System 1.5: Designing Metacognition in Artificial Intelligence

Nick Oh
socius
London, UK
nick.sh.oh@socius.org

Fernand Gobet
Centre for Philosophy of
Natural and Social Science (CPNSS)
London School of Economics
London, UK
f.gobet@lse.ac.uk

Abstract

Modern AI systems excel in both intuitive pattern recognition (System 1) and analytical reasoning (System 2), yet lack the sophisticated metacognitive mechanisms that humans employ to regulate these processes. We introduce System 1.5, a theoretical framework implementing metacognitive regulation in artificial systems that monitors cognitive processes, generates responses, and evaluates outcomes to determine when to rely on intuition versus analysis. Our key contributions include a theoretical foundation bridging cognitive science and AI, a formal framework for metacognitive regulation, and insights into coordinating different processing modes in expert decision-making.

1 Introduction

The divide between intuitive and analytical thinking has long challenged both cognitive science and artificial intelligence. Whilst modern AI architectures have made significant progress in both domains separately, they lack the metacognitive mechanisms crucial for expert decision-making – where success depends on dynamically shifting between these modes of thought. Unlike previous approaches treating System 1 and System 2 as discrete entities, we develop a framework that acknowledges their existence on a continuum, providing explicit mechanisms for coordination between them. Through this work, we aim to advance both our theoretical understanding of human metacognition and its practical implementation in artificial systems, whilst offering new perspectives on how different processing modes can be effectively combined in expert decision-making. We begin by examining influential theories of intuition and cognitive heuristics to establish the Common Model of Intuition, before introducing System 1.5 as a theoretical framework for implementing metacognitive regulation in artificial systems.

2 Cognitive Heuristics: The System 1 Framework

In this section, we first explore four influential theories of expert intuition: Hubert Dreyfus's phenomenological approach [Dreyfus, 1972, Dreyfus and Dreyfus, 1984, 1986, 1996, 2005], Herbert Simon's information processing theory [Newell and Simon, 1972, Simon and Chase, 1973], Daniel Kahneman's dual-process model [Kahneman, 2003, Kahneman and Klein, 2009], and Fernand Gobet's template theory [Gobet and Simon, 1996, Gobet and Chassy, 2009]. Each of these theories aims to explain the phenomenon of expert intuition, from Dreyfus's emphasis on embodied, non-representational knowledge to Simon's focus on chunking and pattern recognition. We then examine how these theories complement and contrast with each other, providing a comprehensive

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

framework for understanding expert intuition across various domains. This analysis will later serve as a foundation for developing the Common Model of Intuition.

2.1 Four Theories of Intuition

2.1.1 Hubert Dreyfus

Dreyfus's theory of intuition, grounded in phenomenology, presents a stark critique of symbolic approaches to intelligence [Dreyfus, 1972]. Central to his framework is the notion that human cognition is fundamentally embodied, situated, and experiential. Dreyfus argues that experts do not rely on symbolic representations but instead perceive their environment and make decisions through holistic processes [Dreyfus and Dreyfus, 1984, 1986, 1996, 2005]. His theory posits a five-stage model of expertise development: novice, advanced beginner, competent, proficient, and expert. At the expert stage, Dreyfus contends that understanding and decision-making become intuitive and fluid, transcending rule-based behaviour.

Despite its face validity, Dreyfus's theory has been challenged by empirical data on several fronts. Contrary to Dreyfus's predictions, expertise in many domains does not necessarily imply a decrease in abstract thought and an increase in concrete thought [Gobet and Chassy, 2009]. The existence of distinct stages in expertise development, as proposed by Dreyfus, lacks robust empirical support [Van der Maas and Molenaar, 1992]. Furthermore, the success of heuristic-search computer programs in achieving expert-level performance in complex games like chess contradicts Dreyfus's assertion that intuition is sufficient for expert performance. Critics argue that the theory underestimates the role of conscious problem-solving in expert performance and that neuroscientific evidence does not support the notion of holistic pattern recognition as envisioned by Dreyfus [Gobet and Chassy, 2009].

2.1.2 Herbert Simon

Simon's theory of intuition is firmly rooted in mechanistic explanations and experimental data. Simon posits that experts operate under the same cognitive limitations as novices, such as limited attention and short-term memory capacity [Simon and Chase, 1973]. The key difference, according to Simon, lies in experts' acquisition of a vast repertoire of perceptual patterns associated with possible actions, termed "productions" [Newell and Simon, 1972]. In this framework, intuition is explained by the firing of a production: recognising a familiar pattern triggers an automatic solution. Simon's theory finds support in experimental data from domains such as chess and physics, demonstrating experts' ability to perceive chunks of information and quickly recognise solutions to routine problems [Simon and Chase, 1973, Larkin et al., 1980]: That is, intuition to Simon is a cognitive mechanism of reducing the search space. Computer simulations have successfully modelled various aspects of this theory, including chunk acquisition and the transition from backward to forward search in problem-solving [Gobet and Simon, 2000].

However, Simon's theory has not escaped criticism. Detractors argue that it fails to explain how a single move is selected when multiple chunks are recognised, and that it does not fully account for the rapid, fluid nature of expert behaviour [Dreyfus and Dreyfus, 1986]. Some researchers contend that the theory does not capture the creative, constructive aspects of intuition [de Groot, 1986]. Technical objections include the suggestion that encoding into long-term memory may be faster than proposed by the chunking theory, and that the proposed chunk sizes may be too small to reflect conceptual knowledge [Holding, 1985, Gobet and Simon, 2000]. Critics also point out that Simon's computer models either failed to reach high levels of expertise or relied heavily on hand-coded knowledge [Gobet and Chassy, 2009].

2.1.3 Daniel Kahneman

Kahneman's dual-process theory of cognition offers a nuanced perspective on intuition, bridging cognitive psychology and behavioural economics. It distinguishes between fast, automatic, effortless, associative, and often emotionally charged "System 1" thinking, and slower, deliberative, and effortful "System 2" processes [Kahneman, 2003, 2011]. The difference in cognitive effort provides the most useful indication of whether a given mental process belongs to System 1 or 2. Kahneman introduces the concept of accessibility, which refers to the ease with which mental contents come to mind [Higgins, 1996]. While Kahneman acknowledges that expert intuition can be highly accurate in environments with sufficient regularities and where experts have had adequate opportunity to learn

these regularities [Kahneman and Klein, 2009], he also highlights potential pitfalls. Even in structured environments, attributes that are routinely and automatically produced without effort (termed "natural assessments") may lead to systematic biases. These biases often result from cognitive shortcuts or heuristics, such as the "prototype heuristic" and "affect heuristics" [Tversky and Kahneman, 1974, 1983, Kahneman, 2003].

While Kahneman's dual-process theory provides insights into both expert and non-expert decision-making, critics have highlighted several issues with Kahneman's dual-process theory. Critics have raised several issues with Kahneman's dual-process theory. The theory has been criticised for oversimplifying human cognition [Keren and Schul, 2009] and lacking strong neurobiological evidence. The distinction between System 1 and System 2 is often unclear [Glöckner and Witteman, 2010], with some arguing that intuitive and deliberative judgments may rely on similar principles [Kruglanski and Gigerenzer, 2018]. Critics argue that the theory overemphasises System 1's flaws while neglecting its adaptive value [Gigerenzer, 2008], and underestimates the role of expertise in improving intuitive decision-making [Klein et al., 1995, Campitelli and Gobet, 2010] (but see Kahneman and Klein [2009]). Additionally, the theory may not fully capture the integration of intuitive and analytical thinking in real-world decision-making [Reyna, 2004]. These critiques suggest that while Kahneman's theory provides valuable insights, it may not fully encompass the complexity of human cognition and decision-making processes.

2.1.4 Fernand Gobet

Gobet's template theory extends and refines Simon's approach, providing a more comprehensive account of expert intuition through complex knowledge structures and rapid pattern recognition mechanisms. The theory posits that experts develop chunks, which are perceptual patterns learned recursively, and which are not necessarily verbalisable and conceptual [Simon and Chase, 1973, Gobet and Clarkson, 2004, Gobet, 2012]. These chunks function as units of both perception and meaning, allowing experts to perceive groups rather than individual elements (e.g. chess players seeing groups of pieces rather than individual pieces). Patterns that recur frequently in the environment evolve into more complex data structures called templates [Gobet and Simon, 1996, 2000, Gobet and Chassy, 2009]. Templates, similar to schemas proposed by Bartlett [1995] and Minsky [1974], consist of a "core" encoding stable information and "slots" encoding variable information. According to the template theory, chunks and templates acquired through domain-specific experience guide attention and action. The theory proposes that expertise develops through acquiring a large number of chunks and templates linked to possible actions, enabling experts to quickly recognise patterns and make decisions in their domain of expertise. Recently, Ruoss et al. [2024] developed a search-free transformer model for chess that achieved Elo ratings of 2895 against humans in blitz games, providing an interesting parallel to the rapid pattern recognition mechanisms proposed in the template theory.

While Gobet's template theory addresses several limitations of Simon's chunking theory – by explaining how experts can rapidly adapt to new situations by updating variable information in templates, and by providing a mechanism for the integration of perceptual and conceptual knowledge – it still faces some challenges. The theory overemphasises long-term memory knowledge structures and provides limited insight into real-time processing, including creative or adaptive thinking [Linhares, 2005, De Groot, 1996]. Although it acknowledges emotions, it fails to offer a detailed account of the interplay between cognition and emotion. Critics argue that it underestimates the role of deliberate, conscious thought, particularly in complex situations [Montero and Evans, 2011]. Developed primarily using chess data, its applicability to less structured domains remains questionable. The theory may overemphasise expertise, potentially neglecting cases where novices demonstrate effective intuition. Lastly, there has been limited empirical research testing the theory's specific predictions on intuition, highlighting the need for further validation of its claims about intuitive expertise.

2.2 The Common Model of Intuition

The complexity of intuition and the ongoing challenge of comprehending human cognition have been extensively explored through various theoretical frameworks. While these approaches universally acknowledge the importance of perceptual learning, they diverge in their treatment of development (stepwise versus incremental) and the role of analytical processes – with Dreyfus uniquely emphasising situatedness and developmental stages. Despite these theoretical variations, scholars generally

agree that a comprehensive theory of human intuition must account for five essential features: rapid perception and processing, opacity of cognitive processes, holistic understanding, general accuracy with notable exceptions, and the intrinsic role of emotions [Gobet and Chassy, 2008].

The Common Model of Intuition focuses on characteristics rather than functions, primarily due to the opacity of the process. This fundamental opacity, which is widely recognised in the field particularly amongst dual-process theorists like Evans [2003], makes it challenging to delineate specific functions, leading researchers to describe intuition through its observable features. By emphasising characteristics, the model accommodates the elusive nature of intuitive processes while providing a framework for understanding and studying intuition.

While the precise mechanisms underlying intuitive processes remain largely opaque, the four theories of intuition discussed above converge on certain fundamental principles governing these phenomena. Intuition is more accurately described as a two-part process: the recognition of perceptual or conceptual patterns in working memory, followed by the rapid access and application of relevant information stored in long-term memory. This process can be modelled as an IF-THEN rule, where the IF component represents pattern recognition, and the THEN component represents the "switch" or activation of associated knowledge and actions. Importantly, *the patterns recognised in working memory encompass both perceptual and conceptual information*, broadening our understanding of intuition beyond mere perceptual recognition to include a wider range of cognitive processes.

These features draw a compelling parallel between human intuition and modern deep neural networks. Both demonstrate: (i) swift processing of complex inputs; (ii) opacity in input-output processes: (iii) holistic understanding of situations; (iv) generally accurate outputs with occasional exceptions; and (v) strong influences from past experiences or training data. Indeed, System 1 processes are most often described as those formed by associative learning processes characteristic of neural networks [McLeod et al., 1998, Evans, 2003], lending strong support to the view that *deep neural networks could be viewed as an artificial implementation of intuition*.

It might be important to end our discussion on System 1 by examining how Bayesian explanations of learning contrast with the approaches discussed above. On one hand, Simon's and Gobet's theories, while explicitly acknowledging human cognitive constraints (e.g., attention and short-term memory capacity), emphasise learning as a gradual accumulation of chunks and productions, with probabilistic information implicitly encoded through chunk connections and frequencies. On the other hand, Bayesian models (e.g., for an overview, see L Griffiths et al. [2008]) frame learning as a probabilistic updating process that integrates new information with existing knowledge (priors) to form updated beliefs (posteriors), emphasising uncertainty and experience in shaping decisions. While Bayesian models do not directly address memory limits, they implicitly account for this by focusing on the most informative evidence. This theoretical distinction raises an important question for future research: whether Bayesian reasoning is built-in through an innate frequency-processing module (System 1) or learned through explicit calculations (System 2) [Cosmides and Tooby, 1996, Gigerenzer, 2003, Girotto and Gonzalez, 2001, Sloman et al., 2003] (cited in Evans [2003]), or even whether System 1 is more naturally understood through a frequentist or Bayesian lens. Resolution of this question would enable subsequent mathematical formalisation of the cognitive processes involved.

3 System 1.5: Towards Artificial Metacognition

From a phenomenological perspective, there appears to be a metacognitive mechanism that modulates the engagement of System 1 and System 2 along an Intuitive-Analytical Continuum. This mechanism functions as a dynamic "switch", regulating the activation of our analytical processes as needed. This interplay between intuitive and analytical thinking forms the basis of what we term "System 1.5".

The nature and operation of this "switch" raise several questions: What cognitive or neural processes govern its activation and modulation? How can we design and implement an analogous system in artificial intelligence? Should it be an explicitly coded feature or an emergent property of the system?

This section addresses these questions by examining several key theoretical perspectives: first exploring the Intuitive-Analytical Continuum as a more nuanced alternative to the strict System 1/System 2 dichotomy, then reviewing theories of metacognition that explain how the mind regulates

its cognitive processes, ultimately leading to our proposed System 1.5 architecture that bridges intuitive and analytical processing in artificial agents.

3.1 The Intuitive-Analytical Continuum

The interplay between intuition and analysis is fundamental to human expertise. Rather than viewing System 1 and System 2 as dichotomous constructs, it is more persuasive to view them as existing on an Intuitive-Analytical Continuum. This concept of an Intuitive-analytical Continuum is supported by various research findings. Evans and Stanovich [2013] acknowledge that while they defend a dual-process model, they agree that the attributes often associated with Type 1 and Type 2 processing are continuous in nature. This supports the idea of a continuum rather than two discrete systems. In the context of expertise, Ericsson and Lehmann [1996] emphasised the importance of deliberate practice, which often involves analytical reflection, in developing expert performance. Hammond et al. [1987] introduced the cognitive continuum theory, suggesting that most decision-making tasks involve a quasirationality[1] between intuition and analysis. Collective research suggests that genuine expertise necessitates a dynamic integration of both intuitive and analytical skills across various domains. In the field of medical diagnosis, Norman et al. [2007] observed that experts employ both non-analytical and analytical reasoning strategies. Likewise, in the realm of business decision-making, Dane and Pratt [2007] advocated for the complementary roles of intuition and analysis.

Thus, the Intuitive-Analytical Continuum refers to the range of cognitive processes that blend intuitive (System 1) and analytical (System 2) thinking. This continuum represents the dynamic interplay between rapid, automatic processes and more deliberate, controlled cognition. It acknowledges that many cognitive tasks involve a mix of intuitive pattern recognition and analytical reasoning, with the balance shifting based on factors such as expertise, task complexity, and available cognitive resources.

This view of cognitive process aligns with several key points in cognitive science: (i) Expertise Development – as individuals develop expertise in a domain, processes that were once effortful and analytical can become more automatic and intuitive [Kahneman and Klein, 2009]. This suggests a shift along the continuum rather than a jump between discrete systems; (ii) Cognitive Load and Resource Allocation – the Cognitive Load theory [Sweller, 1994] suggests that as we become more familiar with a task, we can allocate cognitive resources more efficiently. This could be seen as movement along the Intuitive-Analytical Continuum; and (iii) Global Workspace theory – Baars [1997] suggests that the continuum operates as a workspace where attention regulates access to System 1 components, making them available to System 2 processes only after they receive attention.

The interaction between processing modes raises a fundamental question: what mechanisms control shifts along this continuum? As Kahneman notes, these systems "can be active concurrently" and "compete for the control of overt responses" Kahneman and Frederick [2002, pp. 51–52]. This observation points us toward metacognitive theories that explain how the mind regulates its own cognitive processes.

3.2 Theories of Metacognition

The concept of metacognition in cognitive science has been explored through various theoretical frameworks, each offering insights into how the mind monitors and regulates its own cognitive processes. Metacognitive Theory of Consciousness from Nelson (1990) posits that meta-level processes *monitor* and *control* object-level cognitive processes, demonstrating how consciousness divides into meta-level and object-level components to enable awareness and control over our mental processes.. This idea is complemented by Baars' (1997) Global Workspace Theory, which suggests that analytic reasoning (System 2) operates on information first processed by System 1, with attention serving as a gateway between unconscious pattern recognition and conscious deliberation – only after sensory inputs, working memories, and internal representations enter this attentional workspace can System 2 processes access and manipulate them. And recently, Clark's (2013) Predictive Processing Framework suggests that the brain constantly generates and updates predictions through a hierarchical system that combines knowledge-intensive strategies, integrating higher-level forecasts with incoming

[1]Hammond's concept of quasirationality is distinct from pure rationality. It encompasses various combinations of intuition and analysis, potentially leaning closer to the intuitive end of the cognitive continuum in some instances, and towards the analytic end in others (see Dhami and Mumpower [2018]).

sensory data in a dynamic cascade of cortical processing that guides our cognitive adaptation to changing tasks.

Each theory connotes for the necessity of **monitoring**, **generating**, and **evaluation**: Nelson's theory directly demonstrates how monitoring is fundamental to meta-level control through its empirical studies of metamemory, while Baars' theory shows how planning emerges from the need to selectively allocate limited conscious processing resources among competing processes, and Clark's framework reveals how evaluation is essential through its emphasis on prediction error minimisation. Together, these theories suggest three fundamental functions: monitoring that enables awareness of cognitive states, generating that facilitates attentional resource allocation, and evaluation that guides adaptive behaviour through error correction.

3.3 Designing Metacognition in AI

To define System 1.5, we first highlight that we are not talking about the general notion of metacognition per se. It is more of a metacognition of experts making judgments. The definitions of expertise and judgments are as follows:

- *Expertise*, whether in humans or machines, can be defined as the ability to perform intelligent tasks in a regular environment, obtaining vastly superior outcomes given cues, compared to those obtained by the majority of the population. This cognitively inspired definition, aligns with McCarthy's (2004) view of intelligence as computational goal achievement and Russell and Norvig's (2016) concept of intelligent agents maximising expected performance based on experience and knowledge.
- *Judgement* is informed assessments, conclusions, inferences, or predictions generated by either human or artificial expertise, based on specialised knowledge, training, or programming in a particular domain. Notably, this definition emphasises making judgments in a "regular" environment – one with reliable relationships between cues and outcomes, and continuous, accurate feedback [Kahneman and Klein, 2009, Waters and Gobet, 2024]. Regularity is a necessary condition for experts to identify and learn patterns over time, and perform in a reliable manner with notable exceptions.

These definitions precisely scope the role and requirements of System 1.5 in artificial systems. Since expertise requires superior performance in regular environments with reliable cue-outcome relationships, and judgement depends on accurate feedback for learning patterns over time, this frames System 1.5 not as a general metacognitive system, but as a specialised regulatory mechanism for environments with clear performance metrics and feedback loops. This aligns well with recent attempts in improving application in agentic, multi-step reasoning, such as `Agent Q` [Putta et al., 2024] designed to beat average human performance in the WebShop environment and `V-STaR` [Hosseini et al., 2024] designed to solve code generation and maths reasoning. These systems demonstrate expertise within specific environments with clear regularities and singular goals of maximising judgement utility.

Building on these foundational concepts, System 1.5 represents an architectural framework for metacognitive regulation, specifically focused on mediating between fast, automatic processes (System 1) and slow, deliberative processes (System 2). It is a framework that can interface with different AI components as System 1 and System 2, rather than being tied to specific implementations. Fast, intuitive processes (System 1) could be deep neural networks for visual recognition, language processing, or rapid evaluation; while slower, deliberative processes (System 2) could be logical reasoning modules, symbolic systems or algorithmic approximation of certain System 2 functions.

In this section, we propose System 1.5 as a framework with three critical regulatory functions: **Monitor**, **Generator** and **Evaluator**. Our formulation builds upon the work of Hosseini et al. [2024], which enables direct comparison with other agentic frameworks and helps clarify System 1.5's unique contributions. While `V-STaR` demonstrated the effectiveness of verification-based approaches, our work takes a step forward by providing a more cognitively plausible framework. Notably, we prioritise a theoretical description grounded in psychological and cognitive insights over mathematical formalisation (e.g., Direct Preference Optimisation [Rafailov et al., 2024]), as we believe establishing strong theoretical foundations must precede the development of precise computational implementations.

3.4 System 1.5

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:
$$\mathcal{M}_t = \text{finetune}(\mathcal{M}_{\text{BASE}}, \mathcal{D}_{\text{GENERATE}_{t-1}})$$

"Familiarity" mirrors the brain's prefrontal regulatory mechanisms [Shimamura, 2008], particularly in how it orchestrates between different processing modes based on task demands. Similar to how the prefrontal cortex synthesises diverse signals for metacognitive monitoring and control, we highlight two different types of familiarity: *problem-type recognition* and *solution-pattern familiarity*. The *problem-type recognition* component operates analogously to the brain's pattern recognition systems, quantifying encounter frequencies with similar problems, while *solution-pattern familiarity* mirrors what Miller and Cohen [2001] identified as the prefrontal cortex's role in synthesising both successful strategies and error patterns from past experiences. This approach is neurobiologically motivated by studies showing how the prefrontal cortex integrates both bottom-up perceptual matching (akin to our problem similarity detection) and top-down executive control (similar to our solution pattern analysis) [Evans, 1998, 1999], with differential weighting mechanisms that parallel the brain's belief-bias regulation demonstrated in empirical evidences in experimental psychology [Evans et al., 1983]. Just as neuroimaging studies reveal the prefrontal cortex's role in judgments of learning and feeling-of-knowing states [Schwartz and Bacon, 2008], our system accumulates these familiarity signals over time to guide adaptive processing strategies, effectively implementing a computational analog of what Houdé et al. [2000] demonstrated as the brain's transition from perceptual matching to executive control processes based on learned task contingencies and concepts.

At final iteration T, we obtain three specialised functions:

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

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

The three cases work as follows. When a new problem x is presented, System 1.5 framework first computes its familiarity score f using the monitor $\mathcal{M}_T$, which ranges from 0 to 1. This score, compared against two thresholds (θ_h and θ_l), determines how we approach the problem. For highly familiar problems ($f > \theta_h$), we trust our generator $\mathcal{G}_T$ to produce a single solution directly, similar to how human experts solve routine problems confidently.

For less familiar problems, our approach becomes more deliberate. With moderate familiarity ($\theta_l < f \leq \theta_h$), we generate multiple solutions, where the number $k = \lceil C(1 - f) \rceil$ adapts to our familiarity level (lower familiarity triggers more solutions). For instance, with $C = 10$ and $f = 0.8$, 2 solutions would be generated. When familiarity is low ($f \leq \theta_l$), we not only generate the maximum number of solutions k_{max} but also engage an external system with System 2-like capabilities for additional guidance. In both cases, our verifier $\mathcal{V}_T$ evaluates and ranks all candidate solutions, ensuring we select the most promising one.

This adaptive inference strategy mirrors human problem-solving: for familiar problems, we rely on direct solutions; for less familiar ones, we generate and evaluate multiple approaches; and for unfamiliar problems, we also leverage external knowledge or System 2-like approaches.

4 Discussion

This paper presents System 1.5, a theoretical framework for metacognitive regulation in artificial systems that bridges intuitive and analytical processing. Whilst our work establishes important theoretical foundations, several limitations and future directions warrant discussion.

First, significant aspects of the framework remain to be developed. We have not fully specified what constitutes System 2 processing in our architecture, nor have we formalised the precise mechanisms for familiarity assessment in the Monitor component. Our primary contribution lies in leveraging parallels between artificial and human intelligence to revisit cognitive science theories and derive architectural principles for AI systems. The necessary next step is to develop detailed technical specifications to enable meaningful implementation discussions with the technical community.

It is important to note that our approach does not advocate for neural realism. Instead, aligned with connectionist principles that intelligence can emerge from networks of simple computational units [Goodfellow et al., 2016], we focus on functional mechanisms rather than biological fidelity. Whilst humans consciously regulate their cognitive abilities, System 1.5 represents a computational implementation of these regulatory processes. This functional approach has historical precedent – for instance, Bayesian inference models have advanced our understanding of visual perception [Yuille and Kersten, 2006], whilst combining neurophysiology with feedforward neural networks has helped explain neural encoding in visual object recognition [Afraz et al., 2014]. Such bidirectional exchange between cognitive science and AI can create synergistic discoveries.

Our framework primarily addresses metacognitive regulation and conditional knowledge – the "knowing when and why" aspect of cognition. Whilst declarative knowledge ("knowing what") and procedural knowledge ("knowing how") are typically implicit in the AI models that System 1.5 regulates, our architecture explicitly designs the conditional knowledge that determines when and why to use different processing modes. This aligns with research showing that expert performance relies not just on superior knowledge organisation, but also on highly developed metacognitive skills that enable monitoring progress and adjusting strategies as needed [Berliner, 1994].

A key insight from our work is that search and pattern recognition are deeply interlinked in expert decision-making. Rather than being mutually exclusive, System 1 and 2 are often interleaved processes [Gobet and Simon, 1998]: during System 2's look-ahead search, System 1's pattern recognition suggests possible actions based on both current perception and anticipation. This suggests that future implementations of System 1.5 should consider how to effectively integrate both preferred and dispreferred solutions in familiarity assessment, paralleling how language models can learn from discrepancies between correct and incorrect solutions [Hosseini et al., 2024].

References

Arash Afraz, Daniel LK Yamins, and James J DiCarlo. Neural mechanisms underlying visual object recognition. In *Cold Spring Harbor symposia on quantitative biology*, volume 79, pages 99–107. Cold Spring Harbor Laboratory Press, 2014.

Bernard J Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, USA, 1997.

Frederic Charles Bartlett. *Remembering: A study in experimental and social psychology*. Cambridge university press, 1995.

David C Berliner. Expertise: The wonder of exemplary performances. *Creating powerful thinking in teachers and students*, pages 161–186, 1994.

Guillermo Campitelli and Fernand Gobet. Herbert simon's decision-making approach: Investigation of cognitive processes in experts. *Review of general psychology*, 14(4):354–364, 2010.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Leda Cosmides and John Tooby. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73, 1996.

Erik Dane and Michael G Pratt. Exploring intuition and its role in managerial decision making. *Academy of management review*, 32(1):33–54, 2007.

A.D. de Groot. Intuition in chess. *ICGA Journal*, 9(2):67–75, June 1986. doi: 10.3233/ICG-1986-9202.

Adriaan D De Groot. Perception and memory in chess. *Heuristics of the professional eye*, 1996.

Mandeep K Dhami and Jeryl L Mumpower. Kenneth r. hammond's contributions to the study of judgment and decision making. *Judgment and Decision Making*, 13(1):1–22, 2018.

H.L. Dreyfus. *What Computers Can't Do: A Critique of Artificial Reason*. Harper & Row, 1972. ISBN 9780060110826.

Hubert Dreyfus and Stuart E Dreyfus. *Mind over machine*. Simon and Schuster, 1986.

Hubert L Dreyfus and Stuart E Dreyfus. From socrates to expert systems: The limits of calculative rationality. *Technology in Society*, 6(3):217–233, 1984.

Hubert L Dreyfus and Stuart E Dreyfus. The relationship of theory and practice in the acquisition of skill. *Expertise in nursing practice: Caring, clinical judgment, and ethics*, pages 29–47, 1996.

Hubert L Dreyfus and Stuart E Dreyfus. Peripheral vision: Expertise in real world contexts. *Organization studies*, 26(5):779–792, 2005.

K Anders Ericsson and Andreas C Lehmann. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, 47(1):273–305, 1996.

J St BT Evans, Julie L Barston, and Paul Pollard. On the conflict between logic and belief in syllogistic reasoning. *Memory & cognition*, 11(3):295–306, 1983.

Jonathan St BT Evans. Matching bias in conditional reasoning: Do we understand it after 25 years? *Thinking & Reasoning*, 4(1):45–110, 1998.

Jonathan St BT Evans. The influence of linguistic form on reasoning: The case of matching bias. *The Quarterly Journal of Experimental Psychology: Section A*, 52(1):185–216, 1999.

Jonathan St BT Evans. In two minds: dual-process accounts of reasoning. *Trends in cognitive sciences*, 7(10): 454–459, 2003.

Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013.

Gerd Gigerenzer. *Reckoning with risk: learning to live with uncertainty*. Penguin UK, 2003.

Gerd Gigerenzer. Why heuristics work. *Perspectives on psychological science*, 3(1):20–29, 2008.

Vittorio Girotto and Michel Gonzalez. Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, 78(3):247–276, 2001.

Andreas Glöckner and Cilia Witteman. Foundations for tracing intuition. *Foundations for tracing intuition: Challenges and methods*, pages 1–23, 2010.

Fernand Gobet. Concepts without intuition lose the game: commentary on montero and evans (2011). *Phenomenology and the Cognitive Sciences*, 11:237–250, 2012.

Fernand Gobet and Philippe Chassy. Towards an alternative to benner's theory of expert intuition in nursing: a discussion paper. *International journal of nursing studies*, 45(1):129–139, 2008.

Fernand Gobet and Philippe Chassy. Expertise and intuition: A tale of three theories. *Minds and Machines*, 19: 151–180, 2009.

Fernand Gobet and Gary Clarkson. Chunks in expert memory: Evidence for the magical number four... or is it two? *Memory*, 12(6):732–747, 2004.

Fernand Gobet and Herbert A Simon. Templates in chess memory: A mechanism for recalling several boards. *Cognitive psychology*, 31(1):1–40, 1996.

Fernand Gobet and Herbert A Simon. Pattern recognition makes search possible: Comments on holding (1992). *Psychological Research*, 61:204–208, 1998.

Fernand Gobet and Herbert A Simon. Five seconds or sixty? presentation time in expert memory. *Cognitive science*, 24(4):651–682, 2000.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Kenneth R Hammond, Robert M Hamm, Janet Grassia, and Tamra Pearson. Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on systems, man, and cybernetics*, 17(5):753–770, 1987.

E Tory Higgins. Activation: Accessibility, and salience. *Social psychology: Handbook of basic principles*, pages 133–168, 1996.

Dennis H Holding. *The psychology of chess skill*. Routledge, 1985.

Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners. *arXiv preprint arXiv:2402.06457*, 2024.

Olivier Houdé, Laure Zago, Emmanuel Mellet, Sylvain Moutier, Arlette Pineau, Bernard Mazoyer, and Nathalie Tzourio-Mazoyer. Shifting from the perceptual brain to the logical brain: The neural impact of cognitive inhibition training. *Journal of cognitive neuroscience*, 12(5):721–728, 2000.

Daniel Kahneman. Maps of bounded rationality: Psychology for behavioral economics. *American economic review*, 93(5):1449–1475, 2003.

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

Daniel Kahneman and Shane Frederick. Representativeness revisited: Attribute substitution in intuitive judgment. *Heuristics and biases: The psychology of intuitive judgment*, 49(49-81):74, 2002.

Daniel Kahneman and Gary Klein. Conditions for intuitive expertise: a failure to disagree. *American psychologist*, 64(6):515, 2009.

Gideon Keren and Yaacov Schul. Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on psychological science*, 4(6):533–550, 2009.

Gary Klein, Steve Wolf, Laura Militello, and Caroline Zsambok. Characteristics of skilled option generation in chess. *Organizational behavior and human decision processes*, 62(1):63–69, 1995.

Arie W Kruglanski and Gerd Gigerenzer. Intuitive and deliberate judgments are based on common principles. In *The motivated mind*, pages 104–128. Routledge, 2018.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. Bayesian models of cognition. 2008.

Jill Larkin, John McDermott, Dorothea P Simon, and Herbert A Simon. Expert and novice performance in solving physics problems. *Science*, 208(4450):1335–1342, 1980.

Alexandre Linhares. An active symbols theory of chess intuition. *Minds and Machines*, 15(2):131–181, 2005.

John McCarthy. What is artificial intelligence? Online, 2004. URL `http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html`.

Peter McLeod, Kim Plunkett, and Edmund T Rolls. *Introduction to connectionist modelling of cognitive processes*. Oxford University Press, 1998.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Marvin Minsky. A framework for representing knowledge, 1974.

Barbara Montero and CDA Evans. Intuitions without concepts lose the game: mindedness in the art of chess. *Phenomenology and the Cognitive Sciences*, 10:175–194, 2011.

Thomas O Nelson. Metamemory: A theoretical framework and new findings. In *Psychology of learning and motivation*, volume 26, pages 125–173. Elsevier, 1990.

Allen Newell and Herbert Alexander Simon. *Human problem solving*, volume 104. Prentice-hall Englewood Cliffs, NJ, 1972.

Geoff Norman, Meredith Young, and Lee Brooks. Non-analytical models of clinical reasoning: the role of experience. *Medical education*, 41(12):1140–1145, 2007.

Pranav Putta, Edmund Mills, Naman Garg, Sumeet Motwani, Chelsea Finn, Divyansh Garg, and Rafael Rafailov. Agent q: Advanced reasoning and learning for autonomous ai agents. *arXiv preprint arXiv:2408.07199*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Valerie F Reyna. How people make decisions that involve risk: A dual-processes approach. *Current directions in psychological science*, 13(2):60–66, 2004.

Anian Ruoss, Grégoire Delétang, Sourabh Medapati, Jordi Grau-Moya, Li Kevin Wenliang, Elliot Catt, John Reid, and Tim Genewein. Grandmaster-level chess without search. *arXiv preprint arXiv:2402.04494*, 2024.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson, 2016.

Bennett Schwartz and Elisabeth Bacon. Metacognitive neuroscience. In John Dunlosky and Robert A. Bjork, editors, *Handbook of Metamemory and Memory*. Psychology Press, New York, 1st edition, 2008.

Arthur P Shimamura. A neurocognitive approach to metacognitive monitoring and control. In John Dunlosky and Robert A. Bjork, editors, *Handbook of Metamemory and Memory*. Psychology Press, New York, 1st edition, 2008.

Herbert A. Simon and William G. Chase. Skill in chess. *American Scientist*, 61(4):394–403, 1973.

Steven A Sloman, David Over, Lila Slovak, and Jeffrey M Stibel. Frequency illusions and other fallacies. *Organizational behavior and human decision processes*, 91(2):296–309, 2003.

John Sweller. Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4): 295–312, 1994.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131, 1974.

Amos Tversky and Daniel Kahneman. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, 90(4):293, 1983.

Han L Van der Maas and Peter C Molenaar. Stagewise cognitive development: an application of catastrophe theory. *Psychological review*, 99(3):395, 1992.

Andrew J Waters and Fernand Gobet. Trustworthy experts and untrustworthy experts: Insights from the cognitive psychology of expertise. In *Philosophy, Expertise, and the Myth of Neutrality*, pages 13–28. Routledge, 2024.

Alan Yuille and Daniel Kersten. Vision as bayesian inference: analysis by synthesis? *Trends in cognitive sciences*, 10(7):301–308, 2006.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract claims to introduce System 1.5 as a theoretical framework for metacognitive regulation in AI systems, which is thoroughly developed in Section 3 with specific architectural components and mechanisms. The introduction promises to bridge cognitive science and AI while offering new perspectives on processing modes in expert decision-making, which is delivered through detailed analysis of cognitive theories in Section 2 and the technical framework in Section 3.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper dedicates the first paragraph of the Discussion section to explicitly addressing several limitations, including the incomplete specification of System 2 processing and familiarity assessment mechanisms. Throughout the paper, it also critically examines limitations of existing theories when developing the Common Model of Intuition.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper is primarily conceptual and does not include formal theoretical results or proofs.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The paper does not include experimental results; it focuses on theoretical concepts and analysis of existing approaches.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The paper does not present new experimental results or algorithms that would require code or data.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The paper does not include experiments; it discusses theoretical concepts and existing approaches.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: The paper does not present experimental results that would require statistical analysis.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

Justification: The paper does not include experiments or computations that would require reporting of compute resources.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The theoretical nature of the work does not raise immediate concerns about safety, security, or discrimination. The paper does not involve the release of new models or datasets that would require specific safeguards or licenses.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: The paper is primarily conceptual.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not introduce new models or datasets that would require safeguards against misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites and credits original sources for the theories and concepts discussed.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce new datasets, code, or models that would require documentation.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve research with human subjects that would require IRB approval.