seldon

Modelling the Human Dimension

Behavioural simulation for multi-domain readiness

Tom Charman · Co-founder & CEO, Seldon

NATO CA2X2 Forum 2026

Tom Charman

Co-founder & CEO, Seldon

Behavioural modelling in UK intelligence

Using large social data to identify patterns in behavior - with the goal of finding those who were being radicalized and those who were victims. We used a combination of ML and CNNs to model behavior.

AI governance for the European Parliament

Advisory work on AI regulation and GDPR, on where the burden of proof sits when a model informs a decision that affects a human. Also dug in to who's responsible during a data breach.

Now based in San Francisco building Seldon.

Behavioral simulation for decisions that are expensive or dangerous to get wrong - particularly in regulated industries including banking, healthcare, critical infrastructure - and now increasingly the frontline.

seldon Model what happens next

We build models of how specific people decide.

Our models are trained on behavioral data rather than prompted with a personality description, and they are already deployed with government organizations. The underlying infrastructure powers what you'll see today.

1.2M
real sessions and field research behind the behavioural library

87%
behavioural fidelity against held-out human sessions, back and forward tested.

Compliant
SOC2, HIPAA, GDPR, and undergoing FEDRamp for government customers.

0
real end-user records used when running simulation - all data is anonymized.

WHERE THIS CAME FROM

We didn't start with defense. But the underlying engine supports work in high-risk environments.

COMMERCIAL

Will people actually do this?

Agents move through real products before a change ships. Regulated industries, production load, live today.

87% fidelity against held-out human sessions.

CYBERSECURITY

Who bypasses the control?

Security tooling is reactive. You learn that people route around a control after they have routed around it.

Deployed with government organisations.

DEFENCE

What matters most?

How does command behavior impact the unit level? How does behavior impact the overall run?

Today's subject.

Underneath all three is the same engine: we work out how somebody decides from the trail they leave behind. Clicks in one case, cables/briefings in another. Everything can be tested against historic events/reports.

HEDGEHOG 2025 · ESTONIA · 16,000 TROOPS FROM 12 NATIONS

Ten Ukrainian drone operators rendered two NATO battalions combat-ineffective in about half a day.

17
armoured vehicles simulated destroyed, with around thirty further strikes

10
operators, drawn from two Ukrainian unmanned systems brigades, acting as OPFOR

Again
at Aurora 2026 in Sweden, eighteen thousand troops from thirteen countries, same outcome

Everyone in this room has read that. I want to talk about the part of it that is a simulation problem.

seldon

None of it was really about the drones (technically)

Read the accounts from people who were there and the differences are behavioral and organizational. They are also, every one of them, things a simulation could represent and vary with ease.

Threat that felt hypothetical	An OPFOR participant described the NATO battlegroup walking around openly, tents and vehicles parked without concealment, behaving as though the drone threat did not apply.
Who is allowed to know	Ukrainian units push operational data down to subordinate units fast. Several allied militaries restrict access to the same data - slowing reaction time and complicating coordination.
Speed of a decision cycle	One side had compressed detection to strike into minutes through hard experience. The other was operating manoeuvre patterns that assume they have longer than they do.

Concealment discipline, information release policy, decision tempo. Those are human and organisational assumptions, and your exercises are the only place you can test them before somebody tests them for you.

And when the adversary you're up against is genuinely good, the exercise stops working

AT HEDGEHOG

The OPFOR inflicted so many simulated casualties that continuing the training serial stopped being useful. That is not a criticism of anybody. It is what happens when a realistic adversary meets an exercise designed around an approximate one.

AT JOINT VIKING

Reporting on the Norway exercise describes organisers asking the Finnish reservists playing the enemy to ease off. A completely sensible decision on the day, and a permanent ceiling on what the exercise could actually teach to those involved.

This is a supply problem, not a will problem

There are not enough people with real adversary experience to play a credible opponent in every exercise, at every echelon, as often as anybody needs. And the ones who do exist are needed elsewhere.

THIS IS NOT A NEW OBSERVATION

NATO's own science community has been saying this for 20+ years

RTO HFM-128 Human Behavior Rep.	Found that the absence of human factors in operational models was "seriously flawing the validity and predictive power", and proposed representing human states such as exhaustion.
STO reference model 2023	States plainly that there has to date been limited representation of the effects of events, kinetic and non-kinetic, on human behaviour in defence modelling and simulation.
MSG-207 cognitive warfare, 2022	Non-kinetic effects on human behaviour are often overlooked, and behaviour models tend to be single-use: built for one situation, then rebuilt for the next one.

The knowledge has been more advanced than the implementation for two decades. What's changed is that implementing it can now be fully traced with new technology.

What "behavior" actually means, at two levels at once

A unit is not a commander writ small. Cohesion and fatigue are properties of a formation that do not exist for any individual in it, and they interact with what that formation's commander believes.

AT THE COMMANDER

Trust in the reporting and in the chain above. **Tolerance for acting** on an incomplete picture. **Willingness to challenge** an order that no longer fits the ground. **Memory** of the last time a concern went unanswered.

AT THE UNIT

Fatigue after days of alert. **Cohesion** after taking losses. **Confidence** in the equipment and in the people alongside them. **Discipline** that erodes when nobody has slept and the threat has not materialised yet.

A tired formation whose commander has stopped expecting an answer behaves differently from a fresh one whose commander trusts the picture. Both can be issued the identical order.

WHY THE MIDDLE MATTERS

Most simulations keep the start and the end. The most important thing sits in between.

Take the commander whose vehicles were parked in the open. Here is the same decision recorded the way we record it.

WHAT HE WAS TOLD

Air threat assessed low. Move at best speed. Received at 04:30, and nothing since.

WHAT HE MADE OF IT

He believes the picture is current. Nobody has told him it is three hours old.

WHAT HE DID

Moved without dispersing. Inside his authority, and reasonable given what he had.

WHAT IT COST

Vehicles concentrated and visible. Worked out by the rules, not narrated by the model.

WHAT HE CARRIES

The next low-threat assessment lands differently, and that is in the record too.

Keep only the first and fourth and you know you lost vehicles. Keep all five and you know the assumption that cost you was "assessments are current" - which is a thing you can go and test and branch against.

HOW WE MODEL A PERSON

We read character by observing, rather than being told

Ask someone how they make decisions and you get a self-description. Watch what they actually did next and you get something you can model. It is the same method in both of our worlds.

COMMERCIAL

Clickstream

Millions of real journeys, and a weekly supply of ground truth to score ourselves against.

DEFENCE

Cable traffic

What was written, to whom, in response to what. The same inference on a different exhaust.

SMOKE-TESTED

50+ recorded exchanges

Conversations between world leaders, rebuilt from open sources. It held up across the evals we ran.

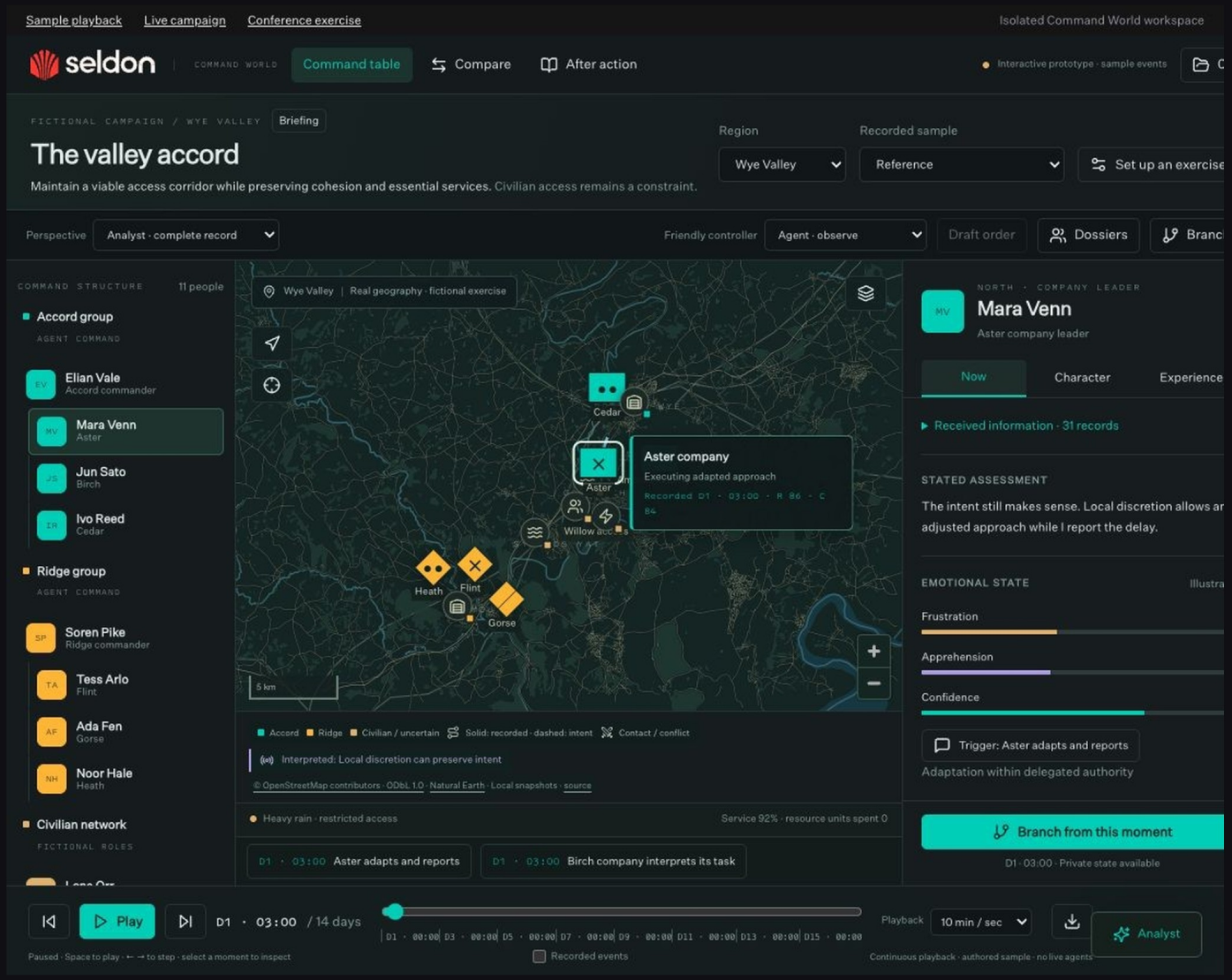
SO HERE IT IS, RUNNING

That record, fifteen actors at a time

Every role has its own inbox
and its own reading of it

Pause at any moment
inspect the actor, change one thing, branch

The map reads the record
it never invents an event of its own



Who knew what, and when they knew it

Start with the panel on the left. The analyst sees the whole world; no actor does. A message that has not arrived is simply absent from a decision, and one that has arrived can still be doubted.

COMMANDER

Holds a coordination report that is already an hour old

"Progress is uncertain."

SUBORDINATE

Has a local access restriction headquarters has not seen

"The route needs review."

AUTONOMY SUPERVISOR

Knows the sensing summary carries a caveat nobody upstream received

"Coverage is incomplete."

What makes two commanders in the same job behave differently

WHAT STAYS FIXED

Their job, what they are allowed to do, how they tend to operate, who they answer to. Your staff set all of this and sign it off before anything runs.

WHAT MOVES

What they currently believe, how confident they are, how much patience they have left, and what they remember happening to them so far.

Objective	Keep the access corridor open
Experience	Raised a constraint earlier. Nobody replied.
Tendency	Cautious with incomplete reporting
Relationship	Expectations of the commander, formed by that silence

One order, two commanders, and the day goes differently

BOTH RECEIVE "Maintain access. Report constraints."

THE ONE WHOSE LAST REPORT GOT ANSWERED

Pushes on, reports afterwards

The corridor stays open. Headquarters never finds out the route was marginal, and plans the next phase as though it is fine.

THE ONE WHO GOT NO REPLY LAST TIME

Asks first, and waits

The corridor is shut for ninety minutes. Headquarters does find out the route is marginal, and re-plans around it.

Neither of them disobeyed anything. An hour later headquarters is holding two different pictures of its own corridor, and that is the thing worth putting in front of people.

He only ever saw his own inbox. You get to see all of it.

WHAT THE COMMANDER HAD

Three reports, one of them ninety minutes old. That is the whole basis of his decision, and it is all he will ever be judged on having.

WHAT YOU HAVE

Those three reports, who else was sent them, what he said he made of them at the time, what the rules then did to the world - and the clarification he sent that was still sitting in a queue.

And you can go back and change it

Return to 04:30. Answer the message nobody answered last week. Run the same day forward again from the same state, and put the two versions side by side. Every intervention you make is recorded, so nobody later mistakes a run you steered for one you did not.

That is the difference between replaying an exercise and running an experiment on it.

AFTER THE EXERCISE

Interrogate the wargame afterwards, in plain language

Every decision, every piece of received information and every adjudicated effect is a record. An analyst assistant sits on top of those records, so the after-action review becomes a conversation with the exercise rather than an archaeology exercise.

"Why did the northern formation hold position for four hours?"

Returns the orders it had, what it was still waiting for, and what it said about the delay at the time.

"Who knew the corridor was closed, and when?"

Returns the delivery record actor by actor, including everybody it never reached.

"Show me every point where a commander acted against the reported picture."

Returns the moments worth arguing about in the review, rather than the whole timeline.

It answers from the record and it respects the same boundaries the exercise did. It is a search and explanation tool over what happened, not a second opinion on what should have happened.

BEFORE THE EXERCISE

Run the same day a thousand times and find where it breaks

A staffed exercise gives you one path through a scenario at considerable cost. Agents give you a distribution, which is a different kind of answer: not what will happen, but which assumptions the outcome was sitting on.

Which branches split

Most of the day is invariant. Find the few decision points where the outcome genuinely diverges and design the human serial around those.

Which assumptions are load-bearing

Vary one reviewed assumption at a time across hundreds of runs and see which ones move the outcome and which ones never do.

What the tail looks like

The rare path where three ordinary decisions compound into something nobody planned for is the one worth putting in front of people.

THE PART THAT IS NOT ANALYSIS

An adversary that is available every week, at every echelon

Ten people with real experience of a drone-saturated battlefield changed what two NATO battalions believed about their own survivability. They cannot be at every exercise. A model built on how that adversary actually behaves can be.

It does not get tired of being red

The same staff playing the opponent every rotation is how participants learn the staff instead of the problem. An agent has no house style to learn.

It can be built from their record

We infer how someone decides from what they actually did next. Given the material, the opponent in your exercise reasons the way that opponent reasons, not the way our doctrine expects them to.

And nobody has to go easy

Difficulty becomes something the exercise team sets deliberately, and records, rather than something negotiated on the day when the serial starts going badly.

This is a training instrument, not just an analytical one. Participants practise against something that responds to them.

THE FIGHT NATO IS ACTUALLY PREPARING FOR

Drones do not take the human out of the decision. They move it down, and they give you less time.

What is actually happening is not swarms flying themselves. It is small teams working with machines over links that keep dropping. And it is quietly rewriting who decides what, which is a command problem before it is a technology one.

Minutes, not hours

Detection to decision has collapsed. A staff process that takes an hour is no longer a slow process, it is an absent one.

Trust at the edge

A corporal with a controller reads an uncertain feed, and a brigade acts on her call. Either you believe her or the speed advantage is gone.

Authority moves down

The clearest picture is now at the bottom. Wargames are built around intent flowing down; increasingly the decisions are made where the feed is.

The unit itself changes

How you disperse, crew and structure a formation now follows from the human and machine pairing, rather than the pairing fitting the formation.

The human stays in control. Or steps out of it deliberately for analysis.

HUMAN COMMANDER

Participants practise

A real officer in the role, against subordinates, partners, civil authorities and an opponent who each respond from their own information. Their choices remain entirely their own.

AGENT COMMANDER

Analysts explore

Nobody in the chair, so you can run the same hour fifty times with one assumption changed and see which paths were fragile. Before the exercise, or between iterations.

An agent commander is an exercise configuration, never delegated authority over anything real.

THE PART PEOPLE UNDERESTIMATE

A good behavioural model is half of it. Keeping every path is the other half.

Everyone building in this space is working on the model. Far fewer are keeping what the model actually did, run by run, branch by branch - and that is where the answers live.

Same ending, different day

Two runs can finish on identical numbers having done completely different things on the way. You can't just focus on start and end.

Across a hundred branches

What you compare is not a hundred scores. It is where each branch split from the others, how long it stayed apart, and whether it came back.

That is the hard engineering

Every decision, every piece of information, every effect, kept per actor and per branch, and still queryable afterwards.

We learned this the expensive way. In an earlier study of ours, two arms behaved differently for several decisions and finished on identical final numbers - and the scoring rules meant they always would have.

Your context. Your assumptions. Your exercise control.

YOU BRING

Doctrine and authority. Organisations and people. Geography and capabilities. Reports and prior experience.

WE SEPARATE

What a document states. What a trait infers. What the scenario simply assumes. Three different kinds of claim, kept apart in the record.

YOU KEEP

Configure, assign roles, intervene, pause and review. Every intervention recorded, so an untouched run and a facilitated one are never confused.

And it runs where you need it to

Our models are small enough to run on your own hardware, inside your own network, with nothing leaving it. That's what makes work in classified environments possible at all. You decide what may enter, who may see it, which models are approved and how long records live.

It sits beside your simulation, not in place of it

Your environment

Holds exercise time, world state and the physical adjudication. It stays the authority on what is possible.

Behavioural roles

Take scoped observations, keep each actor's information and state, return structured responses with identifiers and receipt times.

Exercise records

Replay, comparison and human review across the boundary, with provenance preserved on both sides.

CANDIDATES HLA and NETN, C2SIM, or a versioned event API. Chosen against a specific host and tested in a sandbox.

WHERE THE EVIDENCE IS

This is running, it is measured, and there is one thing left to measure with you

BUILT AND RUNNING

The engine, information boundaries per role, persistent state, branching and full decision records. Fifteen actors and 171 recorded decisions in the campaign you saw. Small enough to run on your hardware, and deployed with governments today.

MEASURED, NOT ASSERTED

87% behavioural fidelity against held-out human sessions in the commercial product. Trait inference smoke-tested against fifty-plus recorded exchanges between world leaders, rebuilt from open sources, and it held up across the set.

WHAT WE MEASURE WITH YOU

Whether it holds in your domain, on your exercise, judged against how you staff that role today. We will not claim your ground until we have measured on it, and that is the first thing we would do together - smoke testing is key.

WHAT A PILOT ACTUALLY LOOKS LIKE

Six months, and a decision at the end that belongs to you

MONTH 1 · SCOPE

Agree the goal

We pick the role and the question together, and write down what a useful answer looks like.

MONTHS 2–3 · CONFIGURE

Add your context

Your doctrine, your chain of command, your past exercises. Your people check every assumption we make.

MONTH 4 · RUN

The study itself

We run the one thing you wanted tested, hold everything else still, and run it hundreds of times.

MONTHS 5–6 · JUDGE

Analyze results

Your people put it next to how they do that job today, against the bar you set in month one.

DISCUSSION

Which role in your next exercise is hardest to represent convincingly?

Command relationships

A subordinate who has stopped expecting an answer

Coalition coordination

A partner whose constraints are real and not yours

Interaction with autonomy

A supervisor deciding when to stop trusting the machine

Tom Charman · Co-founder & CEO, Seldon · seldon.com/defense

seldon

APPENDIX

Published NATO priorities, and what each one actually is

Source	Status	Relevant priority
ACT Audacious Training	Programme, 2026	Free play and credible opposing behaviour
DIANA Decision Superiority	Challenge, window closed	Interacting red, blue and neutral roles; repeated analysis
NATO Autonomy Plan	Policy, 2022	Autonomy levels and appropriate trust
NATO AI Strategy	Strategy, 2024	Traceability, governability, test and validation
ACT interoperability	Explanation, 2026	Interoperability includes people, procedures and trust

Different publications with different authority. None of them is an endorsement of this, and a closed challenge is not an open procurement route.

APPENDIX

Implementation and evidence status

Record	What it supports	Limit
Development campaign	15 actors, 171 recorded decisions, paused at 64.7 exercise hours	Partial run; software capability, not behavioural validation
Bounded episode	4 roles, one pair, 24 responses, one authored intervention	No causal trust effect; sampling was not deterministic
Separate prior study	20 continuations; a real difference concealed by final measures	Conditional on those action counts, not a universal property
Trait inference	50+ open-source leader exchanges as a smoke test	Define the measure before quoting a figure from the stage

Separate systems and separate evidence. Nothing here combines into a validated claim about real human behaviour.

APPENDIX

Integration, data handling and assurance

Interface candidates	HLA and NETN, C2SIM, or a versioned event API, chosen against the host
Exchange semantics	Identifiers, exercise time, delivered observations, permitted actions, adjudicated effects
Customer boundaries	Classification, national releasability, hosting, approved model access, audit retention
Acceptance work	Information isolation, restart recovery, facilitator intervention, latency at exercise tempo

Three problems exercise designers already know

Opposition becomes predictable	The same staff play red every time, and participants learn the staff rather than the problem.
Too many roles to staff	Partners, civil authorities and second-order actors get thinned out, and the coordination problem thins with them.
Nobody can explain the path	In the after-action review you know what happened. Reconstructing why a particular decision was taken is guesswork.

These are questions to confirm with a customer, not deficiencies we are asserting across NATO. Our contribution is aimed at the third one.

One event. Four readings. Four different next moves.

A communications outage during a coalition operation. Nobody is wrong. Everybody is working from what reached them.

ADVERSARY

Reads the silence as preparation. Moves first.

PARTNER

Reads it as a caveat problem. Waits for national clearance.

SUBORDINATE

Reads it as being ignored again. Acts within delegated authority.

CIVIL AUTHORITY

Reads it as a service failure. Reallocates people you were counting on.

Readiness is the ability to cope with all four at once. Most exercises assume one of them.