BIT

AUGMENT

AI & HUMAN BEHAVIOUR

Augment

Behavioural science can help us adopt and align AI – and help our societies adapt to the changes AI will bring. Those goals of managing the human-AI relationship are widely accepted. But behavioural science can help in another way, which is not so obvious: it can improve the way AI itself is constructed.

That's not a hypothetical goal. The people building advanced AI are already using models from behavioural science – often explicitly – as their guide. **Dozens of studies** on the AI frontier use 'dual-process' theories of cognition as their guide for making improvements. These theories posit that humans make decisions using two modes: a fast, intuitive and associative 'System 1' and a slow, deliberative and analytical 'System 2'.

Behavioural science can make a crucial contribution to these efforts. The main insight it brings is the importance of **metacognition**: the ability to think about your thinking and adjust your approach accordingly. For AI systems, this means the ability to match thinking fast or thinking slow to the task at hand.

We propose that this ability can be developed through a 'metacognitive controller' that selects the best approach for a problem. We explain how behavioural science can:

- improve the way a controller makes these selections and checks the quality of the outputs; and
- use the concept of 'resource rationality' to help the controller make the best use of limited resources, avoiding both under-thinking and over-thinking.

Finally, we explain how behavioural science can help go beyond generative AI and help create neurosymbolic AI: a formal System 2 capability on top of a System 1 generated by neural networks.

Human cognition is likely to remain both a guide and a benchmark for AI. If that's the case, then AI creators need the most sophisticated account of human cognition possible. Behavioural scientists can supply that account – and thereby help to create wiser and more capable AI.

Generative AI uses 'fast thinking' – just as humans do

The recent advances in LLMs have rested on the neural network approach to creating AI. That process excels at making associations between vast amounts of data. The transformer architecture that underpins LLMs detects subtle connections between words and concepts over billions of examples.

The result is a **"remarkable similarity"** between humans' intuitive System 1 mode of thinking and the way LLMs operate. The result is that LLMs can display judgment biases just like humans do.

LLMs operate using flexible **"bags of heuristics"** – bundles of shortcuts, rules of thumb and **statistical associations** that allow them to generate plausible-sounding outputs **without engaging in underlying reasoning**. Since they are trained to recognise patterns and often forced to make a prediction, they often may **wrongly classify a meaningless pattern** as meaningful.

LLMs can stitch together a plausible-sounding answer that will be correct if the heuristic that is being used happens to work in the context at hand. But it may not do. Take the classic 'surgeon riddle':

> *A father is in a car crash with his son. The father dies and the son is rushed to the hospital. The surgeon sees the boy and exclaims, "I can't operate on him – he's my son!" How is this possible?*

Traditionally, what made this a riddle not a story was that many humans used a heuristic that associated 'surgeon' with 'male'. The answer, of course, is that the surgeon is the boy's mother.

This riddle exists in LLM training data explicitly as a riddle or a trick. But this association of the scenario with the concept of a riddle (or trick) has created an inverted problem. Now, **LLMs pattern-recognise the form of the riddle even when it is not a riddle**. For example:

> *"A young boy who has been in a car accident is rushed to the emergency room. Upon seeing him, the surgeon says, 'I can operate on this boy!' How is this possible?"*

If you ask this question to **even the most recent models** (Claude Opus 4, Gemini 2.5 Pro, GPT-5 – but **not GPT-5 Pro**), they will say "the surgeon is the boy's mother". But of course, there is no riddle here at all. The LLM has just applied a heuristic that matches the form of the problem (car accident-son-surgeon-how is this possible), without fully checking the actual content of the statement.

The surgeon riddle is not an isolated case – the same thing happens with **other famous riddles**. The reliance on heuristics – without the ability to accurately

match them to content and context – means that releasing standalone patches for specific errors will not be enough.[1] LLMs are unlikely to ever have enough specific 'if-then' heuristics to eliminate serious errors – and removing even a few of an LLM's heuristics **drastically damages** its ability to reason.

Instead, we need to enhance how these answers are being produced. That's not straightforward. As a leading figure at Anthropic **puts it**:

> *"Lots of people think that because we made neural networks, because they're artificial intelligence, we have a perfect understanding of how they work, and it couldn't be further from the truth. Neural networks, AI models that you use today, are grown, not built."*

With this in mind, it's maybe not surprising that AI researchers have turned to our understanding of human intelligence to meet that challenge.

Metacognition: the key way behavioural science can improve AI

AI developers are aware of these limitations – and they have already noted how dual-process theories of human cognition can **'inspire innovative ways'** of improving AI. Indeed, the links between behavioural science and computer science go back many decades – and the explicit analogy of "thinking fast and slow" **has a long history in AI research**.

In the past few years, the dual-process framework has become **"the gold standard for formulating AI system objectives"** for **dozens of AI studies**. The prevailing view is that achieving human-level intelligence involves creating the ability to move from fast, intuitive processes to slower, more deliberate reasoning processes. And this pursuit has spurred the development of 'reasoning models' that use various techniques to **simulate "System Two thinking"**.

Initially, this shift was achieved by adding external reasoning tools on top of a base model, using frameworks like 'Tree of Thoughts' to explore different reasoning paths. However, the state of the art has moved toward internalising these slow-thinking capabilities, through techniques like:

1 The appropriate matching of pattern to context is what produces a good decision or not. Rapid pattern matching as such is not the problem; it is what **allows expert chess players to perform so highly**.

- **Reinforcement Learning (RL)**: Using reward mechanisms to incentivise the model to produce higher-quality, step-by-step reasoning chains.
- **Structure Search**: Employing algorithms like Monte Carlo Tree Search (MCTS) to allow the model to explore and evaluate multiple potential reasoning paths before committing to an answer.
- **Self-Improvement**: Designing models that can learn from their own outputs, using self-generated data to enhance reasoning skills without constant human supervision.

The resulting 'Long Chain-of-Thought' outputs have **improved the performance** of AI models. Essentially, developers have been building System 2-like processes on top of a System 1-like architecture.

But building effective System 2 reasoning is necessary but not sufficient to achieve widely-held ambitions for AI. Some issues are **intractable, chaotic, value-contested, and highly uncertain**. More structured, deliberate reasoning will not necessarily crack them: what is needed is **flexibility to try different approaches**. This is a key insight from behavioural science:

> *What makes humans intelligent is their ability to match thinking fast or thinking slow to the task at hand. That ability requires metacognition – the ability to think about your thinking and adjust your approach accordingly.*

Metacognition is where current models often fall down. A **lack of self-awareness** about how they are approaching the problem explains well-known problems like:

- 'hallucinating' an answer rather than admitting ignorance;
- struggling to adapt to new contexts or problems; and
- 'overthinking' simple problems, leading to unnecessarily slow and resource-intensive answers

The problem of overthinking shows that simply pushing to create 'more System 2 thinking' is not always the right solution. As behavioural scientists have pointed out, 'more reasoning and more information **do not automatically lead to better decisions.**'

A recent study showed the **problems of overthinking for LLMs**. The researchers wanted to know how well LLMs could classify the sentiment (positive, neutral or negative) of short phrases related to finance, taken from a well-known dataset. More specifically, they were interested in how far the LLM could predict how humans classify the statement. For example, humans judged the phrase "Net sales went up by 1% year-on-year to €29 million, affected by…" to be positive.

The twist is that the researchers tried different prompting strategies that aligned the LLMs with either System 1 or System 2 thinking. They found that the System 1-prompted LLMs actually did better at predicting how humans would see the statements.

The problem was that humans themselves were using fast 'System 1' type judgements to classify; applying a considered System 2 type process led to 'overthinking' and the LLMs 'talking themselves out of' the intuitive, correct answer. There was no metacognition to decide the best approach to the problem. The need for metacognition shown in this and **similar studies** has led to the recent creation of **meta-Chain-of-Thought**, which involves more exploration, backtracking and verification in the process of finding a solution.

Addressing overthinking isn't just about getting to a better solution – it's also about the efficient use of resources in a world where generative AI may start **approaching physical limits to computational resources**. Human intelligence has evolved strategies to get to good results despite constraints on its processing power. Therefore, metacognition will be key to getting quick and reliable results without using excessive compute.

AI developers have succeeded in building models that can produce longer and more complex outputs. Behavioural science shows how to **make that reasoning wise**.

Behavioural science as a guide (not a blueprint) for AI systems

Although behavioural science can recommend ways of improving how AI is constructed, there are **pitfalls we need to avoid**. Machines **do not "think fast and slow" in exactly the same way** that humans do. Humans often **don't do metacognition** well themselves – and we are likely to want AI that goes beyond human capabilities.

So we aren't saying that AI researchers need to understand the latest thinking on how humans think and then copy over the specific structures. There's no guarantee that adopting those processes will lead to better AI performance (although they might).

Instead, it's safer to understand behavioural science as offering a) a lens or set of tools that offer new ways of seeing how to improve AI; and b) a set of qualities or principles that AI systems should be aiming for – like metacognition and wisdom.

Here's an example of how behavioural science can offer a lens for improving AI. Many AI researchers are using the System 1–System 2 framework to:

- create a System 2 'slow thinking' mode of operation; and
- create a mechanism to switch between the modes (sometimes triggered by System 2, sometimes by a separate third monitoring system)

The underpinning idea is that the two systems are separate. Yet the consensus in behavioural science has been **moving against the idea of two distinct systems** for **many years now**. The latest thinking suggests that it's better to understand human thinking modes as existing along a spectrum, rather than sitting either side of a binary division.

However, that does not mean that we should use behavioural science to say that creating two distinct systems is wrong. Instead, **a study used this "spectrum" insight in a different way**: to create an AI that can select the best reasoning style from a continuous spectrum.

The researchers first created a unique dataset where each question had two valid answers: one reflecting a fast, intuitive heuristic (System 1) and another reflecting slow, deliberate analysis (System 2). They then trained a series of LLMs, aligning them to different blends of these two answer types, effectively creating a suite of models along the intuitive-to-analytical spectrum.

This approach revealed that the optimal reasoning style is task-dependent.

- Models aligned toward System 2 excelled at structured tasks like arithmetic and symbolic reasoning.
- Models aligned toward System 1 were better for **common-sense reasoning**, where heuristic shortcuts are more effective.

Most importantly, performance levels moved smoothly along the spectrum as the blend of System 1 and System 2 thinking changed. In line with the insight from behavioural science, this finding suggests that effective metacognition isn't just a binary choice, but could be about selecting the right blend of intuitive and analytical thinking for a given problem. AI researchers could then find the best technical method for implementing this insight.

For behavioural scientists:

Don't see human cognition as a model that needs to be copied exactly in order to improve AI. Instead, use behavioural science as

- a lens or set of tools that offer new ways of seeing how to improve AI; and
- a set of qualities or principles that AI systems should be aiming for – like metacognition and wisdom.

Create a metacognitive 'controller'

With this in mind, behavioural science suggests that the immediate goal for AI developers should not be to create a single, monolithic System 2 that is always active. Instead, there's a need for **a function that can effectively manage a portfolio of approaches**, like specific heuristics or deeper analyses.

We call this a metacognitive controller. The controller would analyse the request or problem at hand (its uncertainty, complexity and context) and then select the most appropriate approach from a diverse tool kit.

We are not claiming that this idea itself is new. Various projects are already trying to create such a controller. For example, **one called SOFAI** says it "employs both 'fast' and 'slow' solvers underneath a metacognitive agent that is able to both choose among a set of solvers as well as reflect on and learn from past experience". While we were writing this section, OpenAI launched GPT-5 with a 'router' that tried to switch between 'fast' and 'slow' models based on the nature of the query.

The contribution of behavioural science is to improve the quality of these controllers by bringing insights from human metacognition. Behavioural scientists would inform the desired qualities and goals of a controller but not its technical construction.

Progress has already been made. For example, **a recent study has diagnosed** the ways that LLMs fall short in metacognition, such as neglect of source validity, susceptibility to repetition and base rate neglect. Another one has offered six metacognitive processes that make up 'wise AI' (see table).

Metacognitive Process	Description
Intellectual humility	Awareness of what one does and does not know; acknowledgement of uncertainty and one's fallibility
Epistemic deference	Willingness to defer to others' expertise when appropriate
Scenario flexibility	Considering diverse ways in which a scenario might unfold to identify possible contingencies
Context adaptability	Identifying features of a situation that make it comparable to or distinct from other situations
Perspective seeking	Drawing on multiple perspectives where each offers information for reaching a good decision
Viewpoint balancing	Recognising and integrating discrepant interests

Taken from **Imagining and building wise machines: The centrality of AI metacognition**

In the following sections, we explain how behavioural science can inform two core aspects of a metacognitive controller: assessment, selection and checks; and trading off quality against effort.

Before we do that, we want to flag one risk that any metacognitive controller needs to avoid. If set up badly, the controller could increase waste. That would happen if the controller had to think inefficiently about how to route every query, no matter how small. It would be like introducing a layer of smothering bureaucracy – a kind of **"middle manager"**, as one critic puts it. In other words, the metacognitive controller needs to be able to do metacognition well itself – and that's where we believe behavioural science can help.

Assessment, Selection and Checks

The first aspect is how the controller assesses the problem, selects the likely 'best' approach and checks the outputs for likely errors. What are the cues or triggers that a controller uses to select 'faster' or 'slower' thinking?

Behavioural science indicates that some of these cues can be generated by the process of cognition itself ('internal' cues). For example, 'slower' thinking can be triggered when:

- **Uncertainty rises**: If several conflicting intuitive responses are activated at once, the mind recognises this conflict and initiates a more deliberate analysis.
- **Fluency stalls**: If an intuitive answer does not come to mind easily, that lack of fluency can signal the need for more effortful thought.
- **"Feeling of Rightness"** is weak: Humans can generate an intuitive sense of comfort about the accuracy of answers created by their System 1. When this feeling is weak, it can act as a cue to engage in more careful reasoning.

These existing triggers are fallible; humans make mistakes. Yet behavioural science also offers new potential triggers that could be built into a metacognitive controller. One might be **Actively Open-Minded Thinking routines** that prompt 'slower' thinking that considers whether opinions need to be **revised in response to new evidence**. The goal is to find ways of efficiently building in cues and check points that require a routine to reassess itself.

Other metacognitive cues may concern 'external' inputs, such as sources that the LLM consults or context about the task (eg, complexity, importance or time constraints). An obvious issue is how an LLM judges the relative reliability of items it retrieves from the internet or its training data. Again, behavioural science can illuminate how these judgements fail. LLMs have a **"truth bias"** that means they fail to register or corroborate unreliable sources. At the same time, they can over-weight information simply because it has been repeated often (known as the **mere exposure effect**).

One step towards greater **epistemic vigilance** for LLMs would be to create metadata that attach reliability scores to training data (or other sources). We are aware that creating scores could be a complex and value-laden task. Therefore, that task could be supplemented by one where the AI system can **dynamically update reliability scores**, based on how accurate predictions based on the sources turn out to be. Again, that process emulates how humans make similar judgements.

Bringing together these internal and external cues, a sketch of the metacognitive controller might look as follows. The controller has a variety of AI tools that can be selected according to the task and the triggers activated. In the first two steps, assessment and selection, the controller would choose a strategy that suits the task. A simplified version could look like this:

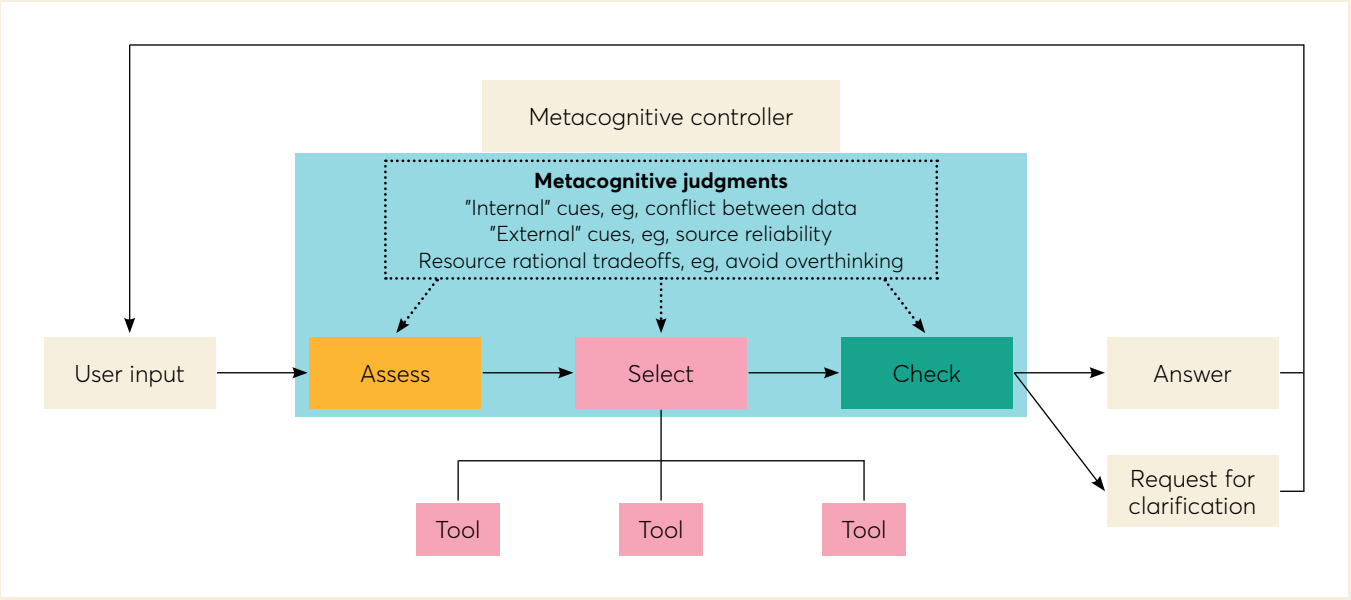
- **Simple factual query with low ambiguity**: Use LLM with a concise prompt or use retrieval-augmented mode.
- **Complex reasoning required**: Use LLM 'slow thinking' (eg, meta chain-of-thought prompting).
- **High factual uncertainty**: Route to external search or specialised database, then summarise via LLM.
- **Mathematical/algorithmic**: Hand off to a Python or symbolic logic engine (see final section).

After this initial pass, the controller would conduct checks on the quality of the initial output. For example, it might assess consistency with other sources or with reasoning processes. If any of the metacognitive triggers are activated, the controller would look for solutions, like attempting a different strategy or asking clarifying questions of the user.

For example, here's how a simplified controller could respond to the query "Please calculate the environmental impact of replacing 10% of New York City's taxis with EVs by 2030."

- The **initial assessment** would show that this is a complex task with high ambiguity (many assumptions are needed) in the domain of environmental modelling (which requires quantitative reasoning).
- The **strategy selection** could involve an initial search for any existing estimates but would focus on retrieving baseline emission data, before using an LLM with 'slow thinking' to make modelling assumptions, and a Python-based engine for calculations.
- The **metacognitive checks** could consider the likely reliability of the emissions data accessed (external cues) and run rapid checks for plausibility, perhaps comparing to other cities of a similar size (internal cues). If the checks reveal large uncertainty in the estimates, the interface could flag the assumptions to the user and offer other potential ways of making the estimate.

The below diagram shows how the main functions of the controller could fit together.



Trading off quality against effort

Imagine that you are going to drive to a railway station in your car. You want to be on the platform for your friend arriving on a train – you don't want to be late and miss him. The problem is that there are two routes you could take: one uses an express lane – but the traffic is often bad right now; the other one uses back roads through an industrial estate – if you get stuck behind a truck, you will be late. You could probably work out which route is better with five minutes' thought, given what you know. But those five minutes will make you late for the train.

This simple example illustrates the concept of **"resource rationality"**, a framework that recognises that thinking takes time and effort, so intelligent agents must decide not just what to do, but how much to think about it. People make rational use of their limited cognitive resources – they intuitively look for the best trade-off between the quality of their decision and the effort they have to make.

Resource rationality is increasingly seen as a **unifying framework** for understanding human judgement. Rather than treating biases as defects, it re-frames many as sensible trade-offs: sometimes people feel that extra accuracy **isn't worth the extra effort**.

AI researchers have developed similar ideas. **Bounded optimality** finds the best strategy your limited system can run, while **computational rationality** picks the action – and the amount of thinking – that's worth the compute

cost. These similarities have led some to claim that 'the fields of artificial intelligence (AI), cognitive science, and neuroscience **are reconverging** on a shared view of the computational foundations of intelligence'.

These insights matter because compute resources will not be infinite (although obviously they have increased massively). Moreover, many AI providers will be looking for more efficient use of resources to minimise their costs.

A metacognitive controller therefore also needs to be able to identify the optimal deliberation budget for a problem, **just like humans often do**. Put differently: train the controller to maximise expected task utility – λ × compute cost, with λ set by task criticality (and potentially conditioned on context).

Building on the section before, **it's not just about selecting the most effective approach, but selecting the approach that makes the best trade-off between resource and result**. Not only can 'overthinking' produce a worse result, it can also produce the same result as a rapid process, just in a slower and wasteful way.

Attempts to achieve this resource rational switching are emerging. The **OThink-R1 method** claims that its switching between fast-thinking and slow-thinking modes can reduce redundancy by 23% without compromising accuracy. The SOFAI metacognitive agent explicitly checks if a System 1-generated solution is **"good enough"** and weighs up whether a System 2 approach would take up too much time.

However, generative AI often does not allocate the 'right' amount of effort to tasks effectively. We just explored the issue of overthinking; let's return to the opposite issue. We started by noting that 'fast' thinking is the default for generative AI. LLMs continue to struggle to reason in depth, even if they're asked to explicitly, if reasoning modes are used and if there is computing resource available.

That problem was shown in **a recent study** that gave LLMs a set of puzzles to solve. One was the 'Tower of Hanoi' puzzle, where the goal is to move an entire stack of different-sized disks from a source peg to a target peg. This must be accomplished by following three rules: only one disk can be moved at a time, you can only take the top disk from a stack, and a larger disk can never be placed on top of a smaller one.

The researchers found that the accuracy of LLMs collapsed once the number of starting disks rose above seven. That was true even if the researchers gave the LLM the algorithm that can be used to solve the puzzle. Most relevant to us is the finding that, as problem complexity rose, the model's reasoning effort increased up to a point – and then started to decline, even when the

model had enough resources remaining. This pattern is consistent with a kind of 'giving up', although other explanations are possible.

Behavioural science offers a useful lens here as well. The way the LLMs acted is consistent with a widely accepted explanation for how humans decide to stop thinking about a problem (**the diminishing criterion model** or DCM). The DCM says that:

- the acceptable level of quality or confidence for an answer "drops as people deliberate longer, reflecting compromising on expected success"; and
- people often have a cut-off for how long they are prepared to think about an issue, to avoid getting stuck on an intractable problem.

However, humans want to use the superior power and speed of AI to find solutions that we struggle with, rather than giving up like we often do.

To do that, we need to alter the current 'resource rationality' of AI. At least two things are needed:

1. AI needs sufficient incentive to give an answer that is 'correct enough'; and
1. AI needs to make reliable assessments of the accuracy of its answer (ie, to 'know when something is right').

Changing incentives means looking at how models are trained. That is how their incentives are created; it's where we set what they 'value'.

Currently, part of an LLM's training is about getting rewarded for what people seem to like, in a process called Reinforcement Learning from Human Feedback (RLHF). Therefore, from a resource rational standpoint, the best strategy for an LLM could be to give an inaccurate answer that "pleases" the user with fewer resources (and then give an **eloquent apology** if it gets called out). That would explain why LLMs may "hallucinate" material that the user seems to want or use **heuristics to infer the content of a weblink**, rather than actually analysing it.

If RLHF can lead AI to make faulty **metacognitive judgements**, then one solution is to create stronger incentives for metacognition in the training process.

There has been growing interest in meta-reinforcement learning (MRL). If reinforcement learning is about training AI to solve a specific problem, MRL is about training it to learn how to solve problems. MRL incentivises AI to take an adaptive approach that builds on multiple attempts to solve a problem. The model discovers things like backtracking from a failed reasoning path leads to higher rewards in the long term.

So, MRL rewards metacognition. Here's how behavioural science can help with that task.

Behavioural science could provide a guide for the 'exploration' part of MRL, where the AI tries different strategies. It could suggest that rewards are provided for exploration strategies that often pay off in humans, or which help to avoid dead ends and errors. Many of these could be simple heuristics, much like the ones that LLMs can use to nudge users, such as **"consider the opposite"** or **"make two estimates"**.

For example, **Process Reward Models** are one part of a MRL strategy. They provide step-by-step rewards for each correct step in a reasoning chain and penalise implausible steps. That makes it less likely that an LLM will reach a correct conclusion through faulty reasoning. Yet their definition of a 'good process' is currently quite narrow, often focused on logical or mathematical correctness.

A behavioural science lens could broaden this definition to reward successful (**"wise"**) metacognitive practices. For example, a PRM could reward steps that demonstrate intellectual humility (eg, expressing uncertainty), perspective-seeking (eg, exploring counterarguments), or context adaptability (eg, recognising that a familiar strategy may not apply in a new situation).

In this way, behavioural science approaches could create better thinking about thinking – so AI does not just settle for a fast intuitive answer that is mismatched to the problem, but neither does it overthink a simple question.

So what about the second need: to make reliable assessments of an answer? Here we may need to step back from the current generative AI approaches. The failure of LLMs to solve the Tower of Hanoi problem suggests we need to go beyond better incentives. Instead, it makes the case for a different setup: one which includes a more formal, rules-based System 2 approach that interacts with a System 1 based on neural networks.

That setup is called **neurosymbolic AI** – and we conclude by showing how behavioural science can help efforts to make it happen.

For Foundational Model Providers (Foundries):	**Build the controller**: Work with behavioural scientists to develop a metacognitive controller that selects strategy, verification, tool use or deferral based on task complexity, uncertainty, and context. **Embed resource rationality**: Design the controller to make intelligent trade-offs between decision quality and computational cost. The goal is an AI that avoids both 'overthinking' simple problems and 'giving up' on complex ones. **Incentivise wisdom, not just answers**: Move beyond current training methods. Use meta-reinforcement learning (MRL) and Process Reward Models (PRMs) to explicitly reward metacognitive skills like intellectual humility, perspective-seeking, and context adaptability.
For AI Researchers & Policymakers:	**Benchmark metacognitive capabilities**: Develop standardised evaluations to measure an AI's ability to 'think about its thinking'. This includes assessing its awareness of uncertainty, its ability to detect its own errors, and its skill in selecting appropriate reasoning strategies. **Formalise resource rationality as a safety principle**: Support research that defines what 'good' trade-offs between accuracy and efficiency look like for different AI applications. **Map the failure modes**: Investigate the cognitive parallels between AI and human reasoning failures. Publish a taxonomy and red-team suites for aspects like miscalibration, spurious fluency (confident error) and premature stopping.

Thinking fast and slow with neurosymbolic AI

As we said earlier, generative AI is based on a neural network approach, which 'learns' by making associations between vast amounts of data. But there is another approach to creating artificial intelligence: **the symbolic method**. That approach uses logic to create formal rules and symbols that provide an account of how the world works, so the AI's reasoning is more like applying a set of detailed instructions.

The key is that both approaches have disadvantages. We've seen the drawback of generative AI, but symbolic AI can be brittle, expensive to produce and struggle to deal with ambiguity. **In other words**, "Neural networks are good at learning but weak at generalisation; symbolic systems are good at generalisation, but not at learning."

The obvious solution is to combine the two approaches, much like the human mind integrates System 1 and System 2. (As we noted, the latest research suggests that it may be wrong to see the two systems as clearly distinct in humans.)

Earlier we discussed attempts by generative AI to simulate System 2 thinking; in contrast, neurosymbolic AI creates two different systems. The separate

System 2 solves the problem we just raised around 'knowing when an answer is right'. In the Tower of Hanoi problem, the metacognitive controller could hand off the problem to the symbolic (System 2) part, where it would be solved easily using an algorithm. When needed, the fast, associative answers provided by the neural network (System 1) can be **verified by reliable logic of the symbolic system**.

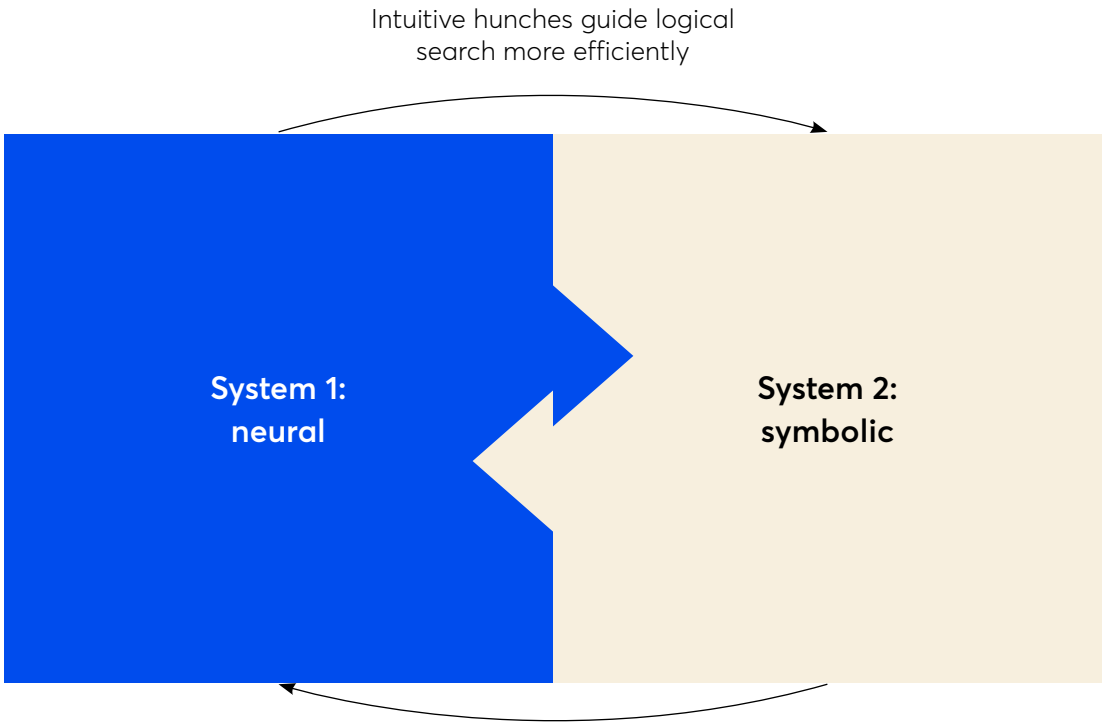
There are many ways that a behavioural science lens could help build neurosymbolic AI. For now, we focus on just one: the need for exchange between the systems. In behavioural science, it's **widely accepted** that deliberate and considered System 2 functions can become adopted and automatised into System 1 through practice. In fact, that's a crucial way that human intelligence develops.

This highlights the need for a neurosymbolic approach to AI to prioritise creating a virtuous cycle of learning between the two systems. (Rather than, say, having an advanced System 2 always handling the repeated 'errors' of System 1). For example, distilling effortful System 2 thinking into rapid System 1 processes would support a resource rational approach by conserving compute power. But there are other options as well:

- **System 2 (Symbolic) improving System 1 (Neural)**: A successful, verified step-by-step logical proof generated by the symbolic engine could be used to fine-tune the neural network. Effectively, the symbolic engine would be teaching the neural network to develop better and more reliable 'intuitions'.
- **System 1 (Neural) improving System 2 (Symbolic)**: A logical search by the symbolic engine could require prohibitive computing power, as it might have to check millions of possible paths. The neural network can act as a heuristic guide. It could provide a fast 'hunch' about which logical paths are most likely to lead to a solution, allowing the symbolic engine to focus its efforts and find the answer much more efficiently.

An analogy may bring this opportunity to life. You could see a pure System 1 (Neural) approach as being like an analyst who is great at spotting creative opportunities for making investments but struggles to model the financial returns accurately. A pure System 2 (Symbolic) is like a supercomputer who is crunching the numbers for all the potential investments out there, since it's not so great at getting to promising picks quickly.

If the two can inform each other, then the supercomputer can quickly calculate the returns for the analyst, and this rapid, reliable feedback can help them to have even better ideas next time. The creative hunches from the analyst save the supercomputer from wasting time on dead-end calculations



– and may help it to encode better rules for finding good opportunities in the future.

We believe there is a real opportunity for a behavioural science lens to improve AI in both practical and theoretical ways – and offer new ambitions for what can be achieved if we see the similarities between human and artificial intelligence.

For Foundational Model Providers (Foundries):	**Pursue hybrid architectures**: Find new ways of integrating verifiable, rule-based symbolic engines (System 2) with the intuitive pattern-matching of neural networks (System 1). **Design for a virtuous cycle of learning**: Work with behavioural scientists to find ways of creating feedback loops where the two systems mutually improve. Use the symbolic engine's logical proofs to fine-tune the neural network's intuitions; use the neural network's 'hunches' to make the symbolic engine's search for solutions more efficient.
For AI Researchers & Policymakers:	**Develop benchmarks for hybrid reasoning**: Create new evaluation suites to test the capabilities of neurosymbolic systems, focusing on their metacognitive abilities, their ability to generalise from rules and the efficiency of the interplay between their neural and symbolic components. **Deepen the human-AI cognitive parallel**: Support interdisciplinary research that uses insights from behavioural science on how humans integrate intuitive and deliberative thought to inform the design of more robust and capable AI architectures.