Research Statement

Yu-Zhe Shi | CS@HUST

We always come up with *new* ideas to deal with the complex world. This seems natural, but too natural to answer *how* the ideas originate—sometimes hypotheses *just come out* when we explain causes for an effect. But there are some cues—our ways of thinking vary in person but efficient in common, while diverse individuals may share the same ideas. We explain differently due to diverse inner preferences; we think efficiently over ideas in proper organizing forms; our minds converge by sharing external knowledge and environments. The philosopher C. S. Peirce summarized thinking like these as ***Abduction***. However, current researches primarily formalize abduction as explanation, leaving the main problem out as stated below.

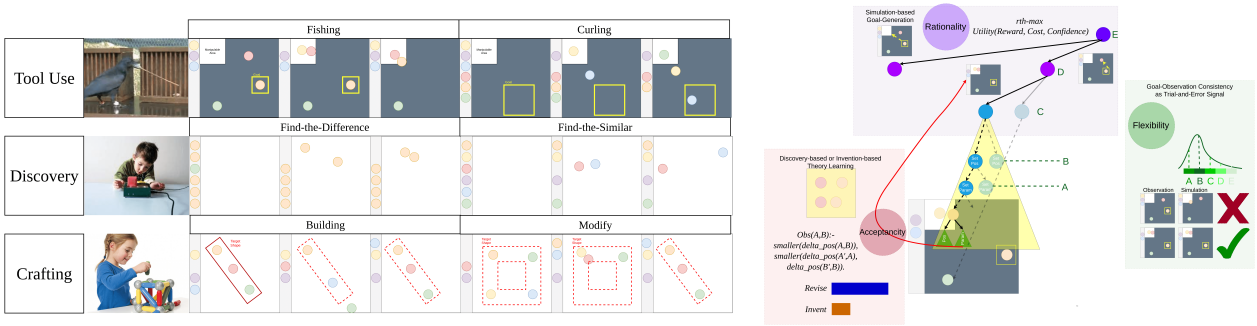
I study *how* to acquire abduction and explore *where* its upper limit is. The former requires looking into humans' development of abduction since childhood. The latter needs to reverse engineer how humans exploit abduction in different scenarios. Both aspects shed light on artificial intelligence (AI) research—I need to understand the mathematical infrastructures for machines to acquire abduction and define the challenging but significant machine abduction tasks. The former needs to formalize the structural commonsense such as recursion, chain, precondition, duplicate, reverse, identical (*i.e.*, analogy), *etc*. The latter requires analyzing the complexity and limitations in computation. More specifically, I describe three questions based on the assumptions about *origins of new ideas*:

Question 1 ***How do humans solve problems by relying on and adjusting inner preferences?***

Question 2 ***Why could humans explain efficiently without explicit probability calculations?***

Question 3 ***To what extent do the minds reach consensus across environments and knowledge?***

Taking the first steps to answer these questions would be the three research thrusts in my doctoral studies. My long-term goal is to develop a computational framework that explains the internal, natural, and external factors of abduction, and thus facilitates human-level AI being able to work efficiently in complex environments with rich and related ideas to solve problems.



(a) Different families of problems (b) Model pipeline

Fig 1: (a) The simulated environment mimics three kinds of problems in daily life. Circles denote objects, and areas with dark backgrounds indicate the inaccessible. These environments include three problem families: 1) tool use, such as fetching targets from inaccessible areas to manipulable areas with the help of other objects; 2) discovery, such as finding the only pair of objects that share the same in property, given a set of objects; 3) crafting, such as arranging objects to resemble a given shape. (b) Work-stream of the model.

Hypothesis 1 Humans trade-off over preferences to solve problems in the way of abduction.

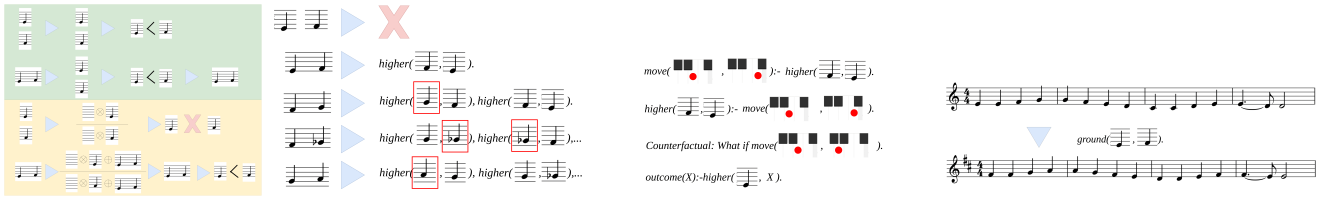
Humans exploit diverse solutions to solve a single kind of problem, whereas sometimes, we use similar strategies to solve different types of problems. However, current problem-solving mainly maximizes cumulative reward, ignoring the astonishing richness and diversity of solutions and strategies. I posit, alternatively, problem-solving is a series of trade-offs that should be evaluated in multiple aspects, *e.g.*, efficiency *vs*. quality of acquired knowledge. For example, a crow may trial flexibly when using a shorter stick to reach a longer stick for meat or persist in reaching it in hand; a child may try to explain the function of a black-box or just invent a meaning; a player may reach the target in the less costly way, or somehow behave irrationally. These compose the richness of problem-solving, and Laura Schulz reports similar phenomena in human play. Intuitively, these are linked to the Gestaltian Problem-Solving hypothesis—humans view problems in holistic representations and solve problems either by internal prior preference (reproductive) or by external objectives according to problems (productive, *i.e.*, insight).

Progress. Currently, I'm working with Dr. Yixin Zhu, Dr. Wang-Zhou Dai, and Dr. Song-Chun Zhu on human problem-solving as the first building block. First, I develop a playground to mimic different kinds of problems—a 2D physical environment with magnets, magnetic objects, and non-magnetic objects (see 1(a)). Even a single move in this world is non-trivial since both physical properties of the objects and inter-object interactions are all unknown to the player. Hence, one solves only by trial-and-error regardless of the problem family. Second, I elaborate a model on the hierarchical problem space indicated by *Rationality (R)*, *Acceptancy (A)*, and *Flexibility (F)*. Flexibility is a distribution over hierarchies indicating to what extent one revises her trials in a loop of intervention. Rationality reorders one's generated goals by the utility. Acceptancy determines whether to explain or to invent a theory given new observations. The three indicators compose a Gestaltian representation for every problem. The algorithm is bootstrapping a mind-simulator of the environment in essence—initialized with a non-magnetic physical engine, the agent takes consistency between simulated outcomes and observations as supervision signals,

learns concepts and theories with infinite models, then updates the simulator with learned physical properties for generating goals in next iteration (see 1(b)). I design a hierarchical evaluation protocol for problem-solving: 1) solution level: efficiency, times of inconsistency; 2) knowledge level: on-the-fly explanations and counterfactual explanations to test the learner; 3) representation level: intra-family and inter-family comparisons on diversity, consistency, novelty, *etc*.

Sketch. I will compare the model with human behaviors and current learning and planning methods. The evaluation contains two parts: 1) given fixed compositions of (R, A, F), rank them independently on multiple metrics; 2) initialize different compositions of (R, A, F), select one or multiple metrics as optimization objective, and observe the updating trajectories.

Significance. This work values not only for a generic computational modeling but also for experimental pipelines and evaluation protocols that may offer a new perspective in problem-solving research. Also, it is a testbed for further investigations into the three trade-offs jointly, facilitating more fine-grained behavioral studies and more advanced algorithms.



(a) Humans excel in relative confirmation (b) There is always a *higher* one (c) Qualitative theories (d) Ground a qualitative system

Fig 2: (a) One can easily tell a higher pitch against a lower pitch before she has trained to acquire perfect pitch (yellow background); however, an audio sensor percepts absolute pitches then differentiate them to figure out relative pitch (green background). (b) When listening to EF, one can tell that `higher(F, E)`—F is the higher, and G is even higher than F after hearing FG. (c) Qualitative theories that support counterfactual reasoning and currying of concepts. (d) Everyone is able to sing out the rhythm of *An die Freude* once listened to it. Most of them capture the relative rhythm, *e.g.*, with a modulation to C. However, they can ground the sentence to D given only a single evidence that E is F.

Hypothesis 2 Humans learn relative concepts and qualitative theories through abductive logic.

Conventional relative Bayesian Confirmation compares posterior of two candidates to infer their binary relationships, *i.e.*, inputs absolute quantitative concepts and outputs abstract relative relations. Humans, reversely, exploit relative relations to estimate absolute instances—they excel in percepting relative relationships such as greater or lesser, especially when prior is weak (see 2(a)). Most interestingly, relative relations are naturally infinite recursive (see 2(b)). Given a relative theory `r(A,B)`, there always exists another `r(A,C)` entailing `r(A,B)` and `r(B,C)`, echoing Dutch-Book property: However likely two competitive concepts are, there must be a constraint over them to distinct one from another. This may help integrate commonsense or subjective preference into Bayesian confirmation.

Qualitative theories describe the correlation of relative relations from different domains, *e.g.*, `the more the better`. Qualitative theories guide the estimation of physical parameters and support inaccurate-but-related counterfactual reasoning (see 2(c)). Once given quantitative observation of a single concept, the entire system is grounded (see 2(d)).

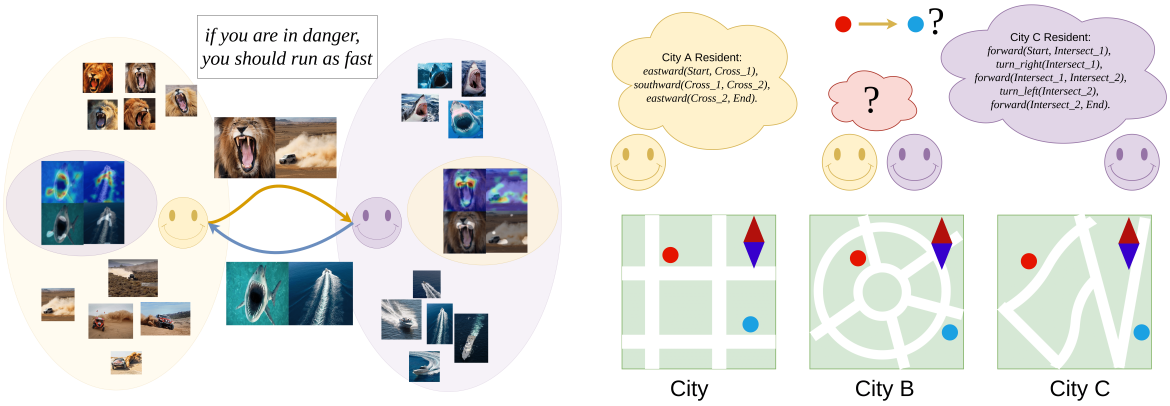
Sketch. The input of the program is perceptions of concepts. Constraints of relative theories come in as commonsense or additional signals of perception. The program drives an embodied agent and makes interventions of the world. The output of the program should be a qualitative simulator of the environment. Investigations should analyze Herbrand properties of the language and learn intuitive theories from visual and audio inputs with the programs.

Significance. Abductive logic has the potential to bridge Josh Tenenbaum's quantitative view and Benjamin Kuipers' qualitative view on physical modeling. This work should lead to further interest in mathematical foundations of relative knowledge representation. If we could understand how humans form relative concepts, we could build machines that reason over a complex system without simulating every detail of it.

Hypothesis 3 In abduction, knowledge aligns experiences and environments shape language.

Abstract knowledge is one's explanation of perceptions from physical environments and interpretation of other minds from social environments—individuals gaining experience in different environments acquire divergent knowledge yet similar in essence. Hence, aligning knowledge or sharing environment become two ways towards common ground. Take 3(a) for example, how can they reach the consensus of the deep causal relation?

There are more subtle cases such as linguistic pragmatics shaped by environments, *e.g.*, residents' way to refer directions correspond to cities' layout (see 3(b)). What is the minimal shared experience or minimal common knowledge required for successful communications? This would be a fundamental problem in computation.



(a) Knowledge aligns experiences

(b) Environments shape language

Fig 3: (a) One growing up in the land without knowing ocean refers to *if you are in danger then you run as fast* by showing two pictures, a tiger showing off sharp teeth and a jeep rushing dirt off sands; and her marine counterpart may use a shark opening mouth and a boat rushing in waves. Neuron activation maps visualize simulations of each other's minds. (b) In city A, people refer to directions with east or north, while they become left or forward in city C. But what language they would use at city B?

Sketch. In 3(a), the two agents share a causal theory as the only common ground. Trying to pair the visual feature sequence `sharp_teeth -> rushing_trace` and the causal theory `scary -> run` as guessing the sender's intention is practical in this case, empirically shown by my preliminary experiments. However, it is likely to fail in complex causal structures; when the messages themselves imply more structural information, such as curve painting and language 3(b). Hence, I should control the complexity of knowledge, perception, and the causal structure of the environment, respectively.

Significance. This work should inspire behavioral studies to combine dual-coding hypothesis and communication tasks, raise novel scenarios for learning theory analysis, and introduce challenging tasks to AI research that require knowledge abstraction from the environments, domain adaptation across environments, and out-of-domain generalization by knowledge alignment.

Personal Background

I am a senior undergrad majoring in computer science. I get top grades in maths such as Calculus, Linear Algebra, Statistics, Probability, Discrete Mathematics, and Statistical Machine Learning. I am inspired by great ideas in philosophy and sciences, including C. S. Peirce's Abduction and Pragmatism, K. Popper's Hypothetical-Deduction, D. Marr's Levels of Analysis, U. Grenander's General Pattern Theory, P. Thargard's Computational-Representational Understanding of Mind, J. Pearl's Causal Intervention, and M. Tomasello's Origins of Communications.

Research. My hypothesis on problem-solving inherits from a representative preliminary work. I model an agent to solve visual deciphering games and Minecraft puzzles with unknown objects—the agent succeeded given prior knowledge about dynamics of the world, solved most problems by maximizing consistency between knowledge and observations in E-M or MCMC iterations, and grounded most unknown objects to composed primitives. However, there is a gap between agents' solutions and their human counterparts—humans not only solve more problems in less time given less knowledge, but also with much greater diversity—and they can always explain why to behave like that, rejecting the hypothesis that unintentional randomness introduces diversity. Hence, I finally decided to rethink problem-solving under Gestalt productive hypothesis. Fortunately, the National Undergraduate Innovation Grants values this shift-of-idea and sponsors me to continue the work.

Skills. I have developed strong capabilities to derive and implement computational models in Logic Programming, Neural Programming, Optimization-based Programming, and Probabilistic Programming. I am also a full-stack software developer—I write both backend environments for algorithmic evaluations and frontend platforms for behavioral experiments. I am learning to design impeccable behavioral experiments with integrity and completeness.

Methodologies. My work begins at intuitive yet under-researched intelligent phenomena. A first-step-work on an abduction problem should be: 1) formalize the problem with plausible evaluation metrics; 2) engineer computational models and run behavioral experiments in parallel, where human results inspire computational improvement and model limitations guide subsequent behavioral studies; 3) analyze complexity and convergence of the model to understand *why* it works; 4) define the problem as an evaluable task for AI community. During doctoral studies, I plan to take the first steps in my three research thrusts and leave deeper investigations for future work. I value working together with psychologists, mathematicians, and AI researchers, ranging from human abduction to computational theories with AI applications that have a broad impact on downstream fields.