality. It represents the tacit knowledge of the person, containing all possibilities, which we will return to later. The rest is an STC-PG: an expression of current time and space. The interpretation of the map contains three parts represented in this figure as triangles, each of which is also its own STC-PG.

Second, the blue triangle on the map represents the current situation. It is also an interpretation of the visual representation of the scene in the 0-t time period.

Third, the green triangle represents the person's intention and action plans. This is also a hierarchical parsing/solution, predicting what he will do next

Fourth, the red triangle represents the person's current object or area of attention.

Putting it all together, we have a representation of the current state of the person and of the near future. They are interpreted using a unified STC-PG and STC-AOG. This is a hierarchy of composition, hence requiring very few samples.

The Robot Imagines Other Minds

One may note that a neural network also has a hierarchy in its depth; this hierarchy can reach more than a hundred layers. But in fact, a deep network operates like it only has one layer. From the input layer to the output layer, the middle layers are all a mystery. You can only explain the last layer that provides the output of object categories, not the processes that lead to it.

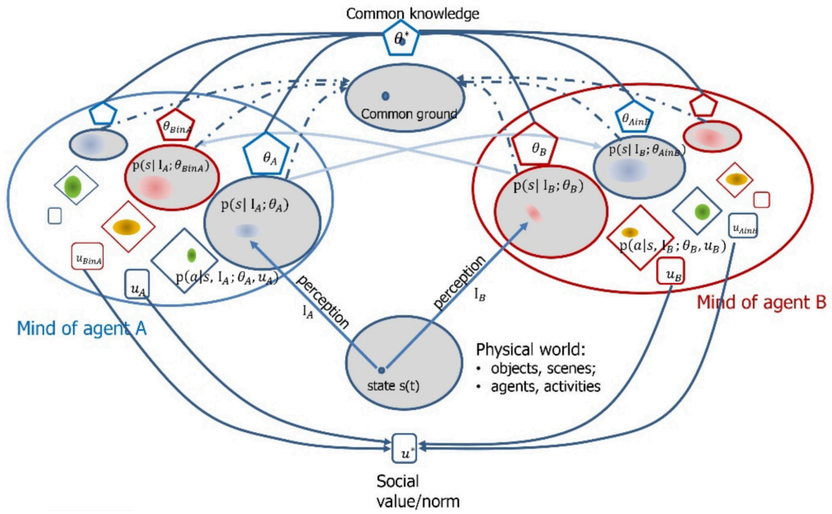
The representation we just discussed is a robot's estimate of the inner state of a person (or another robot). This estimate has a posterior probability. The estimate is not unique, and there is uncertainty surrounding it. The estimate is just an estimate, not definitive truth. In fact, the estimates of different people's observations about the same person, while potentially similar, may not be the same.

So, let's consider an environment containing machines and people. Assume there are *n* robots or people in the scene; in other words, there are *n* selves, or independent minds. If everyone has an estimate of all other minds, we have *n * (n-1)* expressions of minds. Each knows what the rest are thinking, and the number of expressions is at least n squared. If you have 100 friends, you will have some idea of what is on all of their minds; the closer the relationship, the deeper the understanding, the more accurate your estimate.

And this is just first-order reasoning. In a complex, adversarial environment, people are tasked with inferring multiple levels of recursion. When the Greeks had seemingly sailed home from Troy, leaving a giant wooden horse behind, the Trojans imagined the Greeks felt the need to flee battle and accordingly took that fabled horse as spoils of war. But the Greeks, having already predicted what the Trojans would think once they left the beach, hid inside the wooden horse. We know the rest.

Once we set about deliberately deceiving others, the number of minds to be imagined increases exponentially. Yet this is how intelligent beings behave: anticipating the thoughts of others and acting accordingly.

Representations between Two Minds



To summarize: in the illustration above, A and B represent two people, or a person and a robot. The representations inside their heads are shown with a nested recursive structure, each of which represents the mind within a brain.

In addition to the STC-AOG as the knowledge representation and STC-PG as the parsing of the current state, each mind also contains value and decision functions. Values drive actions, and actions, based upon perception, change the state of the world. This is cause and effect at its most elemental.

Look at the oval circle at the bottom of the figure. It represents the real world (the objective truth that we do not know). The middle of the oval is consensus, or the unified, common understandings of a group that grows out of individual perceptions. For example, when the food arrives at a restaurant, everyone can see what the dishes are; it's typically clear whether we all agree on what we are eating, and this consensus provides a basis for interaction. By contrast, in "The Emperor's New Clothes," there is no consensus over what the emperor is wearing. This tale reaches the core of epistemology: how is it that we know what we know?

In philosophy class, epistemology tends to be discussed in abstract, vague terms; here, we see the problem of knowledge concretely.

We have to establish consensus common knowledge to have common values, both within a small group and within a larger community. Where there are common values, social morality, and ethical norms, a consensus on common knowledge can be attained.

As the saying goes, when in Rome, do as the Romans do. When you join a new social group, you may first observe people to see how they interact. For a robot to coexist with people, it must understand the social morality and ethical norms of human groups. Developing a deep knowledge of norms and conventions, therefore, is critical for the healthy development of robots. The crow knows what human beings are doing and thinking, and it uses this knowledge to survive.

So how do groups reach consensus? The tool we use to build consensus is language.

Section 7

Discipline 3: Natural Language Understanding - The Cognitive Basis of Communication

1) Intentional Communication

2) First Glimmers of Cooperative Communication

3) Recursivity ==> Fully Cooperative Communication

The third discipline of AI research is language and dialogue. The human brain's language center is located near the action planning area, which makes sense. After all, why do we talk? The origin and purpose of language is to convey a message from one person's mind to another. This act requires implicit knowledge, intent, and planning. The goal of language is to form a consensus on common task plans so that we can act in concert. In other words, the central purpose of language is to enable cooperation.

Animals have rich communication that is often based upon body language. Humans, similarly, use nonverbal means to convey information about ourselves, from gestures to facial expressions, from body movement to signs. Nonverbal communication is so fundamental to humans that we have developed a special type of art called pantomiming. We already possess rich cognition, including tacit knowledge and consensus values, before we even begin language formation; without this foundation, language would be merely empty symbols. Dialogue could not develop.

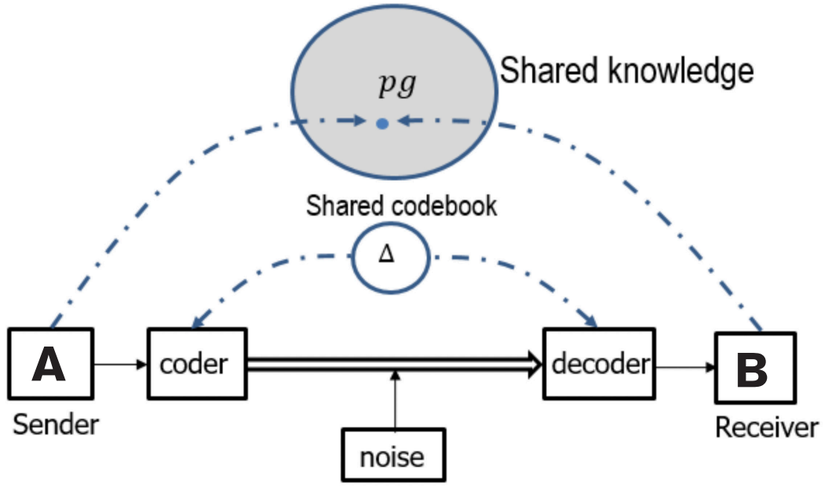
The Lost Gorilla

Developmental psychology experiments show that one-year-old children know how to refer to things by pointing, while babies younger than one cannot. Many animals never reach this level of communication ability.

For instance, you may have heard of the experiment in which a young gorilla child becomes separated from his mother while the two are playing in their zoo. The other gorillas, lounging about in the sun, don't help the mother find her child. If they were human, the sun loungers would most likely be quick to communicate what they know; this is because humans are born to cooperate, and we derive pleasure from helping others. But gorillas have not evolved in this way, and the mother's companions, missing the situation's gravity, are not quick to show her where her child has gone.

The lost gorilla experiment suggests something is absent in gorilla brains that human brains possess. As it turns out, we are more advanced in our cooperative abilities than other animals because there is a cognitive framework of communication (like a multi-layer network communication protocol) within our cerebral cortex. Without this cognitive architecture, the level of communication we are accustomed to would not exist. If people who study machine language fail to study the underlying cognitive framework of communication, they won't accomplish much.

The Importance of a Shared Reality



The figure above from Professor Michael Tomasello, a pioneer of anthropological research and social cognition, illustrates the mutual understanding and sharing of knowledge and intentions. Since visual perception is a critical aspect of the foundation of shared reality, research on language cannot be divorced from it. Otherwise, language is superficial, like a tree without roots. This is why chatbots are so easy to spot and differentiate from humans; they have no shared reality with the humans with whom they interact.

Let's begin our exploration of shared reality with the most basic communications process: the delivery of a piece of information.

Sending a message is a very simple process and was modeled mathematically by Claude Shannon of Bell Labs and published in "A Mathematical Theory of Communication" in 1948. First, a message is encoded to reduce its size and to allow for the fastest transmission. Then, some redundancy is added to enable recovery of the original message if it is somehow lost. Finally, decoding delivers the message.

Two basic assumptions inform this process. First, the two sides share a codebook, without which there is no way to decode the message upon arrival. Second, the two sides have shared knowledge of the outside world that gives context to the message. The sent information is merely a fragment of the parse graph (PG) describing some state of the physical world, which makes sense only when considered alongside other aspects of the world that are part of shared knowledge between both sides. This state may also be ideas and feelings -- fluents inside our mind. Shared knowledge is how one can sometimes have a rich interaction over the phone despite the exchange of relatively few words.

In contrast, if both sides do not have a common understanding of the world, then one side cannot truly understand what the other is saying. Idioms often create this kind of challenge between two speakers of a language, especially when one is a native speaker and the other isn't. A native Spanish speaker, for example, might say "Hay Moros en la costa" (There are Moors on the coast) to mean, in English, "The walls have ears." Without understanding the context of Spanish history and culture, the English speaker would fail to comprehend this idiom's true meaning.

Language and the Epistemological Function

Shannon's communication theory focuses only on the establishment of codebooks (such as video codecs) and communication bandwidth (3G, 4G, 5G). Many mathematically-adept researchers explored information theory after its introduction in 1948 but with few big breakthroughs. Why? Perhaps they ignored the important epistemological issues below:

Common World Models

B might consider if there are any common world models in A's head to permit understanding, once decoded.

Overall Importance of Message

How important is this message? Is it urgent? Will it affect other conditions beyond A's response?

Intent of Sender

B may also consider why she is sending the message. Is this information A doesn't know?

Intent of Receiver

A may want to consider why he should receive this information. Does he really want it?

Desire to Communicate

Does A care about this information? Would A like to have this information? Does B need to communicate?

Possible Presence of Deceit

A may want to consider whether the content of the message is true, or if it is intended to deceive.

Perceived Response

What will be A's response after receiving the message? Will it be desirable?

Possible Presence of Manipulation

A may want to consider whether the message is crafted in such a way as to provoke a certain response desired by B.

Communication theory only covers the sending of information, similar to the way telegraph services collected money to send messages, regardless of motive, content, and consequence. Cognition, which takes shared knowledge into account, takes place beyond the reach of such code-dependent processes. One can see the power of shared knowledge and compositional structure in Chinese pictograms where each word is a picture of something in the outside world, removing the need for code.

Let's look at examples of Oracle bone script from ancient China. Each bone script word is a picture, or in the world of AI, a fragment of a parse graph.

From Vision to Language

Characters (Oracle bone script) = Graphical Model

- Objects and Scenes
- Human Body Pose
- Actions: a HOI or HHI
- Attributes and States
- Space and Time Relations
- Fluents
- Minds
- Causality
- Ethics and Morality

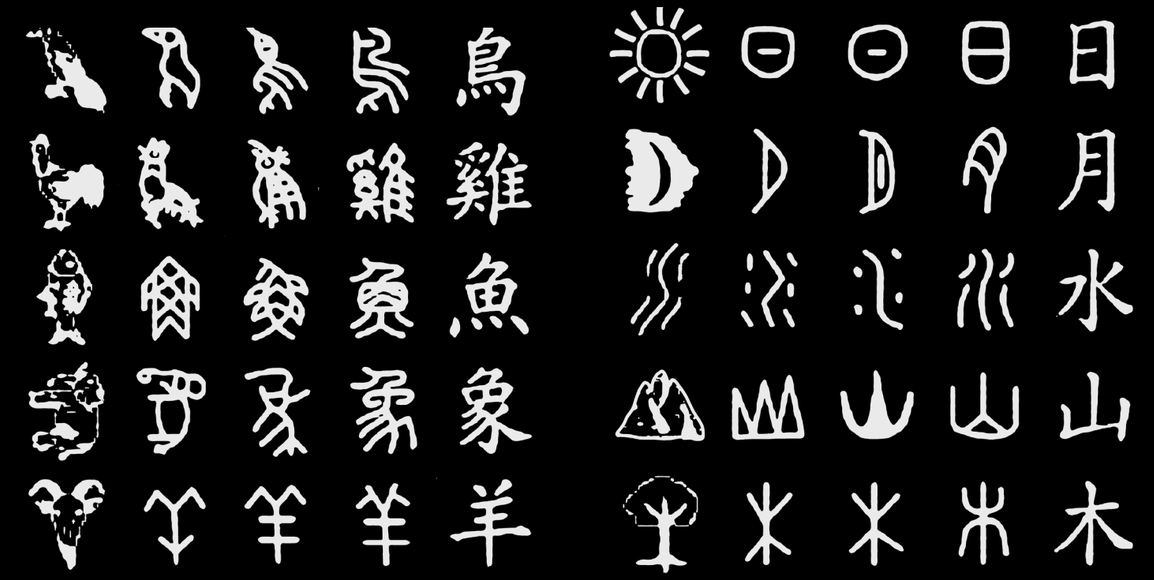
Knowing the written Chinese language has given me particular insight in my research. The above illustration shows the evolution and network of relationships of the Chinese character for “eye” (目). It comes from a fascinating book I found on a visit to Taiwan a few years ago called Chinese Character Tree. In the middle of the diagram, the character for “eye” is a logogram resembling an actual human eye. If you put the character for “hand” above the character for “eye,” perhaps to shield against the bright sun when looking to the horizon, the character for “to look” (看) is formed. There are many other associative compounds built around 目; for example, the character for “self-reflection” (省) puts a small leaf symbol above the character for “eye,” which instructs the fewer to look carefully to see things among the leaves.

From here, abstract concepts that describe how to express time and space evolve. There are words for the beginning and the end, for relationships between people, for states of mind, and even for ethics. As the different individual glyphs combine and evolve, vocabulary expands.

“If we use the same approach to research visual awareness of objects, we would need to look all the way back to the functioning of objects in the Stone Age. Similarly, to research language, we need to look all the way back to its origins.”

Evolving Visual Expression

The following images illustrate how a number of Chinese characters have evolved from their earliest known pictograph to their present forms: 日 for day, 月 for month, 山 for mountain, 水 for water, 木 for wood, 雀 for bird, 雞 for chicken, 魚 for fish, 象 for elephant, and 羊 for sheep. The color chart is a graph model of these objects from our lab created using computer vision technology. The graph model is a kind of variant of the object hierarchy of words in Oracle bone script.

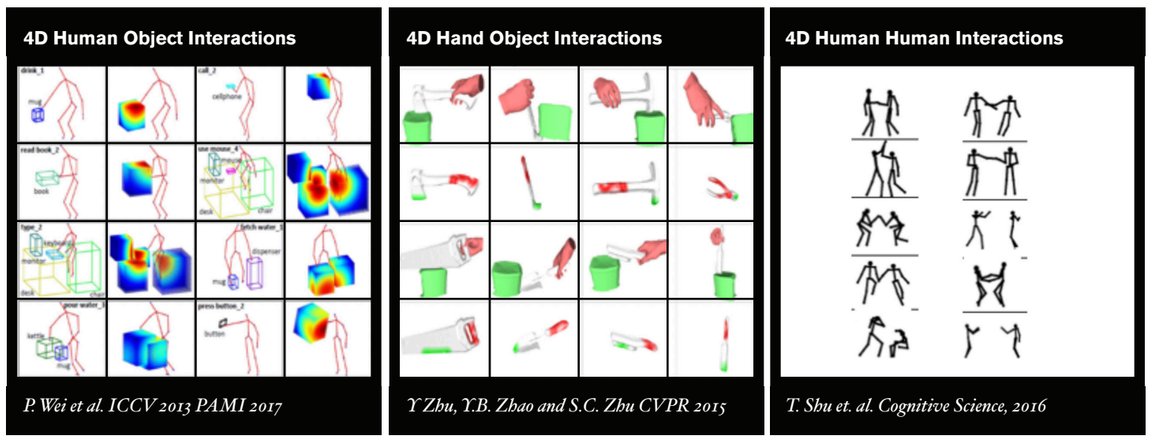


This graph model is from research by Yi Hong, Zhangzhang Si, and other doctoral students on unsupervised learning. Their algorithm discovers bone script characters that represent the bird’s head, body, and feet, along with water waves and aquatic plants. This visual expression model is explainable and intuitive. No code is needed. From the perspective of generative models, the language is vision; vision is the language.

Looking closely at the pictures of early Oracle bone characters describing actions (below), one gains a sense of what the characters mean. Look at the character for “haul;” there are two hands, a rope, and an object on the ground. “Wash” shows two hands in a basin with water; “close” shows two hands closing a door. In “assist,” one hand lends help to another, and in “accept,” two hands facing each other reflects the transfer and acceptance of an object between them. “Compete” clearly shows two hands pulling on the same object in opposite directions. Finally, “chat” depicts two people kneeling face-to-face. In a nutshell, the words already illustrate the detailed actions among people.

How We Communicate Interactions

In my lab today, we can teach computers representations of verbs similar to the Oracle bone script described above.



These symbols express actions between two parties. Those learned by the computer include sitting, using mobile phones, shaking hands, and others. We call these action models 4DHOI (4D Human-Object Interaction), 4Dhoi (4D hand-object interaction), 4DHHI (4D Human-Human Interaction). Remember the core differences between these interactions.

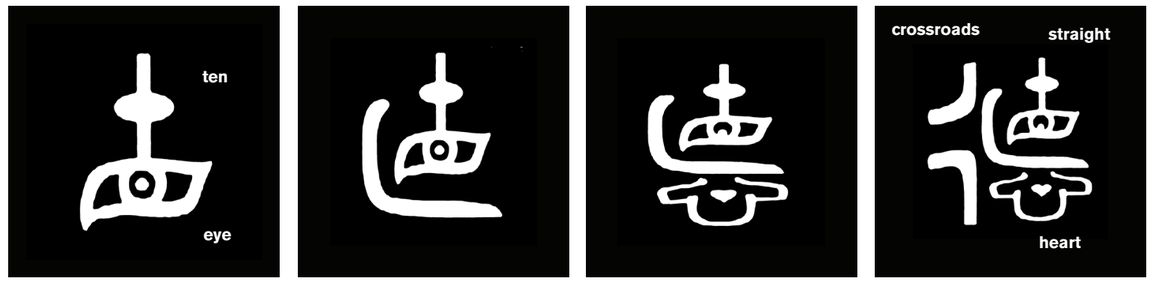
But How Do We Communicate Ethics?

We have discussed nouns and verbs, but there are many other things that must be understood and communicated by machines. Our bone script model can express anything in the world, including complex and abstract ideas such as ethics and intent.

As mentioned in the introduction, as AI systems enter society, many are concerned about harmful effects on humans. In fact, much public discussion of AI begins here. At a DARPA internal meeting on this issue, with professors from fields as varied as social ethics, cognitive science, and AI in attendance, everyone had a different idea. When it was my turn to speak, I said that for this problem, ancient Chinese wisdom actually has something to offer.

What is Morality?

What is the definition of "morality" in the study of ethics? Morality is a relative concept that varies depending upon time and group, as one sees today with issues like abortion and marriage equality.



In the images above, let's look at the early Oracle bone glyph for the Chinese character 德 (pronounced "de"), which means "moral" or "virtuous." In its most ancient form (in the far right box), it combines the symbols for crossroads, for heart, and for straight or upright. The crossroads symbol is on the left and is signified by two lines, or paths. These lines are part of this character because crossroads compel us to make choices, and choices are where morality comes into play. If an old man falls down, for instance, should we help him up? Decisions such as these are made in our hearts, offering us an appreciation for why a pictograph of a person's heart makes up the bottom of the character.

But how can we judge if our inner choice is morally correct? The number of potential actions in society is endless, and one Chinese character can only express so much. But we can see by looking at the third pictograph in the character for moral, which means straight or upright. The pictograph for straight is itself composed of the symbol for the number ten (十) above the symbol for eye (目), together meaning ten watching eyes. This seems to signify that moral correctness is determined by society. If society deems an action acceptable, then it is , typically speaking, moral; if society frowns upon it, it is, most likely, immoral. So when making what we consider moral choices, we consider, sometimes unwittingly, how others might view that choice, and, by extension, what others might think of our character.

One of the critical aspects of today's robots is how they make decisions. If the anticipated reaction is positive, we should go forward with our actions. This rule does not vary by situation. Students of philosophy might recognize this idea as a version of Kant's theory of deontology, wherein people should act in such a way so that their actions create a consistent rule that is easily recognized. If a robot can't recognize what others might think of its actions, then it can't infer morality.

But how can robots know what others think? It must first understand what people around it like and dislike, since not everyone is the same. Humans depend on their knowledge of how people would react to certain words and actions; without this, nobody would know how to interact with others. The same principle, by extension, must hold true for human-level AI.

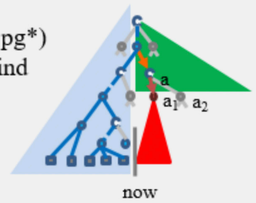
The writers of ancient Oracle bone script understood that words can express problems insightfully. We've used them to ponder the important challenges we must address. But pervasive popular media prioritizes entertainment; and for a human, being entertained often trumps being informed.

Shared Minds

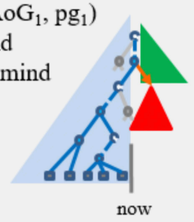
Now we return to the communication protocol problem between people and robots. Let's begin by examining the following cognitive model.

Human Minds

1 H =(AoG*, pg*)
Human's mind

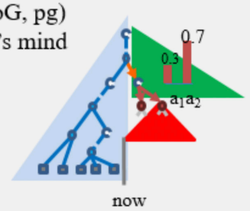


3 R_in_H =(AoG$_1$, pg$_1$)
Robot's mind
in Human's mind

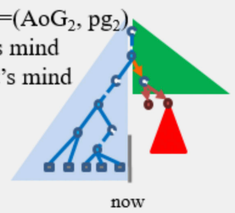


Machine Minds

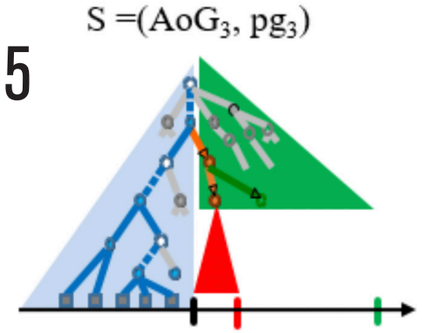
2 R=(AoG, pg)
Robot's mind



4 H_in_R =(AoG$_2$, pg$_2$)
Human's mind
in Robot's mind



5 S =(AoG$_3$, pg$_3$)



Shared and Confirmed

For a person and a robot to communicate, at least five epistemological concerns beyond our core understanding of the conversation's purpose must be understood. They are as follows:

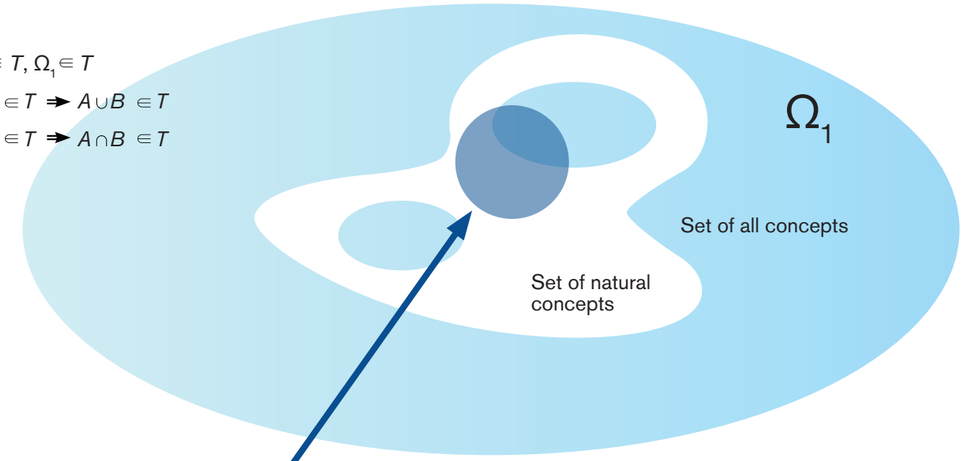
1. **Things I know**
2. **Things you know**
3. **Things I believe that you know**
4. **Things you believe that I know**
5. **Things we know together**

A Math Perspective:

Language is an Algebraic Topology

A Topology on a set Ω_1 is a set T s.t.

1. $\emptyset \in T, \Omega_1 \in T$
2. $A,B \in T \rightarrow A \cup B \in T$
3. $A,B \in T \rightarrow A \cap B \in T$



- $S^{im}(h) = \{I : Hi(I) = h_i, i=1, \dots, K\}$
- $S^{ex}(z) = \{I : g(z;\theta)\ z \in \psi\}$
- $S^{Grammar}(A) = Language(A) = \{I : A \rightarrow_{AOG} I\}$

Let's examine algebraic topology and how it helps us understand language, vision, and the connection between language and vision.

What is topology?

Image space and language space are very extensive. Each of our concepts of the two is usually a subset. One million pixels is a one million dimensional space; each image is a point in it. Meanwhile, a face is a concept, and each one is a subset in this one million dimensional space. This subset and other subsets have a topological relationship. Computer scientists call it grammar, and it corresponds to the algebraic topology.

For example, there is a high probability that a head and neck in an image will be atop a pair of shoulders. The structure of image can be expressed as a kind of grammar, which itself can be described in an STC-AOG.

Grammar can be derived from language, which is a grammatical expression of the total collection of sentences. An STC-AOG is the overall expression of knowledge; each example is an STC-AOG derived from the space-time causal parse graph, or STC-PG. Computer vision uses it, language must use it, cognition uses it, and the robot in its task planning also uses it. This is a unified expression.

Section 8

Discipline 4: Game Theory and Morality - Acquiring and Sharing Human Values

To communicate with humans, a robot must understand human values.

Philosophy and economics have a basic assumption that a rational person's behavior and decision-making are driven by the maximization of his own interests. If we rule out the possibility of deceit, observing a rational person's behavior and choices allows us to reverse engineer his reasoning and learning and to estimate his values.

The concept of utility is a primary principle of modern decision theory: an agent makes rational decisions based on current beliefs and expected utility, known as the principle of maximum expected utility. We believe this simple yet powerful principle is an invisible "dark" force that governs the underlying mechanism of human behaviors. Studying utility should enable AI systems to understand observations more deeply and to generalize better.

By classic definitions, the utility derived from a specific choice is a utility function, a mathematical formulation that ranks the preferences of the individual such that U(a) > U(b), when the choice a is preferred over the choice b. By observing a rational agent's preferences, an observer might construct the utility function that represents what the agent is actually trying to achieve, even if the agent does not know it.

We can express values as a utility function with the symbol u. It usually consists of two parts:

(1) Loss or Reward. What is the net gain?
(2) Cost. How much do you expend to achieve it?

We can define this utility function with fluents. With every action we make, we are changing fluents. Going up in the space defined with u can be considered "appreciation." If we differentiate fluent vector f from function u, then we get a "field."

Applying concepts from calculus, we assume that in a period of time a person's value orientation is not contradictory. If he thought A were better than B, B were better than C, and then that C were better than A, his values are an incoherent "whirlpool." A field without "whirlpools" is called a conservative field. Its corresponding U is a potential function.

The aphorism "[w]ater, without control, flows down in nature; quite the contrary, humans strive hard to move up in the world" describes a difference in social phenomena and physical activities driven by the same essence. In other words, people and water are moving in accordance with their potential energy function. What is the potential energy function that drives people?

Values differ among people; even within a single person, values are always changing. This work does not discuss the social dimensions of values, but sticks with what might be called common sense values, things nearly all can agree upon, such as keeping a bedroom clean. This is a goal for which appreciation is easy to track.

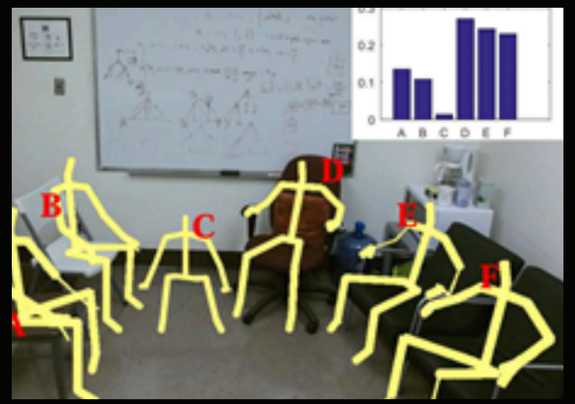
Shared Values:

Seating Choices

The figure below shows a simple experiment. We placed several different chairs and stools in my office (left) and in our lab (right). When my students entered each room, they each chose a chair to sit upon. They could have sat on the floor if they preferred.

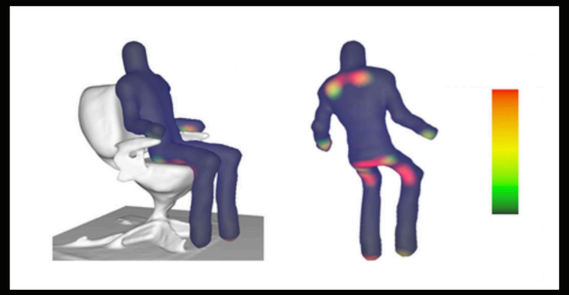
We labeled each chair A through F and watched the students' choices. Our goal was to determine why one chair was better than the others; what qualities led to it being chosen more frequently than others?

The answers reflect a basic value function.



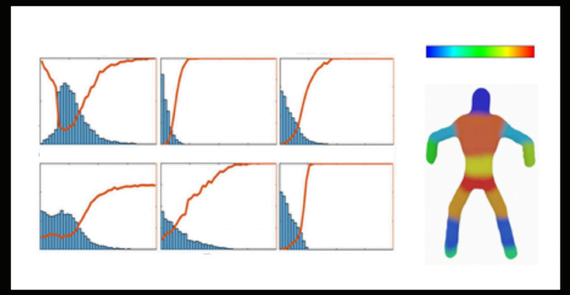
Pressure Distribution Drives Value

Calculating Stress Distribution



To answer these questions, Yixin Zhu and Chenfanfu Jiang (who was recently appointed assistant professor at the University of Pennsylvania), two of my doctoral students who were trained in physics and computer graphics, used the physics-based model in computer graphic form to simulate a man sitting on a chair in different postures. They then calculated the distribution of pressure from the chair on different body parts, including the man's back, buttocks, and head.

Visible Discomfort

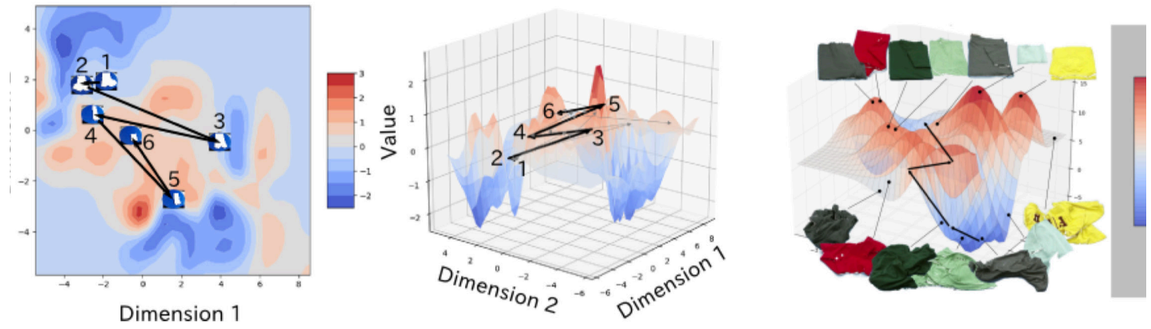


The blue histograms (above) graph pressure on six different body parts, and this information allows us to reduce the value function of each dimension. The six red curves on the graphs are negative value functions. When a person's seated posture results in the pressure falling below the red line, the person is achieving higher value according to their value function; in other words, the person is sitting comfortably. Not all people will have the same value function for comfortable sitting; some people prefer a soft sofa, while others enjoy sitting on a hard stool.

One cannot help but wonder if a similar dynamic applies to other potential physical functions, such as a gravity field. I believe it does. In fact, as we progress further in scientific understanding, we may find additional synergies between physics and other areas of science -- perhaps eventually the theories of Darwin and Newton will be unified.

These dynamics may be common sense for us, but robots must painstakingly calculate these common sense issues. It is not easy for them to put themselves in a person's shoes; the information needed to do so is not readily available to their "minds," like it is for humans. To determine in which chair humans prefer to sit would require knowing which chair most satisfies each person's value function – by placing the least discomfort on the sitter and placing them close to the person or object they are interested in. Meanwhile, the robot would experience sitting in the chair very differently. It will have to journey into the human's mind to imagine what these value functions are.

The Values of Folding Clothes



The process of folding clothes is another example of a common activity that illustrates value functions. If we visualize this conservative potential function as a topographic map, the process of folding a dress is like walking up a circuitous mountain path. Encounter a balled-up dress is like being at the base of the mountain, with the neatly-folded dress as the summit. Every step up the mountain, like every fold of the dress, results in a reward. Through mapping out the mountainous shape of the process of folding clothes, a robot could learn how to perform this intricate process. Through maps of human reward and appreciation, a machine can learn human values.

Much has been made recently about machines that can play Go. Go was invented in China more than three thousand years ago. Like chess, Go is a board game for two players; it differs in its objective, in which each player tries to surround more territory than the other. Machines' success at Go has provoked a mixed reaction.

To learn how to play complex strategy games like Go, machines must map out the value of every possible move, recognizing the potential cost and reward of each step. The machine needs a value function to judge if a move causes value to appreciate. Several other games played by reinforcement learning are quite popular, but studies based on these models are mostly in the simple symbolic space.

Competition and Cooperation - Getting Acclimated

In a multi-agent environment, value functions lead to competition and cooperation. As discussed in a previous section, social norms and ethics form, given the external physical environments and causal constraints, a state of balance in the population of agents in competition. Such states of balance are not fixed rules everyone is required to follow precisely, but rather a probability of behavior, a grammar for how agents will most likely act. In the end, we are still contending with probability, or an expression of STC-AOG.

In the process of social evolution, changes in conditions break old ethics and norms. In society, such changes take the form of technological innovation, policy changes, or other disruptions. New social norms then form, corresponding to another space-time causality diagram or STC-AOG.

If we were to take an STC-AOG representation from one balanced state to another unbalanced state, we would see it “acclimatized” to changes in conditions.

Two Categories of Learning

Inductive Learning

Induction is where the learner is given a series of examples and must intuit the rule connecting them. A large number of data samples from across a period of time, region, or population are observed. In the case of humans, these samples could include the millennia’s worth of culture and heritage. The result of inductive learning is a probability model of time, space, and causality expressed as an STC-AOG. Each space-time action is an STC-PG, or a parse graph.

Deductive Learning

Deduction is where the learner is given a value function (together with physics and causality), directly deriving the aforementioned balanced states. This is also an STC-AOG, and requires a profound, generative model and understanding of the subject of the study. By knowing the values that are pertinent to each task, the mind determines how to engage and interact with other minds.

Human learning is often a combination of induction and deduction. In youth, one does more inductive learning. Deduction is possible later on, when one comprehends the rules sufficiently and begins to develop credible maps of values, actions, and intent. Sherlock Holmes, for example, is a master of deduction because he understands the landscape built by the behaviors of those around him.

AlphaGo, the Go-playing AI program developed by Google’s DeepMind, first learned through induction: it was fed a large number of human-human games, from which it learned the rules. Once the rules had taken firm shape, it learned through deduction. But the mapped space representing all the possibilities for a game of Go in AlphaGo’s memory is nowhere near as spatially complex as the space required for human survival. And this mapped space does not take causal relationships into account. As a result, each move AlphaGo makes has a high degree of uncertainty, and is much more difficult.

Section 9

Discipline 5: Robotics - Constructing a Large Task Platform

In the fourth section, I discussed the "small data for big tasks" cognitive framework that should undergird AI development. Robotics, however, is a platform of large tasks. Not only entailing tasks such as visual recognition, language communication, and cognitive reasoning, robotics also requires the expenditure of considerable effort to change the environment. In this section, we will discuss robotics in terms of the common platforms available in the market.

As we have previously discussed, people and robots perform tasks. Tasks can be broken down into actions, and actions aim to change fluents in an environment.

We further divided fluents into two categories:

1. Physical Fluents

Such as painting, boiling water, mopping a floor, cutting vegetables. Tasks requiring dexterity; often performed alone.

2. Social Fluent

Such as eating, drinking, chasing, helping others; changing biological states and/or relationships; often performed as a group.

Affordance Maps in Planning

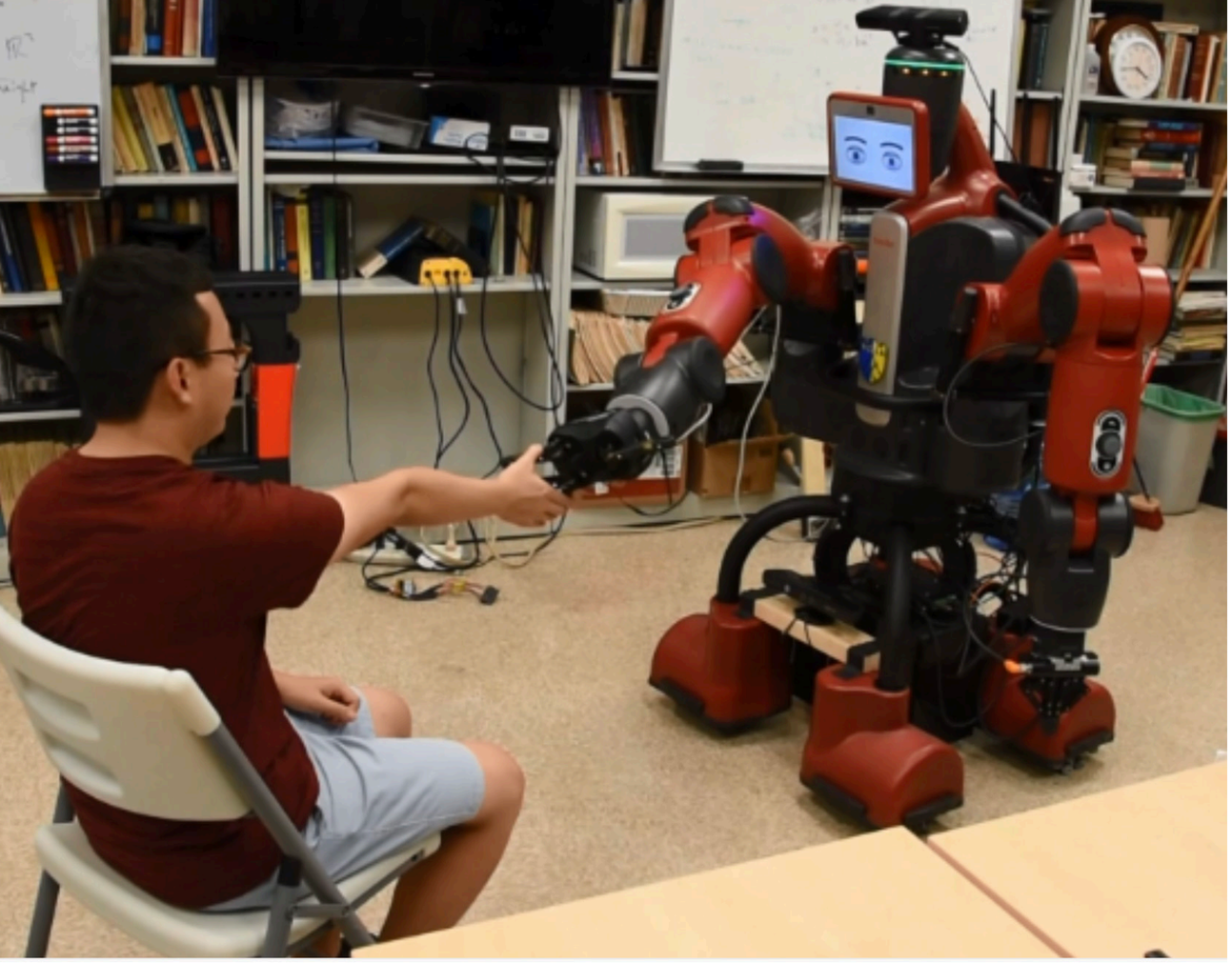
When a robot reconstructs a three-dimensional scene through functional reasoning, it focuses on current or potential tasks: where one might stand, where one might sit, where to pour water, or any number of others. The following figure shows a robot's assessments of where in a room someone could perform certain actions. This is an example of what's called, in robot planning, an "affordance map." It tries to answer the question, "What actions does this scene offer and enable?"

What can this scene give you and let you do?

With these maps of the single basic tasks available to it, the robot can plan a task. The plan itself is a hierarchical pattern of representation, which could be used in myriad ways. Here, I still represent it in the unified STC-PG. Creating this plan is a profoundly complex process because it requires a robot to take actions akin to monitoring and updating its scene to reflect changes in available tasks as a result of its actions. Robot performance of a task such as moving a box would then change these calculations by exposing another group of objects or making more tasks in the affordance map possible.

"The action plan should also consider the cause and effect, the actions and reactions of other agents in the scene. The more agents and environmental factors a robot can consider, the greater the care with which it can interact with a scene."

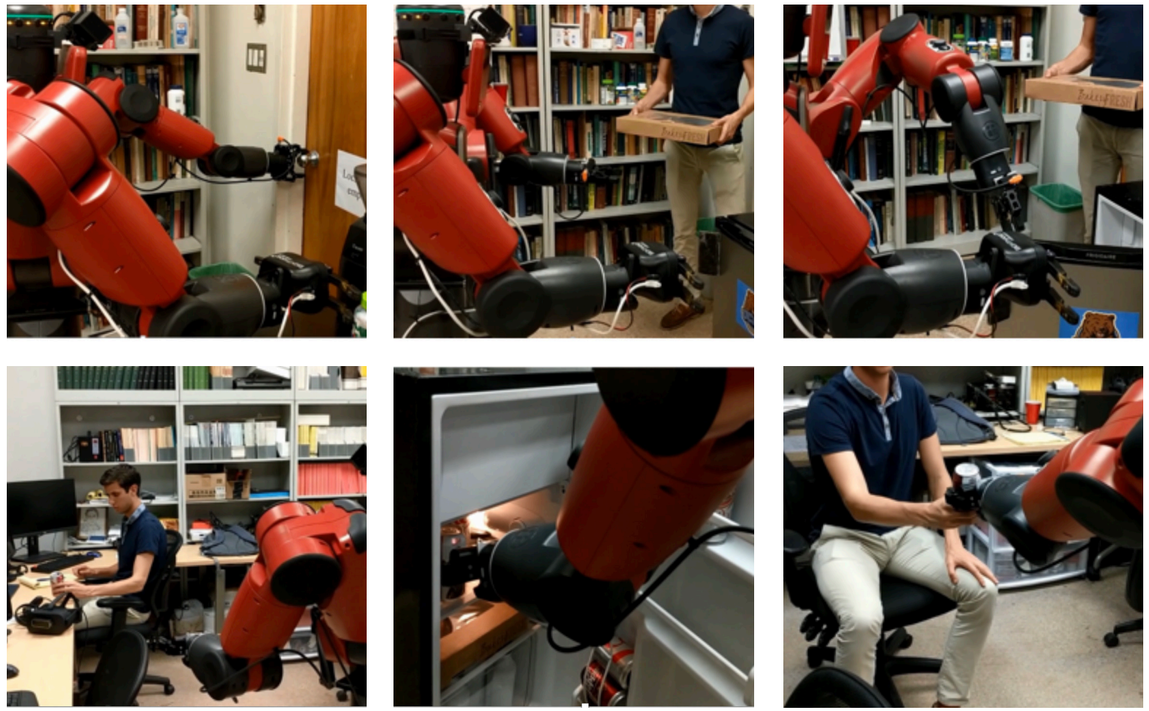
In the DARPA competition discussed in Section 1, the needs of perception and planning were handled by humans operating remote controls in the background. These are abilities we easily employ, without consideration for how complex and dynamic these abilities are.



Connecting With the Robot

In the picture above, doctoral student Tianmin Shu, in an early demonstration of my lab's work, teaches a robot how to shake hands. This is a ubiquitous but sneakily-subtle action; both sides need to be able to sense the intent of the other to avoid the dreaded fate of an awkward handshake. This demo was performed without a remote control and used a standard Baxter robot outfitted with an omnidirectional mobile base, two grippers (one flexible, one strong), and a handful of sensors and cameras. Note the parallel between the two kinds of grippers and certain creatures found in nature: lobsters have one heavy claw for crushing and one serrated claw for cutting. Shu's papers have received media coverage.

Completing a Comprehensive Task



Pictured above is the same robot in my lab completing a series of actions that make up a single complex task. First, it heard a knock at the door and inferred that someone outside the room wanted to enter. Then, it saw a person carrying a box of cake, which the robot interpreted to mean the person needed help of some kind. Through dialogue, the robot learned that the person wanted to put the box in the refrigerator. Finally, the robot opened the refrigerator door for the person to put the cake safely inside.

But the robot wasn't finished. The person sat down, picked up a can of soda, and after giving it a little shake, set it back down. By observing this action, the robot knew the can was empty (detecting an invisible fluent) and guessed that the person wanted another drink. It then went back to the refrigerator, opened the door, pulled out a soda, and handed it to the person.

Of course, this is a limited environment with a limited number of objects and only one other actor. If we were to apply this kind of functionality across scenes reliably, we would have to move closer to replicating a crow's reasoning using available objects in a complex series of behaviors while interacting with others.

Section 10

Discipline 6: Machine Learning - The Limits of Learning and Downtime

The five AI disciplines are groupings of similar kinds of problems. Throughout each section, I've tried to think about each discipline through a single framework, in hopes that we can eventually create a unified representation that addresses all of them.

Machine learning is designed to research and acquire the knowledge necessary to solve the previous five kinds of problems. The five other disciplines are the nails. Machine learning is the hammer.

Of all the hammers in use today, deep learning is particularly useful. Of course, within the five disciplines, there are many different kinds of tools and ways of using them employed today. But deep learning has, in recent years, been the most popular hammer of all.

Machine learning is hotly discussed and debated today. In this section, I will add my thoughts to the ever-unfolding conversation by considering the halting and limits of machine learning. "Halting" refers to the well-known Turing halting problem, which challenges one to determine whether a Turing machine, given a certain input, will stop itself, or run forever.

Based upon our cognitive framework,v;earning should be a continuous process of two-way communication. Given this starting point, under what conditions will the robot learning process terminate? This question is vital because when the learning process ends, no new information can be acquired for accomplishing new tasks. For humans, the learning process can sometimes stop quite early. People become less flexible as time goes on, resorting to older and less effective actions for completing the tasks of a new day. We don't want this to happen to AI systems.

What is Learning?

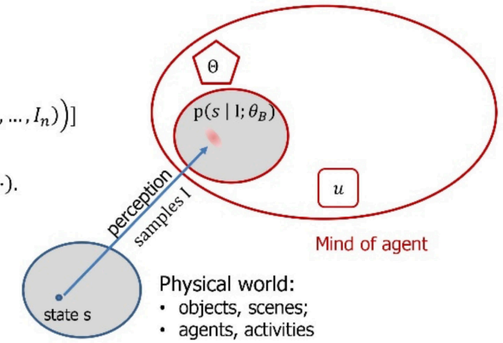
The current dominating stream (not the original one) of machine learning is very narrow. It doesn't represent the entire learning process and could be summarized in the three steps outlined below:

Limited by: *1. the concept class* Θ
2. sample size n

i.e. minimax lower bounds on

$$L(\Theta, n) = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathrm{E}_{\theta}[loss\left(\theta, \hat{\theta}(I_1, \ldots, I_n)\right)]$$

where utility $u_B(\cdot) = -loss(\theta, \cdot)$.



Goal: *a learner acquires concepts from i.i.d. data. It may involve latent state s.*

(1) Define a loss function written as u on behalf of a small task, such as face recognition; reward (+1) if right, and detract (-1) if wrong.

(2) Choose a model, such as a 152-layer neural network with hundreds of thousands of parameters that need to be fitted to the task through data.

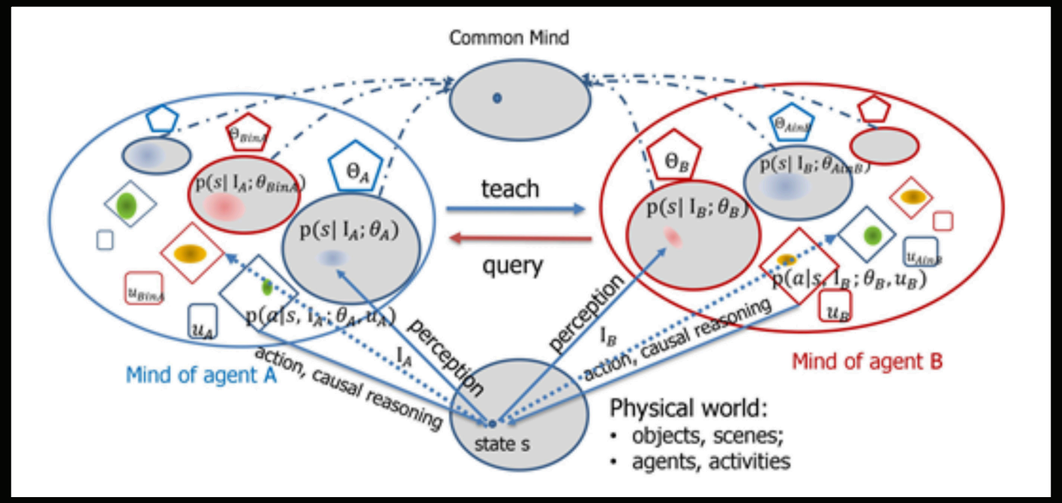
(3) Acquire a lot of data, and assuming someone annotated the data for you, begin to fit the parameters.

This process cannot learn cause and effect. It is purely passive, statistical learning. Nearly all visual and speech recognition systems today fall within these limits.

But real learning is interactive. Similar to the dialogues between Confucius and his students, learning occurs when students pose questions to their teachers, while teachers can pose problems to students, all in a free exchange of thoughts and ideas. Learning doesn't take place from the passive reading of an avalanche of questions and information. It's a two-way street; although I am a professor, I often learn new things from my students.

This learning process, based on the cognitive framework described earlier, is a kind of generalized learning that I refer to as "communicative learning."

Communicative Learning



The above figure describes the exchange between teacher (agent A) and student (agent B), each of whose minds is represented by a large colored ellipse. It's completely symmetrical, reflecting that teaching and learning are equal and bidirectional. Each mind contains three blocks: theta for knowledge, pi for a decision function, and mu for a value function. The small blue oval at the bottom represents the physical world, which contains all knowledge. The small blue oval on top represents the consensus reached by both the teacher (agent A) and the student (agent B) – their common mind.

This communicative learning architecture contains several learning modes. There are myriad learning modes yet to be developed, but the following seven modes correspond to one or multiple arrows in the figure.

7 Learning Models:

1 **Passive statistical learning:** Currently the most popular learning mode for machine learning, one in which large amounts of data fit a model.

2 **Active learning:** Where students take the initiative to ask teachers for the data they need, which was once popular in machine learning.

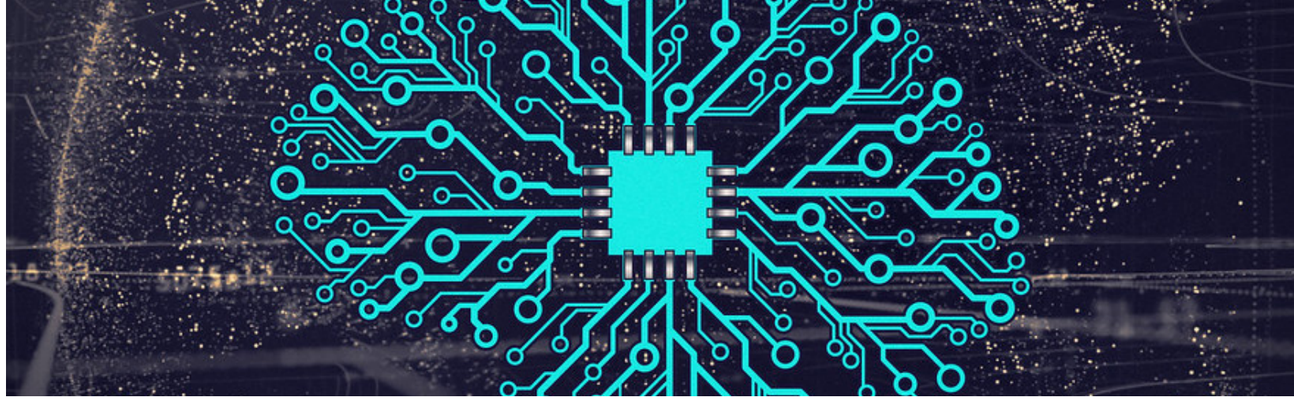
3 **Algorithm teaching:** Where the teacher designs examples to assist learning and tracks the learner's progress. This is ideal, but comes with a relatively high cost of effort.

4 **Learning from demonstration:** Commonly used in the robotics community, a strategy in which robots are taught to mimic the action performed by a person. A variant is imitation learning.

5 **Perceptual causality:** Learning through observing the cause and effect of others' behavior, all without experimental validation. This form of learning is common in human cognition.

6 **Causal learning:** This process enables the acquisition of a more reliable causal model through hands-on interventions, as opposed to perceptual causality, with controlled variables. Scientific experiments usually belong to this category.

7 **Reinforcement learning:** This is a method by which the mind learns a decision function and a value function through positive and negative reinforcement of a large number of actions.



Deep Challenges to the Process of Learning

As mentioned above, deep learning is but a small piece of the broader learning framework. Meanwhile, learning itself is only one discipline within AI. So to equate deep learning with AI is like a frog in a well trying to describe the sky based on the tiny patch it can see.

But what are the ultimate limits of the different forms of learning? What is the "shutdown condition"? In other words, when does learning end?

In passive statistical learning, there is an upper limit to the number of samples. But we want to move beyond passive statistical learning and consider limits outside its confinements. Can a broad learning process converge? And what is its convergence? The halting problem in machine learning is the challenge that occurs when the learning process stops.

Conversation in learning allows information to flow between two minds. It's what is taking place between the two ellipses in the figure on the previous page. Many factors affect the quality of this flow.

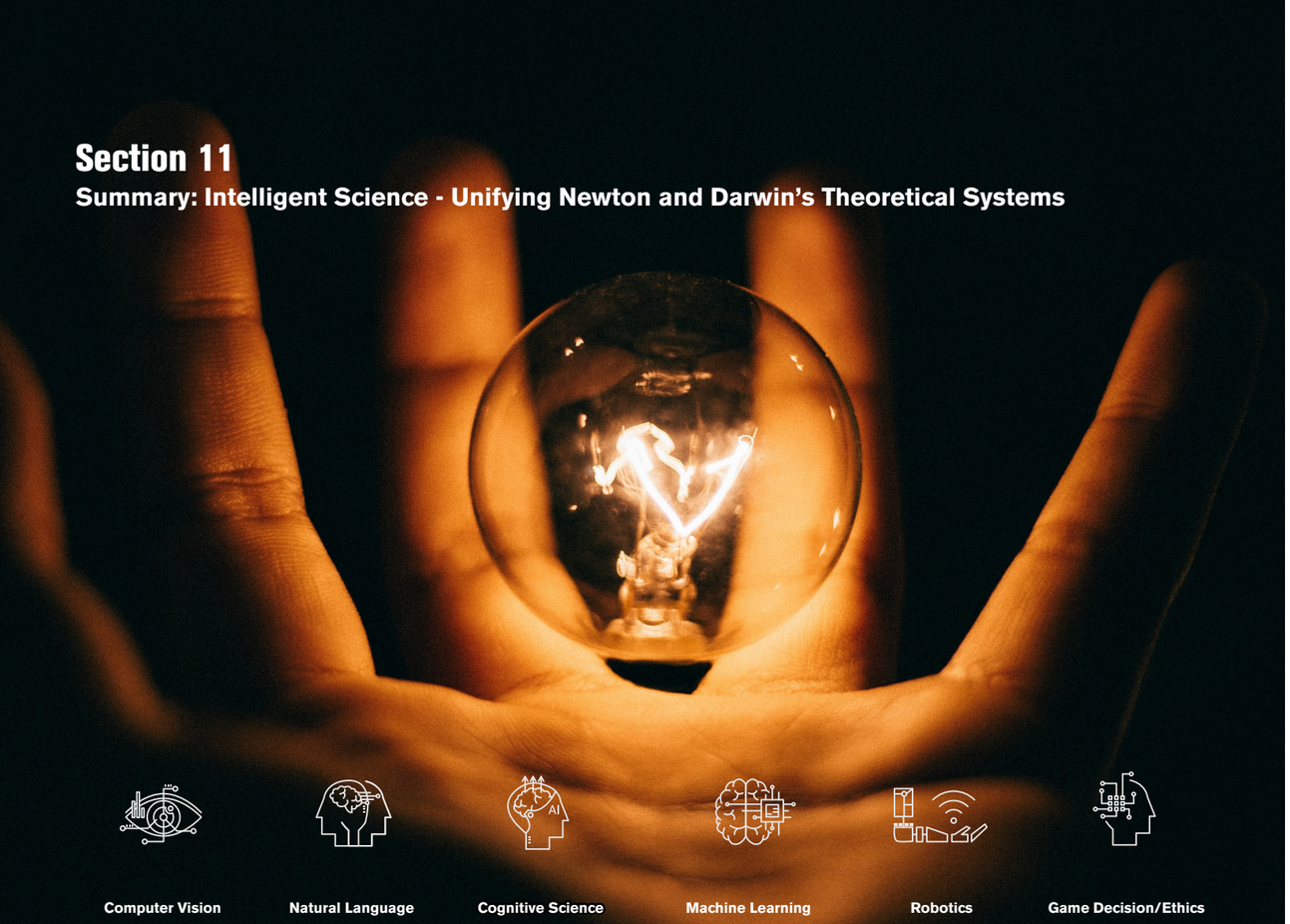
1 **Level of understanding of self and others:**
For teachers to impart knowledge, decision-making, and values to a group of students, they must be confident both that they have all the required knowledge and their students do not. Similarly, when students ask teachers questions, they must understand the overlap between what they don't know and what the teacher does. Both sides need an accurate estimate of themselves and of each other.

2 **Teaching and learning methods:**
If the teacher tracks students' progress, she can provide only knowledge that's new, rather than repeating herself. This is what's taking place in algorithmic learning and perceptual causality.

3 **The IQ Problem:**
How to measure the IQ of a machine? Many animals can't understand certain concepts regardless of how they are being taught.

4 **Value function:**
Students don't usually want to learn about things they aren't interested in. People of different values cannot communicate, let alone listen and learn from one another. For example, if a person in a Facebook group loses interest in its subject or topic, she will be tired of seeing news feed updates from it and leave the group behind.

There are 7.7 billion people in the world, and, among the 7.7 billion people, there are 7.7 billion different brain models. Despite the fact that there are some local shared models, building some small amount of consensus, learning conditions are different everywhere, and people will not all arrive at the same place.
The halting problem is really about how to reach a balanced state in this dynamic process.

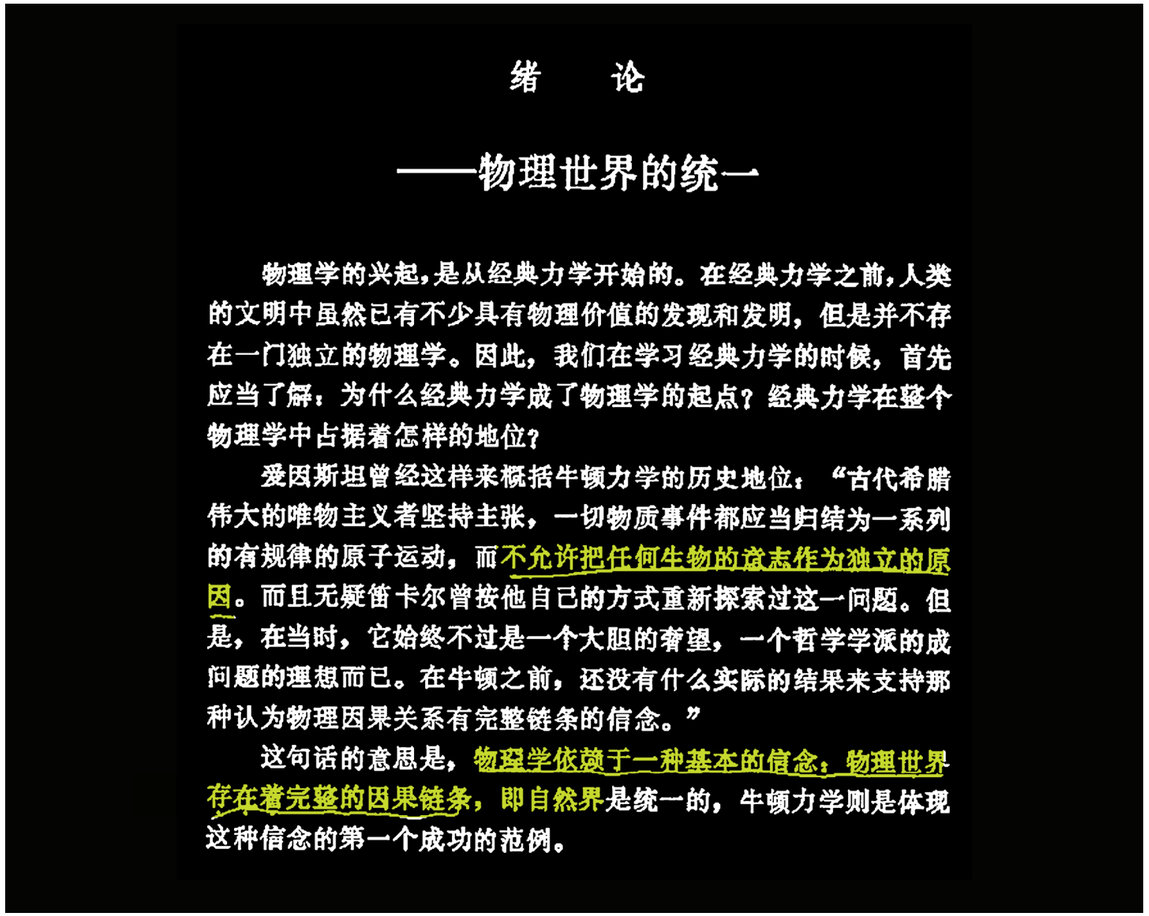


We have now explored the critical issues at the frontier of each of the six disciplines of AI, and I have shown why I believe they are unifying under a common cognitive framework. But how will a mature scientific discipline emerge from these six competing areas? With AI becoming the "science of intelligence," what should this unified scientific discipline be? In my opinion, physics is by far the most developed science; as such, its historical development can teach us about possible paths forward for AI.

"I loved physics so much in high school that when I applied to the University of Science and Technology of China (USTC) in 1986, I wrote down 'modern physics' as my preferred major. But when my brother saw the application form, he worried about the lack of careers in physics; though neither of us had ever seen a computer before, he changed my major to computer science without my knowledge. So I ended up in computer science by accident; but I've always kept a soft spot for the beauty of physics."

Four Sentences to Remember

When I started university, I noticed that my Introduction to Mechanics textbook had been authored by the university's executive vice president and his wife. This was an enduring memory for anyone who attended USTC. And the introduction in the book's very first page stunned me. Here are four sentences that stood and continue to stand out to me:



绪　论

——物理世界的统一

物理学的兴起，是从经典力学开始的。在经典力学之前，人类的文明中虽然已有不少具有物理价值的发现和发明，但是并不存在一门独立的物理学。因此，我们在学习经典力学的时候，首先应当了解：为什么经典力学成了物理学的起点？经典力学在整个物理学中占据着怎样的地位？

爱因斯坦曾经这样来概括牛顿力学的历史地位："古代希腊伟大的唯物主义者坚持主张，一切物质事件都应当归结为一系列的有规律的原子运动，而不允许把任何生物的意志作为独立的原因。而且无疑笛卡尔曾按他自己的方式重新探索过这一问题。但是，在当时，它始终不过是一个大胆的奢望，一个哲学学派的成问题的理想而已。在牛顿之前，还没有什么实际的结果来支持那种认为物理因果关系有完整链条的信念。"

这句话的意思是，物理学依赖于一种基本的信念：物理世界存在着完整的因果链条，即自然界是统一的，牛顿力学则是体现这种信念的第一个成功的范例。

"The history of physics is a history of the pursuit of the unity of the physical world. The first great unity is Newton's classical mechanics, where he, through gravity, established a unified interpretation of the movement of planets of the solar system and the movement of seemingly complex objects. The formation of a scientific system established a firm belief: that there is a complete causal chain in the physical world. The responsibility of physics is to find a unified force that governs various phenomena of nature." It is a dream, to be sure; one almost needs faith to believe in it. But if one does, it is a dream worth working towards. In over three hundred years since Newton's time, physicists are still working, ever so gradually, to discover a splendid universal model.

The Bag of Tricks Model

Compared to physics, AI research so far has paid little attention to the possibility of a unified model. But by resolving some small problems, a unified theory of AI has recently garnered more than considerable attention.

Some renowned professors in the 1980s believed there was no unified explanation for intelligence; rather, they saw intelligence as more like a “bag of tricks.” All intelligent beings did was apply designated problem-solving rules to their corresponding problems. But this perspective, in my opinion, is superficial and short-sighted.

When David Mumford left pure math to study AI in the 1980s, his vision was to build a mathematics of intelligence. This was quite a transition for such a well-known and distinguished mathematician. I have not seen another scientist in any field make such a big leap.

In my statement of purpose for my graduate school applications, I wrote that I wanted to explore a unified AI framework. There was no internet at the time, and I remember that the Department of Science and Technology had just replaced their obsolete dot-matrix printer with a laser model. I had never heard of Mumford.

Most graduate programs rejected me, but I was able to follow my advisor to Harvard. In the same year, Yingnian Wu, a fellow two years my junior from the same department at USTC, was admitted to Harvard to study statistics, and we became roommates. Wu’s understanding of physics and statistics was profound, and his knowledge has benefited me tremendously during my twenty-five years of working with him.

Looking back, I see how lucky I’ve been to have crossed paths with brilliant individuals like Wu at crucial junctures in my own research and learning.

- BERKELEY
- DEEPMIND
- STANFORD
- MIT
- DMAI

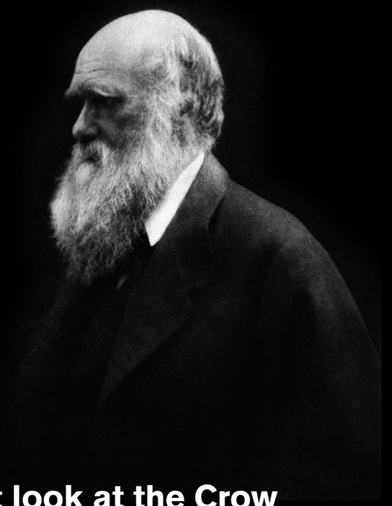
Looking back, I see how lucky I’ve been to have crossed paths with brilliant individuals like Yingnian Wu at crucial junctures in my own research and learning. Work of this great magnitude is rarely, if ever, advanced solely upon the shoulders of one person, which is why, in 2004, I created a non-profit institute in my hometown of Ezhou, Hubei, China, and named it the Lotus Hill Institute. In the following photograph, taken at the Lotus Hill Institute in 2005, we had a mission of collecting annotated image data, which marks the beginning of the data-driven paradigm and statistical learning in AI. Working shoulder to shoulder, we were successful.

From Big Data to Big Tasks

The version of AI we encounter daily is the AI of big data applied to master small tasks like image recognition, speech transcription, language translation. It’s AI that can play chess and video games. It’s parrot mode AI, not the AI of the future.

With ever-increasing realism and faster speed in rendering methods using dedicated hardware, the synthetic data from the virtual world is getting ever closer to the data collected from the physical world. In these realistic virtual environments, one can evaluate any AI method or system from a much more holistic perspective. We can use Virtual Reality using these simulation methods to measure whether a system is intelligent not only by its performance on a single task but through a combination of tasks: the perception of the environment, the planning of the actions, the predictions of other agents’ behaviors, and the ability to adapt learned knowledge to new environments for new tasks. To afford such a task-driven evaluation, physics-based simulation for multi-material multi-physics phenomena (see Figure 37) will play a central role.

Cognitive AI needs to accelerate by adopting more advanced simulation models from computer graphics to benefit from the capability of highly-predictive forward simulations. This acceleration will allow us to grow a crow, to go beyond deep learning, making us capable of integrating the dark aspects of cognition.



One last look at the Crow

Earlier we discussed how we study two basic aspects of physical and biological intelligent systems. To review:

First, the innate tasks and value function in intelligent species.
These are the bare necessities that we are evolved to fulfill. Animal behavior is driven by a range of survival tasks, determined by a value function. This value function is an evolutionary phenotype emerging from a landscape in which the fittest survive. Darwin proposed natural selection as a concept but did not explain it mathematically. Only later did we find that genetic mutation is actually the action of an entire species on its value function on an evolutionary time scale. The value function's topographic map I described previously is, in fact, borrowed from biology.

Second, the objective reality and causal chain in the physical environment.
This is the perception of the physical world and causal chains within it, governed by Newtonian mechanics. Ultimately, for artificial intelligence to become intelligent science, Darwin's and Newton's theoretical systems need to be unified.

Taking Flight

In 2016, I visited Westminster Abbey, where I was surprised to see the graves of Newton (1642-1727) and Darwin (1809-1882), two scientific geniuses who completely changed how we see the world, only a few meters apart.

How long must we wait before we unify their grand visions?

In his "Autumn Song," Tang Dynasty poet Liu Yuxi (772-842) composed lines that capture a bit of the imaginative leap we will need to take to advance this new realm of scientific research and innovation:

Autumn Song

by Liu Yuxi

Since olden days we feel in autumn sad and drear,
But I say spring cannot compete with autumn clear.
On a fine day a crane cleaves the clouds and soars high;
It leads the poet's lofty mind to the azure sky.

Contact

Address: 10940 Wilshire Blvd., 11th Floor, Los Angeles, CA 90024