# Function Alignment:
# A New Theory for Mind and Intelligence

*Part I: Foundations*

**Gus G. Xia**

Music X Lab, Machine Learning Department
Mohamed bin Zayed University of Artificial Intelligence

> *"The created universe carries Yin at the back and Yang in front, and through the union of the pervading principle, it reaches harmony."*
>
> *—Laozi, Dao De Jing, Chapter 42*

H uman perception operates across multiple levels of abstraction simultaneously. For example, when we listen to music, we perceive raw acoustic signals at the most basic level, interpret musical scores at a higher level, and recognize even more abstract structures such as chords and forms. Representations at different levels function like distinct languages, each with its own semantics. Yet, these levels of representation influence one another, and we rely on such interactions to better understand and predict the world. Consider music again as an example: a trained musician with theoretical knowledge can better anticipate upcoming low-level acoustic events, while an improviser attuned to the nuances of low-level musical flow can make more informed decisions about which note to play next.

To model such entangled dynamics of hierarchical representations during perception, I propose **function alignment** as a new **theory of mind and intelligence** in this position paper. Note that this is not a technical paper that introduces concrete solutions but rather a methodological perspective.

As shown in the graphical model in Figure 1, the $y$-sequence represents the true dynamics of physical reality, the $x$-sequence captures the low-level representation in the human mind, and the $z$-sequence corresponds to a higher-level, more abstract representation. While additional layers of representation may exist above $z$, we limit our illustration to two levels for clarity.
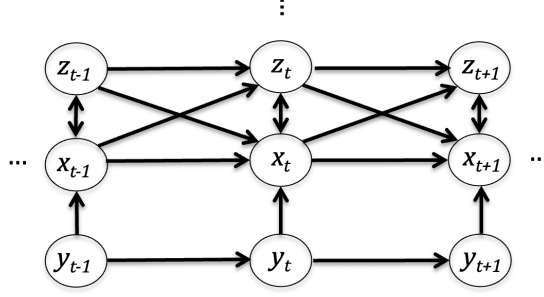
1

Figure 1: An illustration of function alignment: $x$-$z$ is function aligned, while $x$-$y$ is not.

In this framework, the dynamics of $x$ and $z$ are defined as function aligned, characterized by the following three key properties:

1. **Both $x$ and $z$ serve as functional descriptions of $y$**—they encode different levels of abstraction of the same underlying reality.

2. **Both $x$ and $z$ are auto-regressive processes, dynamically influencing each other's predictions.** Unlike typical hierarchical models where higher layers passively summarize lower ones, here $x$ and $z$ actively "listen" to each other, forming a bidirectional alignment.

3. **$x$ and $z$ are aligned in time,** where "time" refers not strictly to physical time but to a generalized logical sequence that governs inference and decision-making.

The second property, represented by the diagonal cross-connections, is the most critical distinction between function alignment and conventional hierarchical time-series models. These cross-level, cross-time influences enable mutual adaptation rather than one-way abstraction.

Notably, $x$ and $y$ are not function-aligned, as indicated by the absence of diagonal connections—$x$ merely passively models the underlying reality $y$ without influencing it. For instance, Newtonian mechanics provides a macro-scale description of physical reality based on underlying microscopic particle interactions, but the law $F = ma$ does not affect subatomic physics. Another example of a non-function-aligned hierarchical structure can be seen in programming: a Python program is interpreted into lower-level C-executable behavior, but the relationship is unidirectional. Such "concealed" hierarchical structures arise from a lack of function alignment. In contrast, the dynamics of the $x$-sequence and the $z$-sequence are deeply entangled and aligned—they are both perceptual representations, interconnected through neural mechanisms, and capable of shaping each other's evolution.

# Neural Hierarchical Representation and Symbolic Thinking

A plausible way to implement the interactions in Figure 1 is through neural networks. As previously mentioned, we illustrate only two levels of representation, but in reality, the hierarchy could extend far beyond $z$. The vertical bidirectional arrows between different levels of representation (e.g., $x$ and $z$) can be interpreted as an encoding-decoding mechanism, where higher-level representations are more abstract.

The encoding process inevitably leads to a loss of detail. At some sufficiently abstract level, representations reach what Hofstadter, in *Gödel, Escher, Bach* (Hofstadter, 1999), calls the "crystallization point", where they transition into symbolic forms and their dynamics become rule-based. This hierarchical transition from abstraction to symbolism is conceptually similar to Escher's artwork, where homogeneous, meaningless sub-symbolic forms at the bottom gradually evolve into distinct symbols with precise meanings at the top.

The focus of this section is not on how symbolic, rule-based processes emerge from sub-symbolic dynamics (this will be explored in the future part of this position paper). Instead, we examine how these two



Figure 2: *Liberation* by Escher.

seemingly distinct processes—symbolic reasoning and neural computation—can function together as an organic whole, a relationship we refer to as function alignment.

As pointed out in the paper *Thinking Like Transformers* (Weiss et al., 2021), rule-based programs can be "compiled" into a transformer. In other words, at the "hardware" level, the dynamics of some abstract representation still run on a neural system, while at the "software" level, the internal semantics can be a symbolic, rule-based program.

Now, consider function alignment among the following three hierarchical levels of representation:

1. The $x$-sequence, representing subsymbolic dynamics.

2. The $z$-sequence, representing a neural-symbolic system, such as a Transformer that processes embeddings of symbols.

3. The $z'$-sequence, another neural-symbolic system, which shares the same symbol vocabulary as $z$ but whose internal dynamics follow a compiled rule-based program.

This framework provides a natural explanation for the **unity of intuition and rationality**—how humans can both follow rules and break them while simultaneously having an intuitive grasp of the underlying rationales. This flexibility lies in the ability to utilize different inference pathways at various levels of representation.

## From Meaning to Explanation: The Bounds of Interpretation

The function alignment framework certainly provides more insights—we can now offer precise and rigorous definitions for some deeply abstract and meta-level concepts central to mind and intelligence, including *meaning*, *interpretability*, *analogy*, and *rationality.*

**Meaning**: What does a particular $y$-sequence (some physical events) mean to us? Its meaning is simply *the totality of hierarchical representations it triggers in the mind.* For instance, what might a short musical phrase mean to a trained musician? It could evoke subsymbolic-level emotional flow, symbolic-level representations such as melody, harmony, and form, and even highly abstract semantic layers, such as "a lullaby my mother used to sing when I was a child," which can be verbalized in natural language.

Depending on context, different levels of representation may carry different weights of significance. For example, in the handwritten message, "Top secret: the password to that old key safe is 9527," the core meaning clearly lies in the symbolic content—those four digits. But if this same sentence were found in a centuries-old letter, its value might shift toward its stylistic aspects, such as the aesthetic quality of its calligraphy—a subsymbolic representation.

Ideally, we would communicate by transferring the entirety of meaning—every level of representation that constitutes our internal feelings and thoughts—from one mind to another. But in practice, we cannot (at least not yet). Instead, we encode only selected layers into communicable media (such as text, speech, or music), and it falls to the receiver to *interpret* the intended meaning based on their own internal function alignments.

**Interpretability**, then, is the ability to express one representational layer (including $y$, the lowest physical layer) in terms of another—whether from subsymbolic to symbolic, symbolic to subsymbolic, or from one symbolic layer to a more abstract one. For example, a parent might interpret their baby's cry as "I need food" or "I am cold"; a singer might interpret a symbolic score into a rich subsymbolic vocal performance; and in Eastern cultures,

the phrase "today's moon is so beautiful" might be interpreted as the abstract symbolic expression "I love you." In the narrowest sense, interpretability is the ability to describe something using natural language—our shared symbolic representation layer $z$.

Function alignment suggests that *interpretability is intrinsically bounded.* Even when layers are aligned, they are not identical, and any interpretation inevitably sacrifices the unique dynamics of the original layer. For instance, interpreting a vocal performance through text will necessarily omit nuanced vocal expressions. Conversely, interpreting a written score through singing may miss explicit structural information such as harmonic progression.

Much has been debated (Fodor and Pylyshyn, 1988; Saba, 2022) about whether artificial neural networks or human minds are "interpretable." Within the function alignment framework, "bounded interpretability" clarifies that we can indeed interpret low-level neural outputs $x$ using high-level symbolic representations $z$, especially when agents share a common symbolic vocabulary. However, this power is limited: certain dynamics at the $x$-level cannot be expressible within $z$, regardless of alignment.

**Analogy-making**, from this perspective, is a form of indirect interpretation. Instead of interpreting a target representation at one layer directly using a source representation from another layer, an analogy uses an alternative target that shares internal representations with the original. For example:

- *"Life is an adventure"* draws structure from purposeful, unpredictable journeys;
- *"Argument is war"* relies on shared representations of attack, defense, and victory;
- *"Your love is like moonlight"* evokes both an emotional tone and an elegant feeling.

In other words, analogy-making is interpretation via *style transfer*: when two targets are juxtaposed, the human mind instinctively extracts their **shared source representation structure**. Sometimes, only the second (metaphorical) target appears, and the mind reconstructs the original target through context—such as in:

- *"He is the father of modern physics"*—where "founder" is omitted, yet readers infer not only his foundational role but also an emotional framing of responsibility, guidance, and care;

- *"I cannot swallow that idea"*—where "comprehend" is replaced by a physical metaphor that conveys not only cognitive resistance but also a visceral sense of stress or discomfort—linking symbolic misunderstanding to subsymbolic embodiment.

This "shared source" view of analogy aligns closely with the notion of "conceptual skeletons" proposed in *Gödel, Escher, Bach* (Hofstadter, 1999) and in *Surfaces and Essences* (Hofstadter and Sander, 2013). And the function alignment framework extends this view

further: the shared representation need not reside solely at an abstract symbolic level. It may span multiple representational layers—embodied, subsymbolic, symbolic—and the more layers involved in alignment, the greater the analogy's explanatory and expressive power.

Many Zen koans and classical poetry draw on natural phenomena not to illustrate abstract notions but to evoke *subsymbolic felt states* across layers. For example, *"the body is like the Bodhi tree; the mind is like a bright mirror"* encodes not just visual and symbolic meaning, but also an introspective experiential resonance. So, too do the aforementioned analogies of "moonlight" and "swallow." In these cases, analogies become a conduit for multi-layered profound experience.

Analogies are not only powerful but also ubiquitous and sometimes hard to be noticed. Consider the simple act of pointing to an object and saying its name—"this is a chair"—is actually a form of analogy. The visual form of the chair is $y$, the acoustic form of the chair is $y'$, and through function alignment, both $y$ and $y'$ will trigger the same symbolic-level representation $z =$ "chair". Despite its effectiveness, analogy is bounded in the same way as interpretability: we rely on partially aligned internal representations to make sense of each other's outputs. When alignment breaks down, interpretation collapses into nonsense.

**Explanation**: We are now ready to define explanation within the function alignment framework. To explain any target is to make sense of it by revealing the causal representations within the function alignment process, through interpretation and analogy. This definition aligns with the causal modeling tradition, where an explanation is the identification of causal structures—typically formalized as directed acyclic graphs (DAGs). Indeed, the function alignment framework itself can be viewed as a causal graph.

However, our view departs from traditional causal modeling in a subtle yet crucial way: we treat the causal graph not as an objective structure to be discovered, but as a high-level representation constructed by the mind. In explanation, what we can actually present or communicate are only the *representations*—the nodes of the graph. The edges—the "arrows" of causality—are not directly expressible. That is, while we can name two events or states, we cannot fully articulate *how* one causes the other, except by invoking abstract symbolic constructs such as "leads to," "influences," or "because." The causal relation itself remains implicit unless encoded in a symbolic system—such as logic, mathematics, or narrative—that renders these dependencies as structure.

In this sense, explanation is always a "symbolic shadow" of an underlying structure, never the structure itself. What remains hidden is the objective causality *as it is*. The explanatory power of any such translation is thus inherently constrained by function alignment and bounded interpretability. We do not explain the graph; we explain *through* it—an attempt to project layered internal representations onto the narrow bandwidth of symbolic expression.

**Rationality**, therefore, is the capacity to explain using logical or symbolic language. As we have argued, such interpretability is bounded—symbolic reasoning typically occurs at higher-level representational layers, while actual behavior and perception are rooted in lower layers. This view provides a structural and representational grounding for Herbert Simon's theory of *bounded rationality* (Simon, 1990). Simon proposed that humans are rational within limits—capable of reasoning, but constrained by cognitive resources. We now have a structural mechanism underlying this description: the limited cognitive capacity for rationality arises from the bounded interpretability inherent in function alignment. When an agent explains a satisficing action using a logic-based framework, it performs an interpretive operation that is, by definition, partial and bounded. People do, in fact, optimize across all levels of representation—including embodied feelings and subsymbolic dynamics—which, in Simon's terms, is to "satisfice." However, such optimization cannot be fully explained at the rational, symbolic level.

In contrast, *complete explanation* is possible only within a pure formal system. In such systems, all reasoning and representation occur within a single symbolic layer. There is no information loss about alignment, because no cross-level interpretation is required; explanation collapses into logical operations, and all causes and effects are expressed in the same formal vocabulary. Thus, the ideal of perfect rationality coincides with perfect alignment—achievable only in formal systems.

## Agent-Based Intelligence and Isomorphic Alignment

We now formalize function alignment within a mathematical framework. Each language model operating at a different representational layer—such as $f$ on $\boldsymbol{x}$ and $g$ on $\boldsymbol{z}$—can be considered an agent. When such agents are functionally aligned, they collectively behave as a single, unified system. We call this principle **isomorphic alignment**.

**Linear case**: Viewing Figure 1 as a first-order linear dynamical system, each arrow across two time steps (horizontal or diagonal) corresponds to a transition matrix, while vertical arrows between variables at the same time step reflect inter-variable coupling—mathematically captured by a covariance matrix.

Let us use degrees of freedom (DoF) to characterize the structure.

Suppose $\boldsymbol{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots\}, \boldsymbol{x}_i \in \mathbb{R}^n$ and $\boldsymbol{z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, \dots\}, \boldsymbol{z}_i \in \mathbb{R}^m$ are two independent linear dynamical processes (or "agents"), each governed by a first-order autoregressive rule. Their degrees of freedom are:

1. $\text{DOF}_{\boldsymbol{x}}$: $n^2$ for transition $+ \ n(n+1)/2$ for initial covariance

2. $\text{DOF}_{\boldsymbol{z}}$: $m^2$ for transition $+ \ m(m+1)/2$ for initial covariance

Now consider a combined process $\boldsymbol{w}_t = [\boldsymbol{x}_t, \boldsymbol{z}_t] \in \mathbb{R}^{n+m}$ modeled as a first-order linear dynamical system. The total degrees of freedom are:

$$\text{DOF}_{\boldsymbol{w}} = (n+m)^2 + \frac{(n+m)(n+m+1)}{2}$$

Next, we account for the structure imposed by function alignment, which includes:

- Horizontal transitions: $\boldsymbol{x}_t \to \boldsymbol{x}_{t+1}$ and $\boldsymbol{z}_t \to \boldsymbol{z}_{t+1}$, which in total contribute $\text{DOF}_{\boldsymbol{x}} + \text{DOF}_{\boldsymbol{z}}$

- Vertical coupling (within-time linear mapping between $x$ and $z$): $nm$ DoF

- Diagonal alignment (cross-time interaction between $x$ and $z$): additional $2nm$ DoF

Summing all contributions:

$$\texttt{Total function-aligned DoF} = \text{DOF}_{\boldsymbol{x}} + \text{DOF}_{\boldsymbol{z}} + 3nm$$

This matches exactly the DoF of the unified system $\boldsymbol{w}$, showing that **full function alignment renders two interacting representational processes structurally isomorphic to a single unified linear dynamical agent**. If any of the alignment arrows—horizontal, vertical, or diagonal—is missing, the system loses degrees of freedom and can no longer function as a fully integrated model. Q.E.D.

**Nonlinear and long-dependency generalization**: In nonlinear or higher-order cases, exact DoF calculations are more complex, but the same principle holds. When all pathways are preserved, the function alignment forms a closed dynamic system, where information and gradients flow bidirectionally across time and abstraction. Any break in the alignment creates information bottlenecks or reduces the capacity of such co-adaptation.

**Implication**: This structure demonstrates that functional alignment is not only a perceptual architecture but also a *mathematically complete condition for agent unification.* It provides a foundation for agent-based intelligence: different parts of the mind (or brain) may specialize in distinct representational dynamics, but through function alignment, they act as a unified agent.

# Beyond Modeling: Insights for Psychology, Philosophy, and Zen

Function alignment offers not only a computational theory of mind but also a lens through which we can deepen our understanding of ourselves. Across psychology, philosophy, and contemplative traditions like Zen, we encounter a core duality: the mind has two systems—one intuitive, one analytical. The challenge has always been not to choose one over the other, but to integrate them in a way that leads to wisdom and harmony.

In cognitive science, the duality is known as **System 1** and **System 2** (Kahneman, 2011); in AI, **Mode 1** and **Mode 2** (e.g., Hierarchical JEPA (LeCun, 2022)); in traditional philosophies, **Yin** (or feminine) and **Yang** (or masculine); in *Zen and the Art of Motorcycle Maintenance* (Pirsig, 1974), **romantic** and **classical** understanding. A harmonic integration of these two modes is not merely an intellectual task, but a living art and experience.

**Split brains and minds**: Psychological studies, especially on split-brain patients, offer striking evidence of function alignment and agent-based intelligence, even at the physical brain level. In one well-known setup (Gazzaniga, 2012): when light is shown to the left visual field, the right hemisphere perceives it, but the left hemisphere (responsible for language) cannot report it at the $z$ level. Still, the subject, at the behavior $x$-level can press the button to indicate "seeing the light." This suggests that each hemisphere can function as an agent, but only the left brain has access to symbolic expression. Without aligned input from perception, reasoning is blind.

**Koan of "mirror becomes the mask"**: If split-brain patients show what happens when symbolic processing is disconnected, our daily struggle is often the opposite: an *over-identification with the symbolic layer*, just as a performer who overemphasizes music theory may lose touch with the underlying flow of music itself. All minds are function-aligned but to varying degrees. A sharp rational mind is smart, but only a deeply aligned mind is wise. My favorite story about this second kind of misalignment comes from Sadhguru:

> *A man, having promised to quit drinking, again returned home late and drunk. On the way, he scratched his face on a branch. Not wanting his wife to know, he quietly applied bandages in the bathroom and sneaked into bed without a sound.*
>
> *The next morning, his wife slapped him: "You drank again!" He was shocked: "How did you know?" She pointed at the mirror. "The bandages were all over it!"*

This is indeed a great metaphor of misalignment: $z$ is the mirror of $x$, and when aligned, $z$ should reflect and serve $x$. But when we are not conscious enough, misalignment arises, and $z$ becomes ego and a distortion we confuse with truth. Despite smartness, we may act

for value detached from experience, argue logic without feeling, and become minds that speak without seeing.

**Towards integration**: The practice of Zen is, in essence, training in experiencing the truth and a deep function alignment. Enlightenment is not knowing more intellectually but **reconnecting to the raw flow of perception $x$** that is as close to $y$ as possible, followed by a non-egoic reintroduction of symbolic framing and deep function alignment. The Zen master's sudden shout, the nonsensical koan, and the silent meditation practice of "just sit" are designed not to teach knowledge *about* truth, but to interrupt overactive $z$-level symbolic thinking and restore access to the unfiltered experience.

As a well-known Zen analogy goes:

*First, mountains are mountains.*
*Then, mountains are not mountains.*
*Finally, mountains are once again mountains.*

These three stages beautifully trace the arc of function alignment:

- The first stage reflects a symbolic-dominant misalignment—where socially conditioned values obscure unfiltered experience.
- The second stage marks disorientation from symbols, allowing one to contact direct perceptual truth.
- The third stage represents deep re-integration, where symbols are rebuilt—fresh, aligned, and transparent.

In brief, this is not a rejection of rationality but a transcendental path where symbols are no longer a substitute for truth but its humble servant. It is the arc of function alignment in full: from confusion, to liberation, to a deep ease of being—a mind no longer at war with itself but aligned in harmony.

## Conclusion and Outlook

In conclusion, this paper proposes *function alignment* as a theory of mind that is not only intuitively compelling, but structurally grounded. Unlike many existing accounts of cognition that rely on pre-theoretical concepts or loosely specified metaphors, this framework makes explicit how meaning, interpretation, and analogy emerge from concrete relationships among representational layers. Each concept introduced, whether symbolic reasoning or feeling-level resonance, corresponds to a definable pattern of interaction within the alignment model. In

this sense, function alignment forms a coherent representational language, capable not only of modeling minds, but of serving as a blueprint for building them.

Moreover, this theory does something unique: *it explains explanation, and it gives meaning to the very concept of meaning.* It shows why interpretation is inherently bounded—meaning is layered, and it must be aligned to be understood. This perspective offers a unified theoretical grounding for many fragments of mind science, such as bounded rationality, symbol grounding, and analogy-making. Once treated as isolated phenomena and concepts, they now emerge as structural consequences of representational dynamics.

Furthermore, function alignment bridges domains too often kept apart. It is not built upon any philosophy or belief system. Rather, philosophies, psychologies, and even contemplative systems like Zen may find themselves reconstructible within it. If symbolic thought is to serve experience rather than obscure it, then we need not only more knowledge, but better alignment. Function alignment offers a shared foundation where logic and perception, explanation and intuition, can meet—not in conflict, but in coherence and harmony.

Finally, this first part has focused on the foundational aspects of the function alignment framework. It leaves several critical aspects for future development: the nature of action, interaction with environments and other agents, the emergence of symbolic language, and the question of how such alignment mechanisms might be realized via AI systems. As a foundational entry, this work sets the stage for these exciting developments to come in a larger program.

## Acknowledgments

# References

Jerry A Fodor and Zenon W Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

Michael Gazzaniga. *Who's in Charge?: Free Will and the Science of the Brain*. Hachette UK, 2012.

Douglas R Hofstadter. *Gödel, Escher, Bach: an eternal golden braid*. Basic books, 1999.

Douglas R Hofstadter and Emmanuel Sander. *Surfaces and essences: Analogy as the fuel and fire of thinking*. Basic books, 2013.

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Robert M. Pirsig. *Zen and the Art of Motorcycle Maintenance*. William Morrow & Company, New York, 1974.

Walid S. Saba. New research vindicates fodor and pylyshyn: No explainable ai without structured semantics, 2022. URL https://cacm.acm.org/blogcacm/new-research-vindicates-fodor-and-pylyshyn-no-explainable-ai-without-structured-semantics/.

Herbert A Simon. Bounded rationality. *Utility and probability*, pages 15–18, 1990.

Gail Weiss, Yoav Goldberg, and Eran Yahav. Thinking like transformers. In *International Conference on Machine Learning*, pages 11080–11090. PMLR, 2021.