Research Statement

Yu-Zhe Shi

The conceptualization of hybrid human-machine intelligent systems that synergistically combine the strengths of human and machine intelligence, is a potent avenue towards achieving Artificial General Intelligence. Humans' distinct capabilities include the production of innovative insights, rational value judgement, and the flexible application of domain-specific tacit knowledge, collectively representative of the human ***mindset***. Regrettably, such merits are curtailed by our inherent cognitive bandwidth limitations. Conversely, intelligent machines excel in discerning a plethora of correlations across voluminous observational data. However, their decision-making ability is compromised when faced with multiple rational, safe and ethically sound interpretations, due to a dearth of physical and social common sense. This dichotomy becomes particularly pertinent within sophisticated but highly-significant domains such as scientific discovery, engineering development, professional education, public governance, and manned-unmanned device coordination in the real-world. In such scenarios, decision-making based purely on pattern recognition is inadequate, while humans' high-level insight, value, and implicit knowledge becomes indispensable. To facilitate super-human capabilities in these high-ending domains, it is both rational and practical to construct hybrid systems that highlight the strengths and circumvent the weaknesses of both humans and machines.

Hybrid systems have demonstrated substantial progress in the realms of visual analysis and intelligent chatbots, thus enhancing various aspects of human decision-making. Regrettably, such systems seldom explicitly and systematically incorporate the myriad variations of human mindsets and the possible configurations of the hybrid systems (see Fig. 1). Notably, in sophisticated applications such as science, education, and governance, human mindsets serve as the primary impetus behind hybrid systems by shaping *how to think*. Mindsets are inherently varied, owing to the individual's unique prior experiences, yet they may converge across individuals due to the shared necessity of addressing the problem at hand. Within a collective, diverse mindsets offer a multitude of perspectives, fostering a more profound understanding of a problem, yet they have to be calibrated to a common ground to facilitate communication between individuals with differing mindsets. Hybrid systems ought to promote the emergence, propagation, and synthesis of ideas from individuals and crowds, irrespective of their mindsets. Therefore, moving beyond merely optimising for convenience in interactions, we must explore for appropriate working environments that foster productive engagement between varied human mindsets and machines.

I propose to study how to construct high-fidelity communication channels between humans and machines in hybrid systems. This is with the aim of advancing sophisticated domains such as science, education, and governance, based on an understanding of the interplay between human mindsets and various configurations of hybrid systems, as viewed through a rationalist cognitive science perspective. This overarching problem can be considered hierarchically in accordance with Marr's level-of-analysis paradigm (Marr, 1982) (see Fig. 1), moving from conceptualized configurations, to three levels of analysis on abstract theories about human mindsets, and finally towards desired hybrid system outcomes in a top-down manner.

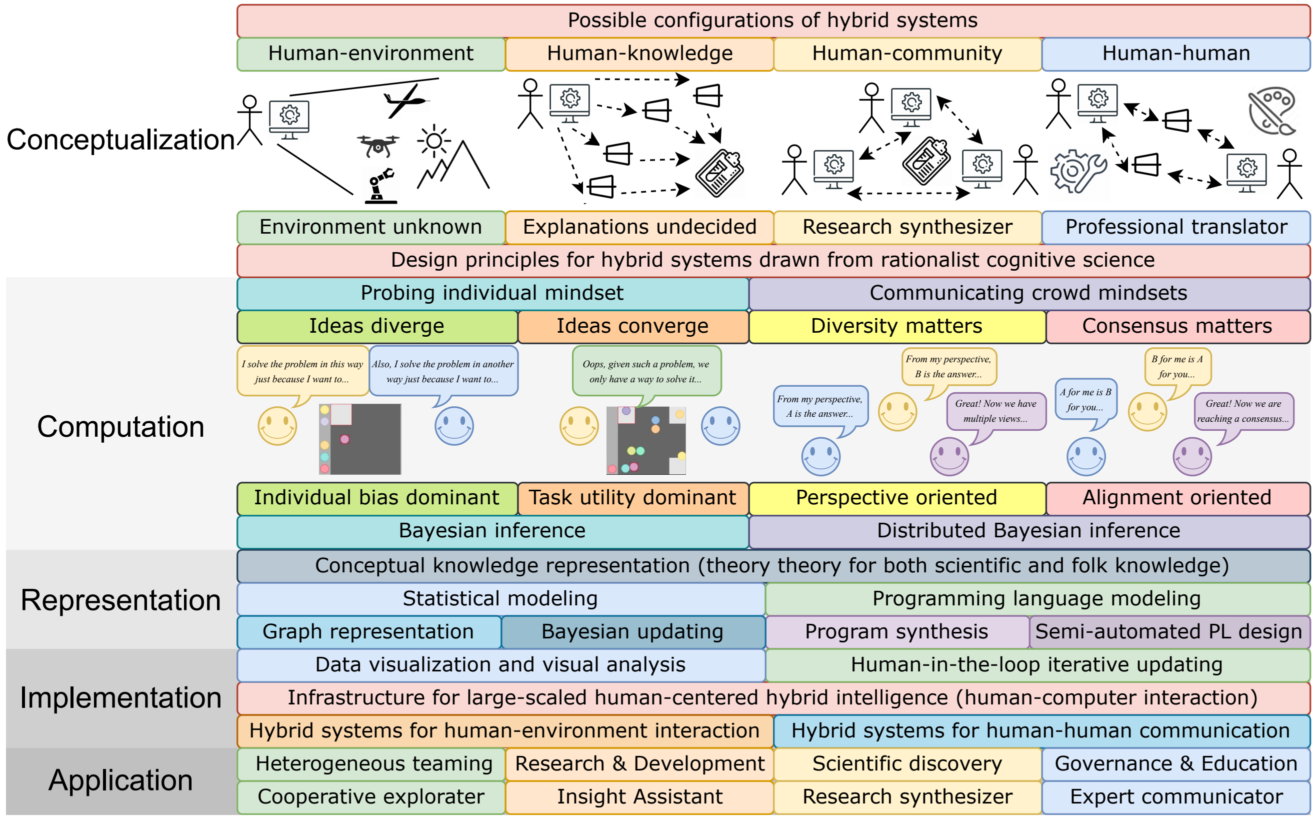

Possible configurations of hybrid systems
Human-environment
Human-knowledge
Human-community
Human-human
Conceptualization
Environment unknown
Explanations undecided
Research synthesizer
Professional translator
Design principles for hybrid systems drawn from rationalist cognitive science
Probing individual mindset
Communicating crowd mindsets
Ideas diverge
Ideas converge
Diversity matters
Consensus matters
Computation
I solve the problem in this way just because I want to...
Also, I solve the problem in another way just because I want to...
Oops, given such a problem, we only have a way to solve it...
From my perspective, A is the answer...
From my perspective, B is the answer...
Great! Now we have multiple views...
A for me is B for you...
B for me is A for you...
Great! Now we are reaching a consensus...
Individual bias dominant
Task utility dominant
Perspective oriented
Alignment oriented
Bayesian inference
Distributed Bayesian inference
Representation
Conceptual knowledge representation (theory theory for both scientific and folk knowledge)
Statistical modeling
Programming language modeling
Graph representation
Bayesian updating
Program synthesis
Semi-automated PL design
Implementation
Data visualization and visual analysis
Human-in-the-loop iterative updating
Infrastructure for large-scaled human-centered hybrid intelligence (human-computer interaction)
Hybrid systems for human-environment interaction
Hybrid systems for human-human communication
Application
Heterogeneous teaming
Research & Development
Scientific discovery
Governance & Education
Cooperative explorater
Insight Assistant
Research synthesizer
Expert communicator

Fig 1: Conceptualization, three levels of analysis, and application of hybrid systems from a rationalist cognitive science perspective

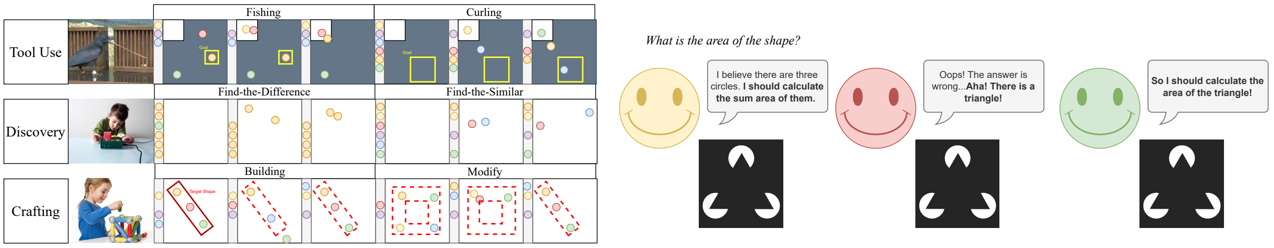
At the **computational level**, I will examine both the individual and inter-individual interaction modes, acknowledging the existence of diverse mindsets. For individual-machine hybrid systems, I ask: How does mindset dictate the interaction between personal bias and problem utility, such that solutions diverge for some problems but converge for others? For the crowd-machine hybrid systems, I consider: (i) How to take advantage of the diverse mindsets to synthesise a deeper understanding of the target problem? (ii) How can a common ground be established through communication, to facilitate idea propagation amongst diverse mindsets? Moving to the **representational level**, I intend to employ statistical models and programming language libraries the as the communication interfaces between human mindsets with a variety of hybrid system configurations. Lastly, at the **implementational level**, I aim to apply theories drawn from the two higher levels as guiding principles in the construction of hybrid systems addressing significant real-world applications.

Computational level

For the exploration of the computational level in individual-machine hybrid systems, I envisage two primary scenarios: (i) the dynamic interplay between mindsets and task utility (Shi et al., 2022b, 2023c,a); (ii) the interaction between mindsets and the domain-specific knowledge inherent to the problem (Shi et al., 2022c). Such scenarios mimic the exploration in uncharted environments and the execution of scientific discovery. In crowd-machine hybrid systems, I propose a pair of scenarios: (i) the communication between differing mindsets orientated towards the same objective; (ii) the fusion of multiple perspectives aligned towards the same objective (Shi et al., 2023b). Such scenarios mimic the processes of idea propagation and synthesis, particularly within the realm of scientific research and professional education.

Individual-bias-dominant *vs.* task-utility-dominant Evidence from cognitive and developmental psychology suggests that individuals solve novel problems by deploying high-level strategies based on context-agnostic prior knowledge. I posit that such strategies are constructed by assigning analogical semantics to the problem's elements, establishing connections with prior experiences. This assignment can be construed as Bayesian inference, taking goals and constraints as priors, and hence implies the diversity and convergence of strategies employed by individuals. To comprehend and model this human capability, I propose the **ProbSol Worlds (ProbSol)** environment (see Fig. 2(a)). This environment, whilst governed by consistent dynamics, can be configured for a range of tasks, such as tool usage, causal inference, and sketching. The incorporation of magnetism-based dynamics within ProbSol reduces the confounding variables of prior semantics encountered in traditional physically-grounded problem-solving tasks. Consequently, the environment-agnostic prior of goals and constraints can be separated from the act of problem-solving. ProbSol possesses substantial potential for conducting large-scale behavioural studies and for benchmarking computational models, thereby probing the diverse understandings of goals and constraints in problem-solving amongst both people and machines.

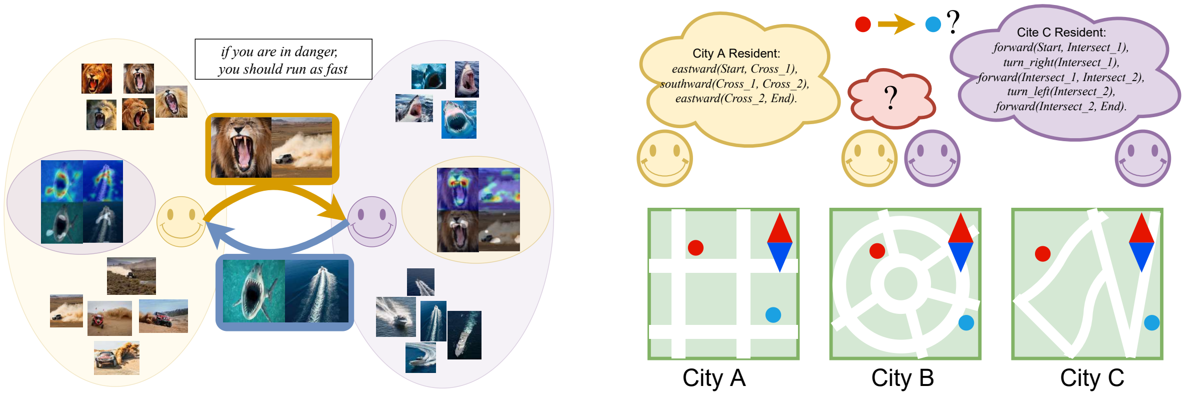
Insight-seeking *vs.* domain-knowledge-relying If scientific discovery is one of the main driving forces of human progress, insight is the fuel for the engine, which has long attracted behavior-level research to understand and model its underlying cognitive process. Unfortunately, extant tasks abstracting scientific discovery tend to concentrate on the emergence of insight, often sidelining the crucial role played by domain-specific knowledge. In this line of research, I perceive scientific discovery as an interaction between *thinking out of the box*—an active pursuit of insightful solutions—and *thinking inside the box*—a generalisation based on conceptual domain knowledge to maintain correctness. Consequently, I propose **Mindle** (see


Tool Use
Fishing
Curling
Goal
Discovery
Find-the-Difference
Find-the-Similar
Crafting
Building
Modify
Target Shape
What is the area of the shape?
I believe there are three circles. I should calculate the sum area of them.
Oops! The answer is wrong...Aha! There is a triangle!
So I should calculate the area of the triangle!

(a) The emergence of semantics

(b) The interplay between insight and knowledge

Fig 2: (a) Overview of the ProbSol Worlds. The simulated environment mimics physically-grounded problems in daily life. Circles denote objects, and areas with dark backgrounds indicate the inaccessible. These environments include three problem families: (i) tool use, such as fetching targets from inaccessible areas to manipulable areas with the help of other objects, reflecting the capability of meta-tool use and planning in both human and intelligent animals; (ii) discovery, such as finding the only pair of objects that share the same in property, given a set of objects, reflecting the people's capability of causal inference in human; (iii) crafting, such as arranging objects to resemble a given shape or transform a shape to another, reflecting the capability of concept abstracting and sketching in human. In these scenarios, magnetism plays totally different roles—helpful tool in tool use, clues and experimental materials in discovery, and harmful destroyer in crafting. **(b) Overview of insight in scientific discovery.** In a classic Gestalt problem, a problem solver first uses domain knowledge to analyze the problem, then seeks for insight once she gets trapped; after reconstructing the problem representation, she again uses domain knowledge to reach the solution. In this case, though domain knowledge constrains the thinking, it serves as the vehicle toward the target.



(a) Knowledge aligns experiences

(b) Environments shape language

Fig 3: (a) Knowledge aligns experiences. One growing up in the land 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) Environments shape languages.** 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?

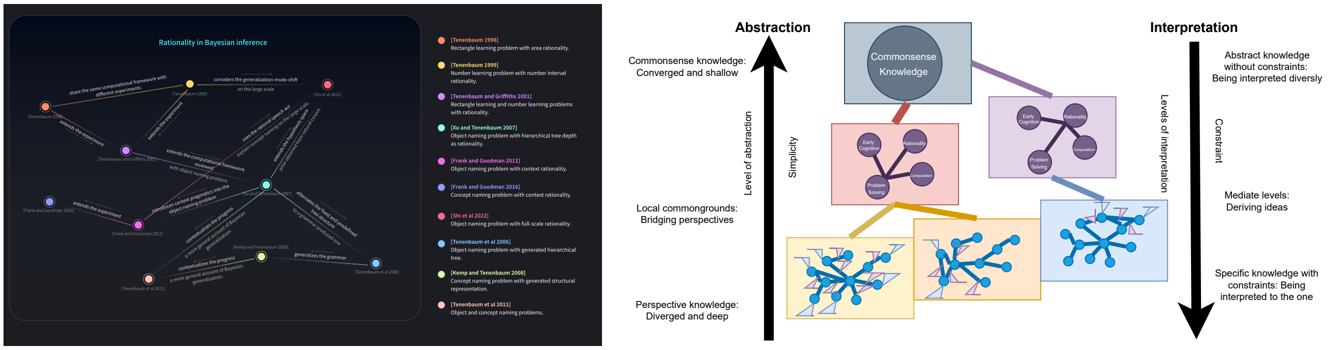
Fig. 2(b)), a semantic search game designed to spontaneously trigger scientific-discovery-like thinking. Mindle serves as a platform for large-scale exploration of scientific discovery, making feasible the reciprocal investigation of meta-strategies for insights and concept usage. Preliminary studies have revealed intriguing observations, sparking the formulation of elaborate hypotheses regarding meta-strategies, contextual factors, and individual diversity for more extensive investigations.

Communicating between diverse mindsets Abstract knowledge pertains to an individual's interpretation of physical perceptions from their environment and understanding of others' mindsets within social contexts. As individuals gather experiences in diverse environments, they acquire divergent yet essentially similar knowledge. Consequently, aligning knowledge (see Fig. 3(a)) or sharing environments (see Fig. 3(b)) present two viable approaches towards establishing common ground. However, such communication modelling necessitates a revision of the existing pragmatics model, which is predicated on the ideal assumption of shared world models between the speaker and listener. Essentially, it is assumed that both parties map object space to the semantic attribute space in an identical manner. Yet, this prerequisite is often unfulfilled in real-world communications, with individuals' world models diverging due to differences in their backgrounds, concept comprehension, and pragmatic understanding. They may, however, share an abstract semantic space that serves as a bridge between their disparate worlds. Therefore, I perceive this issue of speaker-listener coordination as an extension of cross-domain generalisation. The defining insight that distinguishes this coordinated cross-domain generalisation from its original form is the reciprocative learning process that it engenders, rather than a passive one. Hence, from a systemic perspective, the divergence between the two world models is not an obstacle but rather a catalyst for advancement.

Combining multiple perspectives given diverse mindsets Understanding a complex concept through myriad perspectives can be formulated as a distributed Bayesian inference problem (see Tab. 1 for specifics). Firstly, due to limited cognitive resources, a researcher cannot fully grasp a high-dimensional piece of knowledge, which aligns with the *curse of dimensionality* in density estimation. In this regard, let $P(x)$ denote the marginal density of knowledge $x \in \mathcal{X}$ (where $\mathcal{X}$ is the high-dimensional space encompassing all modalities of knowledge), direct estimation of $P(x)$ is inherently infeasible. Secondly, the existence of diverse mindsets amongst researchers, stemming from differing expertise, experiences, backgrounds, and interests, can be viewed as a prior z in a latent space $\mathcal{Z}$, which encompasses all potential perspectives. Here, $P(z)$ represents the distribution of these perspectives. Thirdly, the joint density $P(x, z)$ serves as a representation of human-centric knowledge, which can be generated via $P(z)P(x|z)$ in line with Bayes' theorem. In this relationship, $P(x|z)$ depicts the interpretation of knowledge x, conditioned upon perspective z. Finally, given adequate diversity amongst researchers, and provided we can amalgamate these perspectives, we can achieve a more profound understanding of a piece of knowledge. Formally, this amalgamation of perspectives equates to the marginalisation over $P(x, z)$, *i.e.*, $P(x) = \sum_z P(x, z)$. Unlike conventional research synthesis, this approach explicitly models the generation, interpretation, and amalgamation process. However, this formulation adopts a static view of knowledge. When introducing a temporal dimension to perceive the evolution of science as a dynamic process, it is possible to incorporate the propagation of perspectives into graph dynamics as a constraint.

Table 1: The analogy between perspective synthesis and distributed Bayesian inference

	challenge	prior	generation	merge
Bayesian inference	intractable $P(x)$	$P(z) \in \mathcal{Z}$	$P(x, z) = P(z)P(x\|z)$	$P(x) = \sum_z P(x, z)$
perspective synthesis	limited bandwidth	diverse prior	diverse interpretations	in-depth understanding


Rationality in Bayesian inference
share the same computational framework with different experiments.
considers the generalization-mode-shift
on the large scale
[Tenenbaum 1998]
[Tenenbaum 1999]
[Shi et al 2022]
extends the experiment
extends the experiment
[Tenenbaum and Griffiths 2001]
extends the computational framework
reviewed
with object naming problem
uses the rational speech act
explain concept naming on the large scale.
extends the hypothesis space
prior obtained from natural corpus.
[Xu and Tenenbaum 2007]
[Frank and Goodman 2016]
extends the experiment
introduces context pragmatics into the object naming problem
[Frank and Goodman 2012]
alternates the fixed and predefined tree structure
to a grammar produced one
contextualizes the progress
a more general account of Bayesian generalization.
[Kemp and Tenenbaum 2008]
generalizes the grammar
[Tenenbaum et al 2006]
contextualizes the progress
a more general account of Bayesian generalization.
[Tenenbaum et al 2011]
[Tenenbaum 1998]
Rectangle learning problem with area rationality.
[Tenenbaum 1999]
Number learning problem with number interval rationality.
[Tenenbaum and Griffiths 2001]
Rectangle learning and number learning problems with rationality.
[Xu and Tenenbaum 2007]
Object naming problem with hierarchical tree depth as rationality.
[Frank and Goodman 2012]
Object naming problem with context rationality.
[Frank and Goodman 2016]
Concept naming problem with context rationality.
[Shi et al 2022]
Object naming problem with full-scale rationality.
[Tenenbaum et al 2006]
Object naming problem with generated hierarchical tree.
[Kemp and Tenenbaum 2008]
Concept naming problem with generated structural representation.
[Tenenbaum et al 2011]
Object and concept naming problems.
Abstraction
Commonsense knowledge: Converged and shallow
Level of abstraction
Simplicity
Local commongrounds: Bridging perspectives
Perspective knowledge: Diverged and deep
Commonsense Knowledge
Early Cognition
Rationality
Problem Solving
Computation
Early Cognition
Rationality
Problem Solving
Computation
Interpretation
Levels of interpretation
Constraint
Abstract knowledge without constraints: Being interpreted diversly
Mediate levels: Deriving ideas
Specific knowledge with constraints: Being interpreted to the one

(a)

(b)

Fig 4: (a) Showcase of the rationality perspective for Bayesian inference (visualized by Shi et al. (2023b), a tool for generating frames of perspectives for research); **(b) Illustration of perspective and commonground as two ends of human scientific knowledge**

Representational level

In studying the communication interface between humans and machines, I first define the representation of external conceptual knowledge. I posit that the communication channels within hybrid systems should be conceptualised as operations over this knowledge representation. Two modes of representation are suggested: (i) graph representation and (ii) programming language library representation. Thereafter, I explore the interplay between perspective and common ground, in addition to their roles within the communication between mindsets.

Graph representation

Conceptual knowledge deftly amalgamates both declarative and procedural knowledge, comprising not only facts concerning the concepts but also active processes illustrating how these concepts interact. While numerous viewpoints exist regarding concept representation, I adopt the theory theory as a foundational premise given its acceptance within both scientific knowledge representation and folk psychology (Gopnik and Wellman, 1994). This allows the formulation of domain knowledge in the vein of theory theory. A practical implementation of theory theory posits concepts within a fully-connected network wherein each concept correlates with every other, suggesting each concept can potentially be influenced by all others. This network exhibits high flexibility, permitting various calculi on fully-connected graphs to be employed in formalising operations over concepts. For instance, general pattern theory could be applied, viewing all nodes as words (Grenander, 2012), or the representativeness of attributes could be used, considering all other concepts as attributes of the concept (Shi et al., 2023d). Corresponding well with generative modelling (Shi and Wu, 2022), the graph can be updated through sampling methods. If each node is viewed as a paper projected to a specific perspective, the graph can also represent a frame in scientific research (see Fig. 4(a)).

Programming language library representation

Predicated upon naturalism, the end-user of conceptual knowledge is fundamentally human, whilst the primary objective of conceptual knowledge is to accurately depict the world. Consequently, people engage with knowledge from various specific perspectives, even though the world depicted by such knowledge is singular. This raises a perceived contradiction between the human-centred demand and the content-centred nature of knowledge. Is it possible to strike a balance between these seemingly conflicting needs within any domain? Indeed, it is. Programming languages, fundamental tools for computation, effectively mediate between human users and the world. Specifically, a programming language serves two main functions: (i) it describes or models objects, scenarios, or systems, *e.g.*, 3D masks of human faces, biological experiment protocols, and electronic systems; (ii) it is designed to cater to users' needs and biases, as illustrated by the requirements for object-oriented programming, functions, and stream and procedures (Abelson and Sussman, 1996). The former is embodied by low-level instruction sets focused solely on content modelling, while the latter is represented by high-level programming languages that exclusively consider user requirements, under the assumption that all programming languages can be translated into instructions. The task of bridging these two ends is delegated to the compiler. Therefore, in knowledge representation, we similarly require compilers that translate human-centred perspectives towards a mediator capable of amalgamating or translating other perspectives. These compilers should encapsulate how the perspective is framed from the content, with the mediator providing a common ground that synthesises consensus from different perspectives. Initial efforts have been made to develop a framework that (i) formulates content-centred Instruction Set Assembly (ISA) proposals through statistical features derived from the natural corpus and refines the ISA by incorporating domain expert knowledge in a human-in-the-loop fashion; and (ii) enables human users to express their preferences regarding instructions to form a human-centred program library. This framework, already successfully implemented within the domain of biological experiment protocols (Shi et al., 2022a), holds promise for application across various domains, aiding in the design of domain-specific programming languages tailored both for human usage and for program synthesis.

Perspective *vs.* commonground: rethinking the value of diversity

If personal perspective represents a conditional worldview, a common ground, by contrast, signifies the consensus achieved by a group of individuals, independent of conditioning. In extremities, if the consensus group is sufficiently large, encompassing all global inhabitants, the common ground held by the group transitions to a universal common sense. We propose that perspective and common sense serve as two poles within the expansive knowledge sphere, with varying levels of common grounds – distinguished by group sizes – situated in between (see Fig. 4(b)). Knowledge adjacent to perspective is detailed, focused, and divergent across populations; knowledge closer to common sense is rudimentary, broad, and convergent across populations. Transitioning from perspective (lower levels of common ground) to common sense (higher levels of common ground) involves condensing knowledge to their fundamental and abstract states, thus eliminating condition-specific considerations and personal biases. Conversely, moving from higher to lower levels of common ground signifies the interpretation of knowledge, influenced by biases and constraints. This may reflect findings suggesting that semantic representations of abstract words are more diverse than those of concrete words; compared to their concrete counterparts, abstract words have a "longer journey" towards being grounded, indicating an increased range of interpretations. Similarly, considering the considerable diversity within the scientific community, we must acknowledge that personal biases can contribute to scientific shortcomings. However, such flaws occasionally become springboards for monumental breakthroughs in scientific history, stemming from thoughtful considerations based on both expertise and insight, as exemplified by the discovery of Kekule's structure. Kekule's perspective on organic chemistry was borne out of profound personal bias. Therefore, we should reconsider the value of diversity. Rather than attempting to eliminate all bias in science, we could explore ways of embracing those biases that can be transformed into perspectives, thereby leading to meaningful research directions. After all, while common grounds bridge gaps across scientific disciplines, facilitating communication and learning, it is perspective that propels the frontiers of science, pushing from the deep and narrow.

Personal statement

In my forthcoming doctoral trajectory, I aim to pioneer the application of rationalist cognitive science principles in the design of hybrid systems, specifically within sophisticated realms such as science, education, and governance. Given the complexity of this endeavour, I intend to initially focus on leveraging *VisLab*'s extensive expertise in these domains. Furthermore, I am keen to collaborate extensively with experts in psychology, neuroscience, artificial intelligence, and mathematics, to glean insights from well-established theories and intriguing empirical observations.

My research approach centres around investigating relatively unexplored yet intuitive intelligent phenomena. Introspective studies on hybrid systems should include: (i) formalising the problem with pertinent evaluation metrics; (ii) concurrently developing computational models and conducting behavioural experiments—human results would inspire computational enhancements, while model limitations would guide further behavioural studies; (iii) analysing the complexity and convergence of the model to discern *why* it proves effective; (iv) repurposing the problem as an evaluable task for both the AI and hybrid intelligence communities.

References

Abelson, H. and Sussman, G. J. (1996). *Structure and interpretation of computer programs*. The MIT Press. 4

Gopnik, A. and Wellman, H. M. (1994). The theory theory. In *The Society for Research in Child Development Meeting*. 4

Grenander, U. (2012). *A calculus of ideas: a mathematical study of human thought*. World Scientific. 4

Marr, D. (1982). *Vision*. W. H. Freeman and Company. 1

Shi, Y.-Z., Bi, Z., Zhang, Z., Zhu, Y., and Ma, J. (2022a). The AutoBio Instruction Set Assembly. The BioBirdEye Project Engineering Series. 4

Shi, Y.-Z., Hou, H., Wen, H., Zhang, C., Dai, W.-Z., Ho, M. K., Yang, Y., and Zhu, Y. (2023a). Abductive task abstractions in physical problem-solving without object-level prior. *Machine Learning*. In Minor Revision. 2

Shi, Y.-Z., Hou, H., Wen, H., Zhang, C., Ho, M. K., Yang, Y., and Zhu, Y. (2022b). Semantics emerge from solving problems given abstract prior. *Cognitive Science: ProbSol Concept Paper*. 2

Shi, Y.-Z., Li, S., Niu, X., Xu, Q., Liu, J., Xu, Y., Gu, S., He, B., Li, X., Zhao, X., Zhao, Z., Lyu, Y., Li, Z., Liu, S., Qiu, L., Ji, J., Ruan, L., Ma, Y., Han, W., and Zhu, Y. (2023b). PersLEARN: Research training through the lens of perspective cultivation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 11–30. Association for Computational Linguistics. 2, 4

Shi, Y.-Z., Muggleton, S. H., and Dai, W.-Z. (2023c). Object invention for abductive knowledge induction in the open world. *Machine Learning*. In Minor Revision. 2

Shi, Y.-Z. and Wu, Y. N. (2022). Generative Modeling Explained. In *Statistical Machine Learning Tutorials*. Department of Statistics, UCLA, Summer 2022 edition. 4

Shi, Y.-Z., Xu, M., Han, W., and Zhu, Y. (2022c). To think inside the box, or to think out of the box? Scientific discovery via the reciprocation of insights and concepts. *arXiv preprint arXiv:2212.00258*. 2

Shi, Y.-Z., Xu, M., Hopcroft, J. E., He, K., Tenenbaum, J. B., Zhu, S.-C., Wu, Y. N., Han, W., and Zhu, Y. (2023d). On the complexity of Bayesian generalization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31389–31407. PMLR. 4