System 1.5: **Designing Metacognition in AI**

Cognitive Science # Human and Artificial Expertise # System 1 # System 2

Authors: Nick Oh (socius), Fernand Gobet (CPNSS, LSE)

TL;DR

System 1.5 introduces a theoretical framework for artificial metacognitive regulation that mediates between System 1 (fast, intuitive processing) and System 2 (slow, analytical reasoning). It monitors problem familiarity and adjusts its processing strategy accordingly: For high-familiarity problems, it employs System 1, quickly matching patterns to solutions ("templates"); for moderate-familiarity cases, it adapts known templates and generates multiple solutions; and for low-familiarity scenarios, it engages System 2 processes for deliberative analysis. This approach transcends the traditional System 1/System 2 dichotomy, implementing a dynamic regulatory mechanism that adapts its cognitive strategy based on pattern recognition confidence. The framework provides crucial insights into how artificial systems can effectively coordinate different processing modes in expert decision-making tasks.

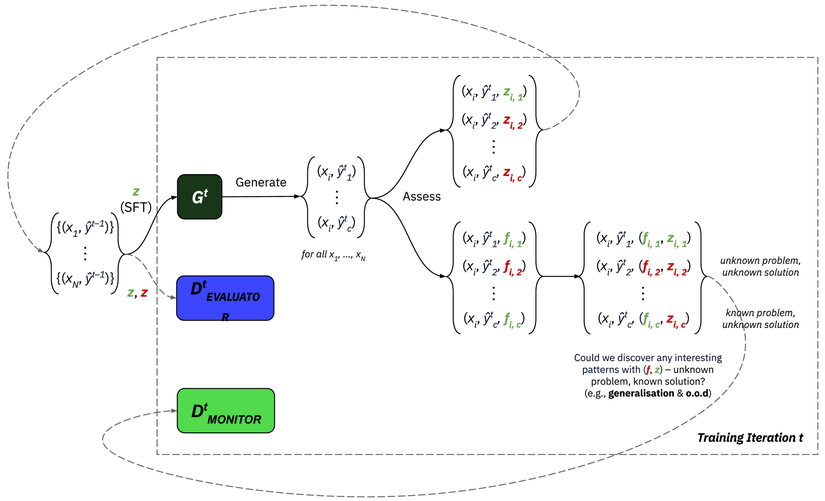
The Common Model of Intuition (System 1)

Synthesising four influential theories of expert intuition — **Hubert Dreyfus**'s *phenomenological approach*, **Herbert Simon**'s *chunking hypothesis*, **Daniel Kahneman**'s *dual-process model*, and **Fernand Gobet**'s *template theory* — we describe the five characteristics of System 1.

Human Intuition	Deep Neural Networks
Rapid Perception and Processing	Swift Processing of Complex Inputs
Opacity of Cognitive Processes	Opacity in Input-Output Processes
Holistic Understanding	Holistic Representationality
General Accuracy with Notable Exceptions	Generally Accurate Outputs with Occasional Exceptions
Intrinsic Role of Emotions and Past Experiences	Strong Influences from Training Data

Cognitive scientists observe System 1 processes emerging from associative learning mechanisms analogous to neural networks (McLeod et al., 1998; Evans, 2003). The Common Model of Intuition is an **attempt to formally describe how deep neural networks serve as a computational instantiation of human intuitive processes**.

System 1.5 (Training)



Initialization: Let $\mathcal{M}_{\text{BASE}}$ be a pretrained base model and $\mathcal{D}_{\text{SFT}} = \{(x_i, y_i)\}_{i=1}^{N}$ be an initial dataset where x_i represents problem description and y_i represents corresponding solution. We obtain:

$$\mathcal{M}_{\text{SFT}} = \text{finetune}(\mathcal{M}_{\text{BASE}}, \mathcal{D}_{\text{SFT}})$$

Iterative Process (for iterations $t = 1, \ldots, T$):

1. For each problem x_i, generate k candidate solutions (Cobbe et al., 2022):
$$\{\hat{y}_{ij} \sim \mathcal{M}_t(y|x_i)\}_{j=1}^{k}$$
2. Construct three datasets:
$$\mathcal{D}_{\text{GENERATE}_t} = \{(x_i, \hat{y}_i)|z_{ij} = \text{preferred}\}$$
where z_{ij} indicates preference
$$\mathcal{D}_{\text{EVALUATE}_t} = \{(x_i, \hat{y}_{ij}, z_{ij})\}$$
containing all solutions with preferences
$$\mathcal{D}_{\text{MONITOR}_t} = \{(x_i, \hat{y}_{ij}, f_{ij})\}$$
where f_{ij} captures "familiarity".
3. Update model:
$$\mathcal{M}_t = \text{finetune}(\mathcal{M}_{\text{BASE}}, \mathcal{D}_{\text{GENERATE}_{t-1}})$$

At final iteration T, we obtain three specialised functions:

$$\mathcal{G}_T = \text{train}(\mathcal{D}_{\text{GENERATE}_{T-1}}) \quad \text{(Generator)}$$
$$\mathcal{V}_T = \text{train}(\mathcal{D}_{\text{EVALUATE}_{T-1}}, \text{preference optimisation}) \quad \text{(Evaluator)}$$
$$\mathcal{M}_T = \text{train}(\mathcal{D}_{\text{MONITOR}_{T-1}}, \text{familiarity assessment}) \quad \text{(Monitor)}$$

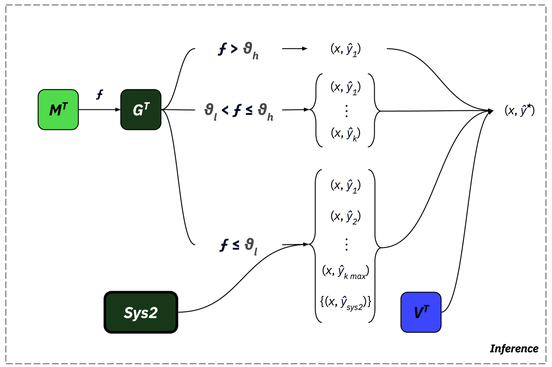
Inference (for input x): At inference time, we adapt our strategy based on how familiar the problem is:

1. Compute familiarity: $f = \mathcal{M}_T(x)$
2. Determine strategy based on familiarity:

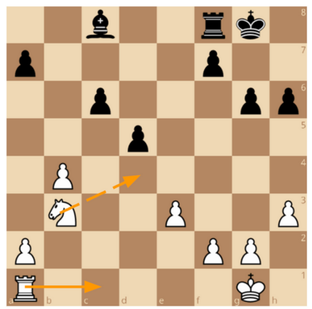
$$\text{out}(x) = \begin{cases} \mathcal{G}_T(x) & \text{if } f > \theta_h \\ \text{argmax}_{y \in \{\mathcal{G}_T(x)_j\}_{j=1}^{k}} \mathcal{V}_T(x, y) & \text{if } \theta_l < f \leq \theta_h \\ \text{argmax}_{y \in \mathcal{Y}} \mathcal{V}_T(x, y) & \text{if } f \leq \theta_l \end{cases}$$

where $k = \lceil C(1-f) \rceil$ and $\mathcal{Y} = \{\mathcal{G}_T(x)_j\}_{j=1}^{k_{\max}} \cup \{\mathcal{G}_T(x|h) : h \in \text{System-2}(x)\}$

System 1.5 (Inference)



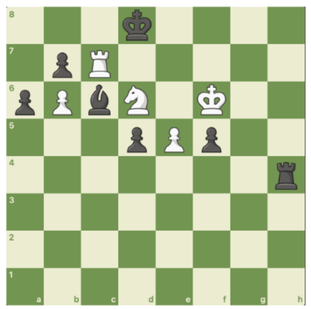
$f > \vartheta_h$



$\vartheta_l < f \leq \vartheta_h$

e.g., $k = C(1 - f)$, where $f = 0.8$ and $C = 10$

$\text{argmax}_{y \in \{\mathcal{G}_T(x)_j\}_{j=1}^{k}} \mathcal{V}_T(x, y)$



$f \leq \vartheta_l$

$\text{argmax}_{y \in \mathcal{Y}} \mathcal{V}_T(x, y)$

$\mathcal{Y} = \{\mathcal{G}_T(x)_j\}_{j=1}^{k_{\max}} \cup \{\mathcal{G}_T(x|h) : h \in \text{System-2}(x)\}$