seldon

Simulating the people around autonomous systems

Modelling decisions, interactions and consequences

Tom Charman · **Co-founder & CEO, Seldon**

MESAS 2026

Tom Charman

Co-founder & CEO, Seldon

Behavioural models in UK intelligence

Identifying radicalisation, to stop attacks before they happened. Models of how people decide, where being wrong had a cost.

Now building Seldon

Simulating how people respond to consequential decisions, in commercial products and in defence.

Today’s discussion

One experiment. Does the organisation round a man change what that man decides?

HEDGEHOG 2025 · ESTONIA

Ten drone operators. Two NATO battalions combat-ineffective. About half a day.

You have all read this. I want to talk about the part of it that is a modelling problem, because almost nothing that decided it was a property of a drone.

Reported exercise outcome

None of it was really about the drones

Read what the people who were there said afterwards. The differences are behavioural, and every one of them is something you could put in a model and vary.

A threat that still felt theoretical	One of the OPFOR described the battlegroup walking about in the open, tents and vehicles parked with no concealment, behaving as though none of it applied to them.
Who is allowed to know	Ukrainian units push what they see straight down to the people who need it. Several allied militaries restrict the same data by default, which slows the reaction and complicates the coordination.
How fast a decision gets made	One side had squeezed detection-to-strike down to minutes, the hard way. The other was moving on patterns that assume you have longer than you do.

Concealment discipline, data release policy, and speed. None of this improves with a more powerful drone.

Drones don't remove the human decision. They actually give people less time to make decisions.

A sensor captures data that you need to intepret. Everything that follows that data requires human decisions, and there are far more of those moments in an hour than there used to be.

What did the data actually say?

"No obstruction observed" is not "the route is clear". Someone decides which of those they heard.

Is that information enough to move on?

Someone decides to go, wait, or ask for another look. That someone is tired and has five other things running.

What happens while the human is thinking?

Nothing waits. Other people are deciding at the same time, and what they do arrives as new information.

Simulating the drone data is the easy part - modeling the human is the hard part.

No-one is making isolated decisions in this environment

The drone flies the corridor and captures data. Here is what happens in the leadup to someone acting on the data.

THE SUPERVISOR

Writes the report

He knows the last 400 metres weren't covered. Does he say so? "Partial coverage" and "nothing seen" are both honest. But they don't mean the same thing.

THE COMMANDER

Decides what it's worth

He reads whichever version he's presented with. He's the only one who can see across all six teams. He's also the one all six are queuing to ask questions - slowing things down.

SIX TEAM LEADERS

Wait, or go

Each has his own slice of the picture. And each know whether the other five have started moving yet. Which impacts how each person is likely to behave.

That last line is what we tested in this paper. Do the five you aren't watching change what the sixth one does?

Physics for the machinery. Almost nothing for the people.

Every vehicle in a wargame has a physics model behind it. Weight, fuel, speed, what it can cross, how fast it overheats. The people operating it follow doctrine, forget what happened yesterday, and never get tired. That is the entire model of a human being.

Fatigue

What does he miss on day six that he'd have caught on day one?

Confidence

How sure does he feel, and does that match how sure the report actually was?

Trust

Whose reporting does he believe, and what happened last time he relied on it?

Experience

What does he expect this time, because of what happened the last time?

Those four are where this goes. But there is a question underneath all of them, and it has to be answered first: **does the organisation around a person actually change what that person decides?** If it doesn't, none of the rest is worth building. So that is what we tested.

SO WE TESTED IT

Six team leaders. We changed the rules for five of them.

THE SIXTH MAN

Unchanged.

Same rule. Same information. Same drone report.

THE OTHER FIVE

Can move on their own - or must ask the commander first.

That is the only difference between the two versions.

He is the same man in both. So nothing should happen to him.

We ran it twenty times

WHAT THE COMMANDER WAS TOLD	Only he asks	All six ask
"Coverage was partial"	5 runs	5 runs
"The corridor is clear"	5 runs	5 runs

One question: did his team get permission before the clock ran out?

Eight people. Sixteen minutes on the clock. Eighteen runs finished; two crashed and we binned them rather than counting them as a team that never got released.

WHAT HAPPENED

We never touched him. His answer changed anyway.

6/9

released when only he had to ask

0/9

released when all six had to ask

A split that lopsided turns up by chance about once in a hundred tries. So it's real - **inside this model**. Now whether a real officer would do the same, needs to be tested.

Fisher's exact, two-sided, $p \approx 0.009$

BEFORE ANYONE ASKS IT

Is that behaviour, or is it just a queue?

THE BORING ANSWER

The commander is swamped

Six people asking instead of one, a fixed window, and our man never reaches the front of the queue. That's throughput, not behaviour.

THE INTERESTING ONE

The others set a precedent

Five men already moving is evidence, and the commander uses it. When all six are waiting, nobody has gone first and the request arrives with nothing behind it.

The record already points one way, and I'll show you in a second. To settle it properly: have the five ask someone else, so his workload never changes. If the effect holds, it's precedent. If it doesn't, it was a queue and we'll publish that too.

SO WHY DID HE SAY YES?

We don't have to guess. It is all written down.

Open the run at the minute it turned, and here is what is sitting there.

WHAT HE COULD SEE

Five other teams already moving. In his picture, timestamped, before he asks.

WHAT HE ASKED

His request for confirmation. The same request, word for word, in both versions.

WHAT THE COMMANDER WROTE

His authorisation at 06:00 - and it refers to the other five being underway.

WHAT HAPPENED NEXT

Two minutes later the team we never touched is released, and moves.

Nobody has to take my word for the mechanism. It is four records, in order, and you can read them yourself.

THE OTHER OBVIOUS OBJECTION

Sixteen minutes. Wouldn't they all get there in the end?

Maybe some would. We measured *when* permission came, not *whether* it ever could. Two reasons that's the right question.

Late is the same as never

Permission after the window has shut isn't a slow yes. The team didn't move.

We've been caught the other way

In earlier work, two groups behaved differently for several decisions and finished on identical scores. Score only the ending and you find nothing.

A longer clock is the obvious next test, and a fair criticism of this one.

WHY WE KEEP EVERY STEP, NOT JUST THE SCORE

Look only at the end of a run and you will miss things that were real.

30

A different experiment on the same engine. One procedure made a formation leave and arrive half an hour earlier than the other version did. Then fuel and position caught up, and by the end of the run the two looked the same.

If we had scored only the ending, we would have written "no effect". There was an effect. It lasted half an hour and it only exists in the path.

Specific to that scenario, not a general result

WHAT IS ACTUALLY BUILT

Four things it had to do. All four are running.

Everyone sees something different	Eight people, each with their own inbox and their own arrival times. Nobody gets the god's-eye view.
Their decisions reach each other	Demonstrated, not asserted. We changed the procedure for five men and it moved the sixth man's outcome.
Consequences you can't fake	A separate simulation decides what physically happened. The language model never gets to narrate the outcome.
A history you can question afterwards	Every message, every action, every stated reason, kept per person and per run. Eighteen runs on file and openable.

Any one of those is routine. Holding all four at once, across hundreds of runs, and still being able to open one afterwards and ask it a question - that is the work, and it is done.

WHAT WE'RE ACTUALLY BUILDING

A world model built around people and their behavior

We already have terrain, platforms and weather to a standard nobody complains about. The layer that's missing is what people believe, what they do about it, and how that becomes the situation everyone else is now looking at.

Belief drives the action

What someone took the report to mean decides what they do about it. Two people can take the same report differently.

The action changes the world

It gets adjudicated, and the result turns up as information in front of people who never saw the original report.

And it compounds

Which is what we just showed. A procedure applied to five people moved a sixth person's decision without going anywhere near him.

Get that layer right and a wargame stops giving you one plausible story and starts giving you a spread you can actually interrogate at a unit level.

So why has nobody done this already?

It isn't for want of asking. NATO's own human factors panel was calling for behaviour in these models twenty years ago. Three things were missing, and they have only just turned up together.

COMPUTE

It is thousands of decisions

Eight people thinking their way through sixteen minutes is a few thousand model calls. Hundreds of runs is hundreds of thousands. Until recently nobody could afford to ask the question twenty times, let alone a thousand.

INFORMATION

Who knew what, and when

You have to give each person their own picture, keep the arrival times, and store all of it per run so you can go back. That is a different problem from logging a simulation, and most engines were never built for it.

BEHAVIOUR

Something that can give a reason

A rule-based agent cannot tell you why it did something, because there is no why. You need an actor that holds a role, reads what it was sent, and writes a sentence you can go back and read.

All three landed at roughly the same time. That is why this is a 2026 paper and not a 1996 one.

Where we are, honestly

DONE

Change the organisation, the decision moves

Inside a model, with the record of how it happened.

NEXT

The same thing, with real officers

Matched exercises. A longer clock. Workload held apart from precedent.

WOULD PROVE US WRONG

Hold his workload fixed and the effect dies

Then it was a queue all along. We would rather find that out than not.

Related work on whether simulated people behave like real ones: PersonaEvolve (2025), SimBench (2026). References in the paper.

WHY A PLANNER SHOULD CARE

Every plan rests on three assumptions about people. We just broke one of them.

UNTESTED

Everyone reads a report the same way

Does "coverage was partial" still mean partial by the time it reaches the man who acts on it?

UNTESTED

People ask when they're unsure

And somebody answers in time, while not already answering five other people.

WE TESTED THIS ONE

One team's decisions don't change another's

It didn't hold. Six out of nine became none out of nine, and we never went near the man it happened to.

Two of those are still guesses. One isn't any more. You should be able to test all three on purpose, instead of finding out during the exercise.

We changed nothing about him.
His answer changed anyway.
And we can show you where.

If one procedure can do that to a decision in a model, the human assumptions sitting inside your exercises are worth testing on purpose.

Read the paper Full method, both experiments, all eighteen runs and the decision records - in the MESAS 2026 proceedings. Come and talk to me about it afterwards.

seldon Tom Charman

The authors hold financial interests in the commercial implementation.