Mark Boyer

Northwind Managed Services: the operating model

Your diagnosis, your design and your first ninety days, assembled from the modules you have completed. Software defines the service, AI runs it, and human beings are why customers choose yours.

ASSEMBLED 2 AUGUST 2026 · 6 OF 6 MODULES · FAILURE DEMAND 41% · FLOW EFFICIENCY 1.1%

Complete. This is a model you can hand to a board.

WORKED EXAMPLE

A complete Operating Model for Northwind Managed Services, a fictional UK managed service provider, assembled from all six modules. Every number below is computed by the kit from the data in those modules, and every one of them reconciles, so check them. The assumptions they rest on: 21 working days a month, so the 3,666 contacts that arrive are 174.6 a day and 175 a day are closed, which is the stability Little's Law asks for; a ten-hour service window, so elapsed time is time on the SLA clock and lead times are in working days; a desk of eight people at 140 productive hours each, which is 67,200 minutes of capacity a month against the 58,255 minutes of handling in the demand table, and that ratio is the 86.7 per cent utilisation; and £16 fully loaded per contact, which is where the £288,000 comes from. Nothing here is saved to your machine.

·· HOW THIS WORKS

One argument, built in six moves

This document is the thing Shift Right makes. Not six reports in a folder: one operating model, written from your own numbers, that you can hand to a board or work from on Monday.

Each module answers one question and writes one part of it. They are grouped by the three movements of the argument: software defines the service, humans differentiate it, and shift right. You can run them in any order, and each is worth the hour on its own, but the sequence below is the one that compounds: you cannot sensibly decide where work should sit until you know what the work is and how it feels.

MOVE	MODULE	THE QUESTION IT ANSWERS	STATUS
i SOFTWARE DEFINES THE SERVICE	The Demand & Flow Reality	What is actually happening, and should this contact exist at all?	In the model
ii HUMANS DIFFERENTIATE IT	The Moments That Matter	What does it feel like, and who is holding the moments they remember?	In the model
iii SHIFT RIGHT	The Operating Model Canvas	Where should the work sit, frontstage and backstage?	In the model
iv SHIFT RIGHT	The Service Sensor Layer	What moves the service before a ticket exists?	In the model
v SOFTWARE DEFINES THE SERVICE	The AI Landing Zones Map	Where does AI land rather than stall?	In the model
vi HUMANS DIFFERENTIATE IT	The Language Architecture Audit	Does the catalogue sell confidence, or demand translation?	In the model

Where you are now. 12 of the 12 sections of the model are written. Every module is in. This is the complete argument, in your own numbers.

How to read what follows. Sections one and two are the argument you are joining. Three to eleven are your evidence, written as prose rather than dashboards. Twelve is what to do about it, sequenced. The board case at the end is one paragraph you could read aloud without apologising for it.

01 THE THESIS

Software defines the service. Humans differentiate it. Shift right.

For thirty years the industry ran on one idea: shift left. Push work towards cheaper resolution, automate more, deflect more, keep humans away from demand. It was a sensible optimisation that hardened into doctrine, and then AI arrived and did not beat it at its own game. It made the game irrelevant.

When software can sense, decide and resolve before a person ever sees the problem, there is no "left" worth shifting towards. Which leaves the question shift left was never designed to answer: where do human beings create value that nothing else can?

Shift right places human attention deliberately at the moments of highest stakes, highest feeling and highest judgement, and lets software absorb the rest. Four modes, not three tiers. Not human as fallback: human as differentiator.

By the time a ticket appears, the experience has already failed. We have spent twenty years optimising what happens after the ticket arrives. This model optimises to stop the ticket being needed at all.

Read as a provider: the routine work you win on today is the work that is about to be commoditised. What a customer will still pay a premium for is judgement at the moments that carry risk, and the confidence that somebody competent owns them.

02 THE PRINCIPLES

The ten this model is built on

1. **Software defines the service.** If a machine can sense it, decide it or resolve it, it should, invisibly, within policy, before anyone needs to ask.

2. **Humans differentiate it.** Deploy people where they change how the moment feels, not where the process happened to run out.

3. **Shift right, deliberately.** Human attention belongs at the moments of dread, judgement and consequence, not the password resets.

4. **Feelings are the product.** Until services are designed for how people actually think, feel and behave, they will keep working on paper and failing in practice.

5. **Sell confidence, not time.** Nobody buys hours. They buy the feeling that the outcome is safe in your hands.

6. **Design the uncertainty out.** It is rarely the waiting; it is the not knowing. Clarity and expectation design are operational disciplines.

7. **Language is architecture.** If customers need a glossary to understand your catalogue, the problem is not their vocabulary.

8. **The floor is not where you compete.** AI raises everyone's floor at the same speed. Strategy built on efficiency alone is a plan to become average, faster.

9. **Steal from outside.** The next great idea for services will come from behavioural science, marketing, design and hospitality.

10. **Rigour on the question, not the answer.** The leaders who matter next are uncertainty experts, not certainty performers.

03 THE DEMAND

What actually arrives, and whether

it should

3,666
contacts a month, scored

41%
failure demand: work created by earlier failure

73%
of volume scores for self-heal

75%
sits in the wrong mode today

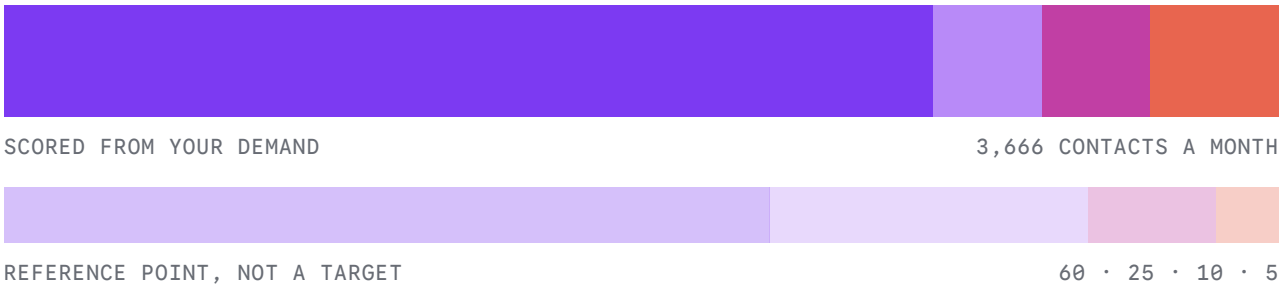
£288k
a year, the cost of that failure demand

1,500 contacts a month exist because something already went wrong. Not work: the cost of earlier failure, arriving dressed as work, and almost certainly paid for per ticket. Handling it more efficiently only makes the underlying defect more affordable to keep.

- Chase an update on something already logged · 640/mo
- Report the same fault again · 310/mo
- Ask for a report we should have sent · 180/mo
- Something you changed has broken something else · 145/mo
- Escalate because nobody came back to me · 130/mo

Where your demand scores

What your own scoring routes to each mode, against a well-shifted reference of 60 / 25 / 10 / 5



Self-heal 73% Guided 9% Assisted 8% White-glove 10%

Where your demand scores, mode by mode. This is what the work is, not what you have chosen to design: section 08 is the design, and the gap between the two is the decision you are actually making.

MODE	YOUR VOLUME	SCORED SHARE	REFERENCE
Self-heal	2,670	73%	60%
Guided	315	9%	25%
Assisted	310	8%	10%
White-glove	371	10%	5%

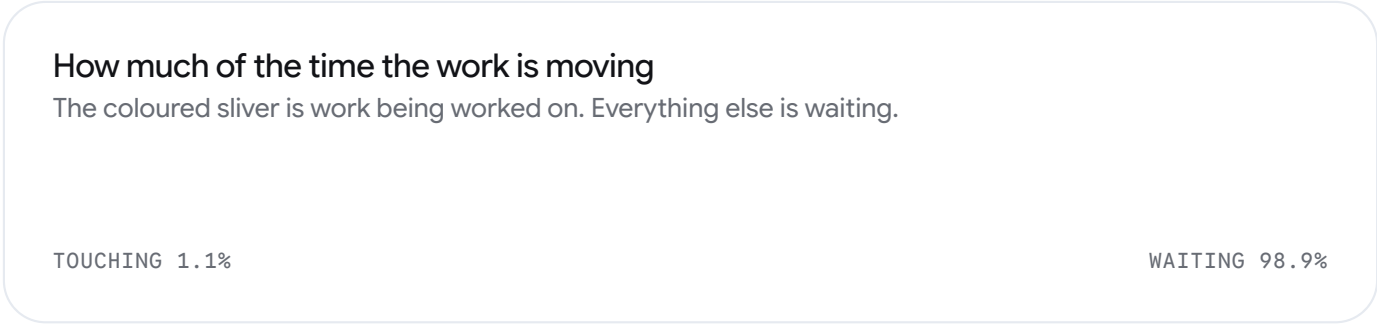
The reference is what a well-shifted operation trends toward. It is a reference point, not a target: your mix should follow your demand.

04 THE FLOW

How much of the time the work is actually moving

1.1%
flow efficiency: all touch time ÷ all elapsed time

98.9%
of elapsed time the work is simply waiting



Your work is being worked on for 1.1 per cent of the time you hold it. The rest it waits: in a queue, for an approval, for someone back from leave, for an answer you could have asked for at the start. That figure is total touch time divided by total elapsed time

across every line you timed, not the average of the per-line percentages, which a few fast high-volume lines would flatter.

This is the number that reframes the conversation, because every other measure in a service report, tickets per head, cost per ticket, utilisation, first-line resolution, describes how busy the resource is. None of them can see how long the work waits. Service and knowledge work typically runs between one and ten per cent, so this is probably normal. Normal is the problem.

05 THE ARITHMETIC

What the queue makes inevitable

2.4	350	×6.5
days average lead time, by Little's Law	the WIP cap that would deliver your promise	queue multiplier at 86.7% utilisation, against ×4.0 at 80%

You carry 421 items in progress and close 175 a day. Little's Law puts average lead time at 2.4 days. It holds regardless of how work arrives or in what order it is done, and it asks one thing of you: that the system is stable, with about as much work leaving as arriving. If the backlog is growing, this figure is the optimistic end of the truth. The promise is 2 days. That gap is arithmetic, not effort, and it will not close until work in progress falls to about 350 or throughput rises.

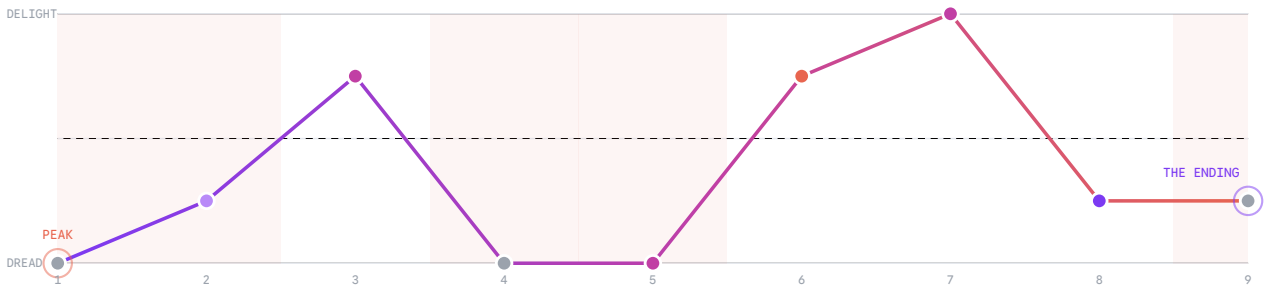
> Utilisation is set at 86.7 per cent. The queueing term is hyperbolic: ×4.0 at eighty per cent, ×6.5 here. A target above eighty is a decision to have slow response times, and a contract that prices for this utilisation while promising tight responses is internally contradictory. The queue will decide which commitment wins.

06 THE MOMENTS

The two moments they will remember

How the journey feels, and who is holding it

Each point is one step, coloured by who handles it. Shaded where the customer does not know what happens next.



Self-heal Guided Assisted White-glove Nobody

Retrospective memory is dominated by the most intense moment and the final one; duration contributes surprisingly little. Of every step in Our finance system is down, two days before month end, two are doing most of the work on what your customer says about you afterwards.

MOMENT	STEP	FELT	HELD BY
The peak	It stops working and we cannot close the ledger	1/5	Nobody
The end	Root cause report, six working days later	2/5	Nobody

Your ending is unattended. The last thing that happens is nobody, and the ending carries disproportionate weight in memory.

> This is what shift left actually did, stated plainly. It pushed people away from demand, and the two places it pushed hardest were the moment of dread and the moment of closure. Those are precisely the two that form the memory. The average may be fine. The memory will not be.

The journey averages 2.4/5, but peak and end together average 1.5. You are being judged on the lower number while reporting the higher one.

07 THE UNCERTAINTY

The not knowing, the waiting after, and what you promise

5
steps where they do not know what happens next

2
post-process waits, which feel longest of all

3/9
of the wait designed, against Maister

+38
points between reality and perception

Uncertainty is not a softer version of a bad outcome; it is worse than one. Stress responses track irreducible uncertainty rather than the likelihood of the bad thing, and are lower when something bad is certain than when it is merely possible. Every uncertain step is cheap to fix with information and expensive to leave. Certainty beats optimism.

> Perception trails reality by 38 points. If customers already believe you are worse than you are, more delivery improvement is the wrong investment: satisfaction is what people perceive minus what they expected, so a gain nobody notices does not move it. Prevention is invisible by definition, so the language and the reporting have to do the showing.

Of 6 commitments classified, 3 are must-be quality: they cause dissatisfaction when absent and earn almost nothing when present. Improving a must-be attribute past adequacy costs real money and buys very little. That is where a great deal of service investment quietly goes to die.

08 THE BLUEPRINT

Where the work should sit

MODE	DESIGNED	REFERENCE	READ
Self-heal	55%	60%	in line
Guided	22%	25%	in line
Assisted	16%	10%	above reference
White-glove	7%	5%	in line

This is the design, not the diagnosis. Section 03 showed where your demand scores: 73 per cent of it for self-heal. This table shows where you have chosen to put it: 55 per cent. Neither number is wrong. The 18-point gap between them is the decision you are making about how much of what could be automated you are willing to automate, and by when, and it is better made deliberately here than discovered by a board later.

Where it hurts today. 1. Chase-ups and re-work, roughly 41% of volume.

What the ticket hides. Two account teams keep private spreadsheets because they do not trust the tool. Customers ring the engineer they like directly. Several have stopped raising the small stuff and just live with it. That hidden layer is the real operating model: the workarounds are where your people already redesigned the service without permission.

The human dividend. Two roles moved into a named major-incident and month-end team. That is what the freed capacity buys, and we will say so before the first automation lands. Naming it matters: unnamed capacity is read as slack, and then cut.

09 THE SENSING

What moves before a ticket exists

5	Instrumented	1
real sensors, of 6 rows	maturity on the sensing ladder	still dashboards, not sensors

A sensor completes one sentence: when a signal moves, we act, decided by an authority, within a time. 5 of your rows do. They read 1 from users, 2 from systems, 2 from business.

- Failed logins spike on one account (Systems) to Force step-up auth and notify the user, within 2 minutes. Owner: Identity lead.
- Ticket passes half its target with no update (Business) to Send the customer a real update with a next step and a time, within Immediate. Owner: Service owner.
- Error rate rises within 60 mins of a change (Systems) to Hold the change and call the customer before they call us, within 15 minutes. Owner: Change manager.
- Repeat contact from the same user in 7 days (Users) to Route to a named engineer, not the queue, within 1 hour. Owner: Desk lead.
- Month-end approaching for a finance customer (Business) to Pre-brief the white-glove team and freeze non-urgent change, within Weekly. Owner: Account lead.

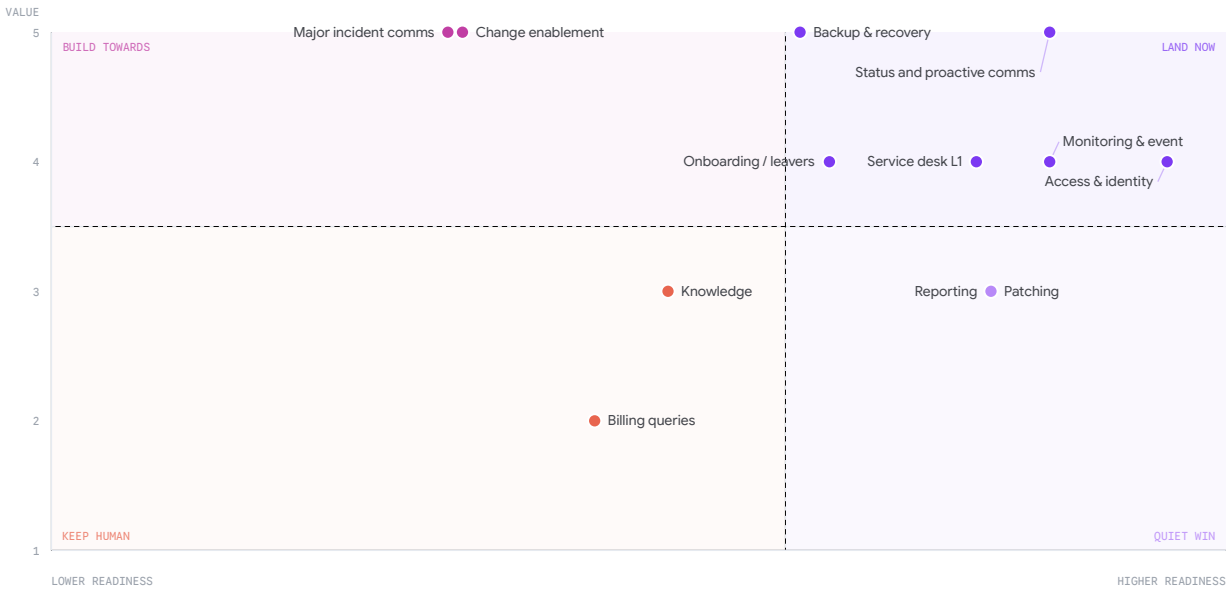
10 WHERE AI LANDS

Landing zones, in sequence

64 per cent of UK organisations say they are using AI; only around 24 per cent are anywhere near mature adoption. The gap is organisational design, not technology, which is why this sits after the operating model rather than before it.

Where AI lands

Readiness across the bottom, value to the business up the side. The lines cross at 3.5.



Land now Quiet win Build towards Keep human

FUNCTION	READINESS	VALUE	ZONE
Status and proactive comms	4.40	5	Land now
Backup & recovery	3.55	5	Land now
Access & identity	4.80	4	Land now
Service desk L1	4.15	4	Land now
Monitoring & event	4.40	4	Land now
Onboarding / leavers	3.65	4	Land now
Patching	4.20	3	Quiet win
Reporting	4.20	3	Quiet win
Change enablement	2.40	5	Build towards
Major incident comms	2.35	5	Build towards
Knowledge	3.10	3	Keep human, for now

Billing queries	2.85	2	Keep human, for now

The sequence is not negotiable. Land where volume is high, the process is deterministic and the blast radius is small. Bank the quiet wins next. Build towards the high-value work that is not ready. Never lead with the moment of dread, however tempting the visibility.

Your first landing is **Status and proactive comms**, at readiness 4.40 and value 5. Take it to production with a policy layer and a rollback, and let it earn the next one.

11 THE LANGUAGE

Selling confidence, not acronyms

3.3
insider terms per ten words, across all 6 lines audited

2.3
before, on the 3 lines you rewrote

0.0
after, on the same 3 lines

3/6
lines rewritten

The comparison that means something is the middle pair, and it is like for like: on the lines you rewrote, density fell from **2.3** to **0.0**. The 3.3 is the average across every line audited, including the 3 you have not touched yet, so it is the size of the job rather than a measure of the improvement. Setting one against the other would overstate what you have done.

BEFORE	AFTER
P1 Incident Resolution within contracted SLA, including RCA and Problem Management	When something critical stops, a named engineer is on it within fifteen minutes and stays with it. If it happens twice, we find out why and stop it.
Proactive Monitoring & Event Management across the estate	Most problems are fixed before you feel them. Each month we show you exactly what never reached you.

Request Fulfilment via the Service Catalogue with defined OLAs	Ask for what you need in one place, and it arrives the same day without you chasing it.

Clean language does not stay clean. One owner of words, a read-aloud gate, a quarterly strip test and customer sampling twice a year. Watch the four places jargon re-enters: contracts, the ticket-tool surfaces customers actually see, bots trained on the old knowledge base, and new joiners who learn the acronyms before they learn the customer.

12 THE PLAN

The first ninety days

Three sequencing rules hold this together, and getting them in the wrong order is how shift right fails. **Instrument before you automate. Automate before you promise.** And the one everyone gets backwards: **create the white-glove roles before you remove the routine work**, because people need somewhere to shift *to*, or the change reads, correctly, as headcount reduction with better branding.

DAYS 1 TO 30

Sense and map

You have already run the demand and flow work, so use this month to validate it against a live month and baseline your signals.

- Baseline failure demand and flow efficiency for one service.
- Walk the mapped journey with three real customers and re-score it.
- Kill the noisiest useless alert, publicly. Removing noise earns the right to add signal.

DAYS 31 TO 60

Design and prove

Design the target, then prove one thing works before promising anything to anyone.

- Remove the cause of Chase an update on something already logged, your largest source of failure demand. Do not make it faster to handle.
- Take one self-heal candidate live with a confidence threshold and a rollback right.
- Put a person back on the ending. A short human confirmation that it is genuinely fixed, before closure.
- Pilot the language rewrite on your top few catalogue lines.

DAYS 61 TO 90

Shift and tell

Allocate the modes, switch the metrics, and then do the part almost everyone skips.

- Set mode allocation v1 for the service and staff the white-glove roles by name.
- Cap work in progress at 350 and drain towards it gradually.
- Switch this service's reporting from SLA-thinking to the mode metrics below.
- Publish the prevention internally. Show the organisation what never reached it. That is the story your board has never been told, and it is what funds the next ninety days.

The metrics that replace SLA-thinking

An SLA measures whether you kept a promise about speed. It says nothing about whether the work should have existed, or whether the customer left. Measure each mode by what it is actually for.

Self-heal

Prevented contacts, time-to-detection, and autonomy rate reported beside rollback rate. If you cannot show what never happened, you are measuring deflection, not prevention.

Guided

Completion without escalation, and a real effort score. Did they get there themselves, not did a ticket close.

Assisted

Resolution quality, and the customer's confidence at the point of close. Not "was it fixed" but "do they believe it is fixed".

White-glove

Outcome delivered, and relationship depth. These are the moments you are judged on for years.

RETIRE THESE ON SIGHT

- × Watermelon SLAs. Green outside, red inside. The report is met and the customer is furious.
- × Deflection worship. Counting contacts avoided as a virtue without asking why the contact existed.
- × Averages that hide the two per cent. The mean looks fine while the cases that decide your reputation burn.

One governance rule underneath all of it: an agent acts within policy only above a stated confidence threshold, a human approves below it, and a human investigates below that. **Every automated change carries a rollback, every action leaves an audit trail, and a named human can revoke autonomy at any time.** Say those three out loud and most of the fear in the room goes with them.

·· THE BOARD CASE

One paragraph, your numbers

WRITTEN TO BE READ ALOUD

From the cost curve to the value curve

41%

of demand created by our own earlier failure

1.1%

of elapsed time our work is actually moving

£288k

a year spent handling our own earlier failures

2,765

contacts a month handled by people who should not see them

“Northwind Managed Services is moving from a cost curve to a value curve. Today, effort scales with demand: more tickets, more people, more cost. But 41 per cent of what arrives is failure demand, work that exists because we failed someone earlier. A share of our revenue is therefore paid to us for handling our own defects, which is a commercial model that quietly rewards the thing we say we want to eliminate. On our own costing that is about £288,000 a year. And our work is genuinely moving for only 1.1 per cent of the time we hold it, which means the problem is not how hard people work. It is how long work waits. We are putting software on the volume that never needed a person, and our people, deliberately, on the moments that decide whether customers renew. When every provider runs the same models, competence is a commodity and judgement is the only ground left where we are not competing on price. We are starting with the ending, which we currently leave to a machine, and which is most of what anyone remembers.”

AND WHAT NOT TO PROMISE YET

Do not promise the automation is safe before it has run with a rollback in anger; promise that it is reversible and audited. Do not promise headcount savings; promise capacity redirected, and name where it goes. Do not promise the metrics switch overnight: the old measures and the new will run side by side for a while, and saying so is honest rather than weak. And do not put a single one of these numbers in front of a customer until the contract has been read against them. Some of what is in this document is an argument for renegotiation, and it is better to arrive with that deliberately than to have a customer find it first.