The Decision Flow Problem: Why Your Supply Chain Planners Spend 95% of Their Time Not Planning

How BCG's Rules of Response apply to information flow — and why autonomous agents change the economics of supply chain decisions.

In 1987, George Stalk Jr. of the Boston Consulting Group published a deceptively simple observation about corporate operations. He called them the "Rules of Response," and they described how time moves through value-delivery systems. The insight was this: **products and services receive value for only 0.05 to 5 percent of the time they spend in a company's system**. The rest — the 95 to 99.95 percent — is waiting.

Stalk was talking about physical goods. A heavy vehicle manufacturer takes forty-five days to prepare an order for assembly, but only sixteen hours to actually assemble each vehicle. The vehicle is being worked on for less than 1 percent of the time it spends in the system. The remainder is waiting: waiting for a batch to complete, waiting for the batch ahead of it, and — critically — waiting for management to make and execute the decision to move it to the next step.

That third category of waiting is where things get interesting for us.

The Invisible Batch: Decisions

Stalk's three sources of waiting time divide almost equally. A third of the waste comes from waiting for batches to fill. A third from waiting for prior batches to clear. And a third from waiting for someone to decide what happens next.

Now consider what a supply chain planning organization does. It doesn't move physical goods. It moves *decisions* through an organization. A demand planner detects a shift in customer ordering patterns. That signal must flow to the supply planner, who adjusts replenishment. The change propagates to the MPS manager, who resequences production. The allocation manager redistributes available-to-promise. The procurement analyst places revised purchase orders.

Each of these is a decision. And each decision, like Stalk's physical goods, spends almost all of its time waiting — not being made.

The .05 to 5 Rule applies to decisions too. A procurement analyst might spend five minutes deciding to expedite a purchase order. But the information that triggered that decision may have been sitting in a report for three days, waiting to be noticed. Before that, it waited two days in an exception queue. Before that, the underlying demand signal waited a week to be aggregated into the next planning cycle.

The decision itself takes minutes. The time it spends in the system? Weeks.

Why Decisions Wait

In 2018, Ajay Agrawal, Joshua Gans, and Avi Goldfarb — economists at the University of Toronto's Rotman School of Management and co-founders of the Creative Destruction Lab — published *Prediction Machines*. Their thesis was elegant: **AI is best understood as a drop in the cost of prediction.** And when the cost of prediction drops, the economics of every decision that depends on prediction changes.

But they made a subtler point that's often overlooked. Every decision, they argued, has an anatomy: data goes in, prediction happens, judgment is applied, and action comes out. When prediction was expensive, organizations batched decisions. They gathered data weekly, ran forecasts monthly, and held planning meetings quarterly. This wasn't because monthly was the right cadence for decisions — it was because prediction was expensive enough that you had to batch it to justify the cost.

© 2026 Azirella, Inc. All rights reserved. | azirella.com

Sound familiar? **Planning organizations batch decisions for the same reason factories batch production runs: to amortize the setup cost across enough units to make it economical.**

The "setup cost" for a supply chain decision isn't a die change on a press. It's the time a planner spends gathering context. Pulling up inventory positions across six warehouses. Cross-referencing the forecast with the latest sales orders. Checking supplier lead times. Reviewing the capacity plan. Reading three emails from the regional sales manager about a customer promotion.

By the time the planner has assembled enough context to make a good decision, forty-five minutes have elapsed. The decision itself — increase the safety stock target by 200 units at the Dallas DC — takes seconds.

This is Stalk's .05 to 5 Rule, applied to information work. **The value-adding act (the decision) occupies a tiny fraction of the total cycle time. The rest is information gathering, context assembly, and organizational coordination.**

What Happens When Decisions Become Cheap

Agrawal, Gans, and Goldfarb observed that when technology makes something cheap, you use more of it. When electricity made lighting cheap, factories didn't just replace gas lamps — they redesigned buildings, extended operating hours, and invented the assembly line. The resistance to adoption, they noted, only breaks when the technology makes something unambiguously cheaper, faster, or better.

So what happens when you make supply chain decisions cheap?

Not cheaper in quality — cheaper in *cycle time*. What if the elapsed time from "demand signal changes" to "corrective action taken" dropped from days to seconds?

This is the promise of autonomous supply chain agents. Not that any single decision is better than what a skilled planner would make. A well-trained agent making an inventory rebalancing decision and an experienced supply chain analyst making the same decision will arrive at similar conclusions. The atomic decision quality is comparable.

The difference is everything that surrounds the decision.

The agent doesn't spend forty-five minutes gathering context. It already has context — it's continuously monitoring inventory positions, demand signals, supplier status, and capacity constraints across every site in the network, simultaneously. It doesn't wait for the weekly planning meeting to surface an exception. It detects the exception the moment the underlying data changes. It doesn't wait for a handoff between the demand planner and the supply planner and the allocation manager. It operates across the full decision chain in a single pass.

A human planner can process perhaps twenty to thirty meaningful decisions per day. Not because they're slow thinkers — because each decision requires extensive context gathering, cross-functional coordination, and organizational navigation. The decision itself is the easy part. The hard part is everything before and after.

An autonomous agent processes the same decisions in seconds, twenty-four hours a day, seven days a week. Not because it's smarter. Because it has eliminated the waiting time between decisions.

Stalk's Rules, Reframed

Let's revisit BCG's Rules of Response through the lens of decision flow:

The .05 to 5 Rule for Decisions: In most planning organizations, actual decision-making represents 0.05 to 5 percent of the total elapsed time. The rest is gathering information, waiting for the planning cycle, coordinating across functions, and shepherding the decision through approval workflows.

The 3/3 Rule for Decisions: The waiting time divides roughly into thirds: waiting for enough information to accumulate (the "batch"), waiting for other decisions ahead in the queue (the "sequence"), and waiting for

© 2026 Azirella, Inc. All rights reserved. | azirella.com

organizational bandwidth to process the decision (the "capacity").

The 1/4-2-20 Rule for Decisions: For every quartering of the decision cycle time, labor productivity doubles and costs fall by 20 percent. A company that moves from weekly planning cycles to continuous agent-driven planning doesn't just make faster decisions — it fundamentally changes the economics of its planning organization.

The 3 x 2 Rule for Decisions: Companies that compress their decision cycle times grow at three times the industry average with twice the profit margins. This is the competitive moat: not better algorithms, but faster *organizational metabolism*.

The Real Argument for Agents

The case for autonomous supply chain agents is not that they make better individual decisions than humans. That argument is difficult to prove and easy to contest. A seasoned planner with thirty years of experience and deep institutional knowledge will sometimes outperform any algorithm on a complex, ambiguous, novel situation.

The case is that agents **eliminate the waiting time between decisions.**

A supply chain doesn't fail because one person made one bad call. It fails because a thousand small signals went unnoticed for too long. A slight uptick in demand at three retail locations. A two-day delay at a port that affects four inbound shipments. A quality hold on one production batch. Each of these, detected and addressed in hours, is a minor adjustment. Left unattended for days or weeks — because the planning organization only reviews exceptions on Tuesdays, or because the analyst responsible was on vacation, or because the signal was buried on page four of a forty-page report — they compound into stockouts, expediting costs, and missed service levels.

Agents don't sleep. They don't take vacation. They don't have forty-five-minute context-gathering sessions. They don't batch decisions into weekly cycles because the setup cost of each decision is effectively zero.

They make the system better not by making any single decision better, but by making *all* decisions sooner.

The Economics of Continuous Correction

Agrawal and his colleagues would frame it this way: agents make prediction cheap enough that you can afford to predict — and therefore decide — continuously rather than periodically. When prediction was expensive (gathering all that context manually), you predicted weekly. When prediction is cheap (agents monitoring everything in real-time), you predict continuously.

And Stalk would frame it this way: the company that compresses its decision cycle time from weekly to continuous has applied the 1/4-2-20 Rule not once, but repeatedly. Each quartering — weekly to daily, daily to hourly, hourly to real-time — doubles productivity and cuts costs by 20 percent.

The multiplication is what matters. Not 20 percent once. Twenty percent compounded across four quarterings of cycle time. The fast-response planning organization doesn't just plan better. It plans in a fundamentally different mode — continuous correction rather than periodic replanning.

This is what the building materials manufacturer in Stalk's paper achieved when it cut order-to-delivery from five weeks to one. Not by working harder. By eliminating the waiting time in its system. Its growth rate exceeded the industry average by more than three to one. Its return on assets was double the competition.

The same economics apply to the flow of decisions through a planning organization. The company that detects and corrects course in seconds will outperform the company that detects and corrects course in weeks — not because its planners are smarter, but because its decision cycle time is shorter.

That is what agents make cheaper, faster, and better. Not the decision. The flow.

© 2026 Azirella, Inc. All rights reserved. | azirella.com

This is the first in a series on the economics of autonomous supply chain planning. The author is the founder of Autonomy, an AI-native decision intelligence platform for supply chain.

References:

- Stalk, G. Jr. (1987). *Rules of Response*. The Boston Consulting Group, Perspectives No. 317.
- Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Review Press.
- Agrawal, A., Gans, J., & Goldfarb, A. (2022). *Power and Prediction: The Disruptive Economics of Artificial Intelligence*. Harvard Business Review Press.

© 2026 Azirella, Inc. All rights reserved. | azirella.com