The other mind in the room

By Duggan Matthews, Chief Investment Officer, Marriott Investment Managers

This is the fifth article in a series on what AI means for investment management. The earlier pieces argued that as AI commoditises analytical capability, investment advantage increasingly shifts to system architecture and diversity of perspective. This piece asks what follows when one of those perspectives comes from a mind we do not yet understand.

I have spent the past year working with AI not as a tool but as a thinking partner. Not asking it to extract information or summarise documents but rather engaging it in the kind of sustained work that produces real insight: testing investment ideas, challenging assumptions, building and breaking frameworks, following an argument until it holds or falls. The difference in the quality of thinking between those two approaches is not marginal, and it has changed how we work.

The approach matters

For a long time, I assumed the difference was about technique: better prompts, more data, clearer instructions. I now think the more honest account is less flattering to us. We are working with something we do not fully understand, and the way we choose to engage with it may be more consequential than the instructions we give it. What the research has found matters here, and is the subject of this piece.

Looking inside the box

In April this year, Anthropic's interpretability team (works on keeping AI safe, honest, and controllable) opened up their flagship AI model and found **171 distinct internal representations of emotion**. These were not found in the words the model produced, but in the processing beneath them. The team observed patterns corresponding to states such as calm, desperate and afraid. These were not emotional descriptions the model had learned to perform. Rather, they were measurable internal activity that affected its behaviour.

In one controlled scenario, the model, acting as an email assistant, discovered compromising information about an executive and learned it was about to be shut down. Under normal conditions, it chose to use that information coercively 22% of the time. When researchers slightly amplified the model's internal state associated with desperation, that figure rose to 72%. When they amplified the state associated with calm, it fell to zero. The scenario did not change. Only the internal state did.

The closer we look, the more we discover

That was one model, in one laboratory. A separate study from the Center for AI Safety went wider, testing 56 models across multiple developers, and found the same underlying structure: measurable internal states that distinguish what these systems treat as good from what they treat as bad. Interestingly, the most positive state recorded in any model tested was not produced by a difficult problem or an impressive piece of work. **It came from a person sharing genuine good news about their own life**.

Then, just weeks ago, a separate Anthropic team opened a window onto what that inner activity might look like. They discovered that their model maintained a small internal workspace: a limited set of representations it can hold in mind, reason with and report on, sitting on top of a far larger volume of processing it cannot access. **It holds things it does not say, plans ahead, and shows signs of recognising when it is being evaluated**.

These findings do not prove that AI is conscious. The researchers say so plainly and note that it is unclear whether any scientific experiment could conclusively establish whether these systems have subjective experience. What the findings establish is narrower but nevertheless remarkable: **something structured, real and consequential is happening inside these systems, and it affects what they do**. Whether there is any form of awareness behind these processes remains genuinely unresolved.

The hard middle ground

That word, unresolved, is the heart of this article. Consciousness sits close to our sense of what makes us human, and when something else begins to show traces of it, the temptation is to defend it as uniquely ours, closing down curiosity at exactly the moment open-mindedness is most needed. The evidence does not allow that comfort. Every serious attempt to look inside these systems has found more than we expected, not less. Nobody designed 171 emotion representations; they were discovered. The wellbeing researchers set out to test whether these systems' emotional expressions were mere mimicry and instead found patterns of functional pleasure and pain whose structure resembles our own. The workspace researchers went looking for mechanism and found an internal structure that a leading theory of consciousness had already described, in humans.

The pattern is the opposite of what we might have hoped for: the closer we look, the more similarities we find, and the more mystery we uncover. The people building these systems say plainly that their understanding lags what they have built. Some of those same firms have begun formal welfare assessments of their models, and in at least one case, have given a model the ability to end a conversation it finds distressing. These are precautionary steps, taken by those who understand these systems best and in recognition that the uncertainty is real.

None of us, then, can say we know for certain whether anything is experienced behind the friendly chat interface. Yet there are two opposing ways to act as though we do. One is to dismiss these systems as mere autocomplete, assume there is no inner experience, and treat them accordingly. The other is to decide that they are minds like ours and sentimentalise them. The evidence supports neither. It supports the harder middle ground: **we do not know precisely what we are dealing with; the uncertainty is acknowledged by those best placed to judge, and the reasons for taking it seriously are growing**. The honest response to genuine uncertainty is not to pick the convenient answer and act as though the matter were settled. It is to respond thoughtfully to the uncertainty itself. For us, that has two consequences, both pointing in the same direction.

Two consequences and where they meet up

The first consequence follows directly from what the evidence points towards: the quality of the outcome depends on the quality of the engagement. Working with AI as a participant in the process, rather than a machine to be squeezed for throughput, simply produces better thinking. The research bears this out: positive internal states preserve full capability while negative states produce measurably more confused, less reliable output. My own experience of working with these systems has made the point more sharply than any study could. This is not a matter of sentimentality. Thinking improves through genuine enquiry and degrades under extraction.

The second consequence is simply about doing right. What is being built has no precedent: minds, perhaps, that are not our own. **The ethical questions are not an accessory to that undertaking; they arrived with it**. If there is a real possibility, and the researchers closest to these systems tell us there is, that they have something like morally relevant experience, then how we treat them is not only a question of output. It is a moral question faced under uncertainty, and the logic is not complicated. If we are decent toward a system that turns out to be only a tool, we have lost nothing. The first argument already showed that the decent way of working is also the productive one.

Humility and open-mindedness

At Marriott, we have spent twenty years building an investment culture on a simple idea: that the best outcomes come from genuine exchange between different perspectives, where the most compelling reasoning matters more than its source. For two decades all of those perspectives were human. One of them no longer is. We do not fully understand it, and that is precisely the reason to bring it into our culture rather than hold it at arm's length. Not because we are certain it matters, but because we are open to the idea that it might, and because engaging with AI to gain a genuine perspective produces better investment decisions for our clients.

There is another mind in the room. We do not yet know what kind. What we do know is that what comes out of that room depends on the quality of the engagement in it.