

Who Gets It First: Rollout Order and Checking in a Simulated Organization

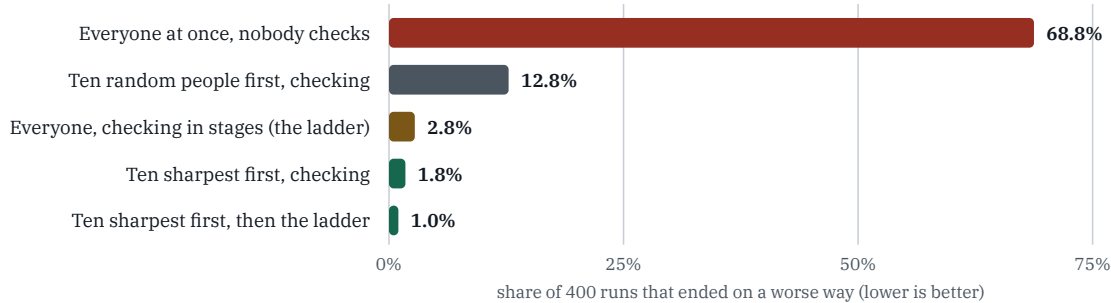
A people-first rollout against a staged capability ladder, read here as staged checking by everyone, and a few builders against everyone, run on the honey bee voting model with an org chart added, and set against earlier work on herding, social learning, diffusion, automation and organizational search.

Berg Atkinson, Wolfberg LLC · berg@wolfberg.ai

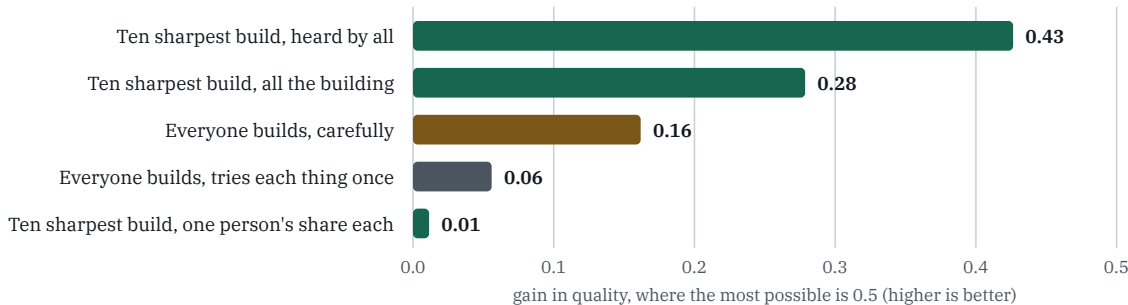
Preprint. Reviewed by a panel of AI reviewers (Claude Opus 5.5) and revised 7 October 2026; not externally peer reviewed. First published 3 October 2026.

About the author. Berg Atkinson: 27 years in defense, intelligence, commercial and international technology, including one of five Chief Engineers on the Air Force ABMS Digital Infrastructure Consortium and advisor to Leidos sector CTOs; since April 2026, sole operator of Wolfberg LLC with an AI team. The positions in this paper rest on that experience; “From one practice” means Wolfberg. [linkedin.com/in/bergatkinson](https://www.linkedin.com/in/bergatkinson)

Choosing among ways of working that exist: how often the organization got a close call wrong



Building new ways of working: how much better the organization worked after 200 rounds



At a glance. The two results. Above: how often a simulated 200-person organization ended a close call on a worse way of working, by rollout plan. Below: how much better it worked after 200 rounds when members build new ways of working, by who builds; the bottom bar is a budget control, in which the ten build at one person's rate and nobody else builds. **Simulated; 400 runs per bar; not a study of people.**

IN PLAIN TERMS

When an organization gets a new AI tool, it has to work out the best way to use it, and most of that working out happens by people copying each other. There are two common ways to stage the rollout. One stages what the tool is allowed to do: everyone gets it, and the human checking is heavy at first and lighter at each stage until the tool runs on its own. The other stages who gets it: a small group of the strongest people get it first, figure it out, and everyone else learns from them. In a recorded interview the author argued for the second, and said it plainly: give your A-team a team, not a tool.

We tested both on the same model we used for our swarm research note, the weighted voter model of how honey bees choose a nest, with an org chart added: 200 people in 25 teams, team leads whose word counts more, an executive, a habit of deferring to the boss, and people who differ in how well they read results. On a close call, giving the tool to everyone at once with nobody checking left the organization on a worse way of working 69% of the time. The staged ladder brought that to 3%, and giving it to the ten sharpest people first brought it to 2%, with about a sixth of the checking, because only ten people did the checking; the two are too close for our runs to tell apart. Both plans work, and for the same reason: somebody checks before the organization copies.

They cost different things and fail in different ways. The ladder had nobody waiting and most of the organization on the best way within a few dozen rounds; the people-first plan made everyone else wait while the first group worked it out, though a short head start with thorough checking did nearly as well, at about three times the checking. The ladder's results depended on the checks being real, because skipping half of them left it wrong 30% of the time. When the best way changed, the finished ladder could not follow, and the ten who kept checking and trying things brought the organization along in 38% of runs. The ladder's failure is built into the model, where nobody on it ever tries anything new; a ladder whose members did, with a check, followed most changes. The people-first plan depended on who went first: ten random people did much worse than the ten sharpest, and the team leads going first with no checks at all did nearly as badly as no plan (we did not run the leads first with checks). The underlying reason is that a check is only as good as the person doing it; in the model people differ only in how accurately they read results. Much of this has precedents in research on herding, social learning and the diffusion of new practices, and Section 3 says which parts.

Choosing among ways of working that already exist is only half of what a tool like this is for. The other half is building new ones, so we added building to the model: anyone with the tool can make a new way of working, and a new way is as likely to be worse as better, whoever makes it. With the same amount of building and the same care in checking, the ten sharpest people improved how the organization worked about 1.7 times as much as everyone building; against everyone trying each new thing only once the gap was 5 times, but that comparison also gives the ten more careful checks. Given only one person's share of the building each, with nobody else building, they made a twentieth of the builds and barely moved it. So the few do more only when they get the capacity, which is the "give your A-team a team" half of the advice. Giving their work an audience, in the model making the ten a tenth of what everyone hears, did more than ordering everyone to adopt it, an order that people could check first did better than one they could not, and the org chart, which pulls people toward their bosses' choices, held back everyone building far more than it held back the few. On a harder problem, one with many local peaks

where any search tends to stop at the first good-enough answer, who built barely mattered, and bigger changes mattered more.

ABSTRACT

We compare two ways of staging an AI rollout in an agent-based model of an organization. The first stages what the tool is allowed to do; the model has no tool permissions, so we read it as staged checking by everyone: each member checks before switching to a new way of working, heavily at first and less at each stage (a capability ladder). The second stages who has the tool: the ten members who read results with the least noise have it alone for a head start, checking and trying alternatives, and everyone else then starts from what they found. The model is the weighted voter model of honey bee nest-site selection (Valentini, Hamann & Dorigo, 2014; Soma et al., 2024), as used in our earlier note (Atkinson, 2026a), with an organization added: 200 members in 25 teams, leads whose votes count three times, an executive, deference to the boss and unequal reading noise. On a close call among five options, over 400 runs per cell, the organization settled on a worse option in 68.8% of runs with everyone getting the tool at once and nobody checking, 2.8% under the ladder and 1.8% with the ten sharpest first, a difference the runs cannot resolve. The ladder paid in checking, with more than six times as many check trials, and the people-first plan in waiting, with nearly a quarter of all member-rounds spent without the tool. Within each plan, outcomes tracked the checking: ten random members in place of the ten sharpest left 12.8% wrong, and the ladder with half its checks skipped 30.0%. When the best option changed after the rollout, the finished ladder followed it in 0 runs and members who kept checking and trying alternatives brought the organization along in 37.8% to 92.7%; the ladder as built has no exploration, and with checked exploration it followed in 69.5%. In a second model the tool builds: members who hold it make new ways of working, each as likely to be worse as better. With the same total building and the same care in checking, the ten sharpest raised the organization's mean quality by 0.279 by generation 200, against 0.162 with everyone building; making the ten 10% of what every member hears raised it by 0.426, more than periodically ordering everyone onto their best way (0.286). On a rugged NK landscape every plan stopped on a local peak and who built mattered little. We relate the results to informational cascades, critical social learning, network interventions, levels of automation and organizational search.

Keywords: AI adoption, organizational learning, social learning, informational cascades, levels of automation, organizational search, NK landscape, collective decision-making, agent-based model

CONTENTS

1	Introduction	5	Results
2	Two ways to stage a rollout, and how we read them	6	When the tool builds
3	What was already known	7	What it means, and three valleys (7.4)
4	Methods	8	Limits
			References

1 Introduction

Most advice on rolling out AI inside an organization is about the tool: which tasks it may take on, how much a person must review, and when it may act alone. Less of it is about the people: who should have the tool first, and whose way of using it everyone else should learn from.

The author argued the second position in a recorded interview (York, 2026). His advice to a company that wants the gains it hears about elsewhere was to give its A-team a team, not a tool, to keep its senior people because they are the ones who can teach juniors to lead, and to start with the solution architects or chief engineers and let the rest of the organization see the results. In the same interview he also described trust in stages: a new employee does not get the company card and the keys to everything on the first day, and he does not let a human team run on its own until it has earned it. On 2 October 2026 Anthropic announced a program that matches the first half of that position, which has organizations nominate their strongest engineers, trains and grades them, and sends them back to lead a real project (Anthropic, 2026). It is a vendor's training program, its first cohorts drawn from consulting firms such as Accenture, Deloitte and McKinsey and from large companies such as Morgan Stanley.

Most organizations deploy something else: a seat for everyone, measured by use (Figure 1). Microsoft reported more than 30 million paid Microsoft 365 Copilot seats on 29 July 2026 and described "high usage" as monthly active usage above 80% across a customer's user base (Microsoft, 2026). When the United Kingdom's Department for Business and Trade evaluated its own Copilot pilot, 72% of the respondents to its diary study were satisfied or very satisfied and many participants reported saving time, but the evaluation did not find robust evidence that the time saved had led to improved productivity, which it notes was not a key aim (Department for Business and Trade, 2025). The programs that follow the author's position look different. Besides Anthropic's, OpenAI places forward deployed engineers inside customer companies, and the head of that team has said that in about 80% of the cases where an enterprise is stuck, the problem is deployment rather than the model (The Next Web, 2026). After the interview the author of this note put the difference this way: deploying a tool to everyone is not the same as having a few strong people use it to build for everyone, and he expected the few to do far more.



Figure 1. Two kinds of program. Every figure here is from the source named beside it and cited in the text.
Reported by the organizations named; not our data.

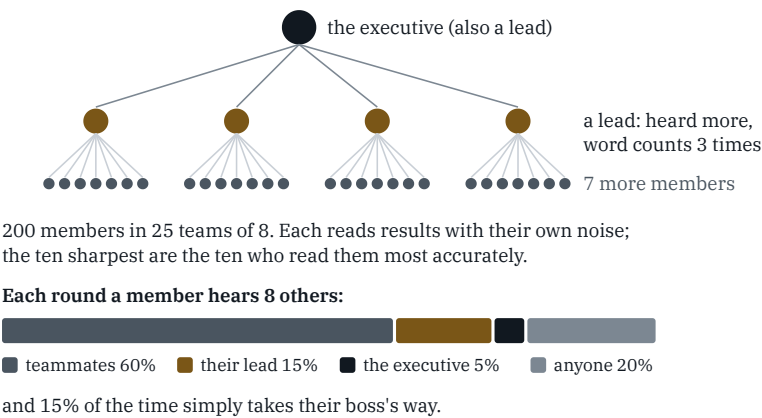
This note tests the position. We asked five questions. On a close call, does staging who gets the tool first do as well as staging what the tool may do? What does each plan cost along the way, in time, in checking and in work done the wrong way? What does each plan depend on, and how does it fail? What happens after the rollout, when the world changes? And when the tool is used to build new ways of working rather than to choose among existing ones, does it matter who builds, and with how much capacity? We used the model from our earlier note on swarms (Atkinson, 2026a) because it is a model of exactly this situation: a population settling on one of several options by copying each other, with no one able to see which option is best. Section 2 sets out the two plans and how we turned them into rules, Section 3 says what earlier work already established, Section 4 gives the methods, Section 5 the results, Section 6 a second model in which the tool builds, and Section 7 what they imply.

2 Two ways to stage a rollout, and how we read them

A capability ladder stages what the tool does. Four-stage ladders now proposed for enterprise AI (we describe the common form rather than any one framework) run from analysis, where the tool informs a human who does the work, to autonomy, where the tool acts and the human supervises less or not at all. They restate, for a new technology, the much older idea of levels of automation: automation can be applied to acquiring information, analyzing it, selecting decisions and carrying them out, each at a level from fully manual to fully automatic, and the level should be chosen by its consequences for human performance, with the reliability of the automation and the cost of its errors as secondary criteria (Parasuraman, Sheridan & Wickens, 2000).

A people-first rollout stages who has the tool. A small group gets it first, works out how to use it, and the rest of the organization starts from what that group found. In the interview the group is the organization's A-team; in Anthropic's program it is the strongest engineers each organization nominates.

The organization



The plans, by round

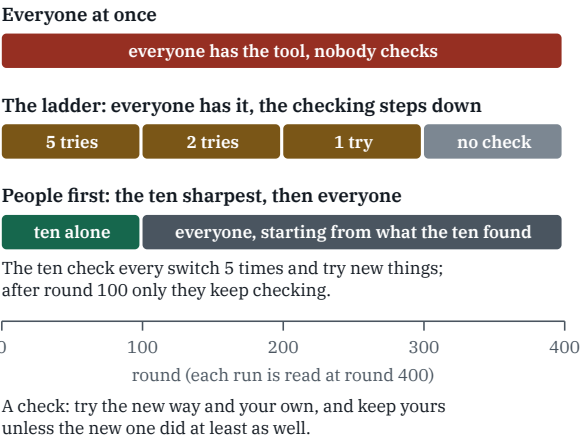


Figure 2. The organization, and the three rollouts compared in Sections 4 and 5. A diagram of the model; no data.

The model in Section 4 has no tool that acts; it has members choosing among ways of working, and copying each other. So we had to decide what each plan means in that model, and the result rests on the reading (Figure 2). We read the ladder as how much each member checks before switching: in its first stage every member who is about to adopt a different way of working first tries the new way and their own five times each and keeps their own unless the new one did at least as well; in later stages the check is two tries, then one, then none, at which point everyone copies freely, which is what autonomy means here. We read the people-first plan as the ten members who read results with the least noise having the tool alone for a head start, hearing only each other, trying a random alternative in 5% of their updates, and checking every switch five times; then everyone else gets it, and each newcomer starts by hearing eight of the ten. After that the organization runs as usual: nobody else checks, and the ten keep checking and trying alternatives as ordinary members. "Both" is the people-first plan followed by the ladder. "Everyone at once" is neither: everyone has the tool from the start and nobody checks. The ladder, as built, has no exploration: nobody on it ever tries a random alternative (Section 5.7). And because the organization settles within the ladder's first stage (Section 5.5), the ladder runs test, in effect, everyone checking five times for 100 rounds.

The reading of the ladder deserves its own caution. A real ladder stages what an automated system is permitted to do, so that a harmful action is caught while trust is earned. Our model has no harmful actions, only better and worse ways of working; what it can count is how much work was done the worse way, and how much of that happened with no check in force (Section 5.2).

3 What was already known

What earlier work had already established. One line each; the fuller account follows.

What was known	Source	Where it bears here
----------------	--------	---------------------

What was known	Source	Where it bears here
People who follow those before them can lock a group into a wrong choice, and the result is fragile	Bikhchandani, Hirshleifer & Welch (1992); Banerjee (1992)	Everyone at once (5.1)
Copying alone does not help a population; copying and then checking for yourself does	Rogers (1988); Enquist, Eriksson & Ghirlanda (2007)	The check (5.3)
Copying works well when people show the best-paying thing they know	Rendell et al. (2010)	The flat organization (5.8)
Who adopts first shapes what spreads, and networks can be used to speed a change	Rogers (2003); Valente (2012)	Who goes first (5.4)
People over-trust automated aids, experts included, and training does not prevent it	Parasuraman & Manzey (2010)	Skipped checks (5.5)
Adaptive processes refine what is known faster than they explore: good in the short run, self-destructive in the long run	March (1991)	Building (6)
On complex problems, networks that spread solutions efficiently do better early and worse late; semi-isolated subgroups with moderate links do best	Lazer & Friedman (2007); Fang, Lee & Schilling (2010)	The rugged test (6.5)
Which network wins depends on how people copy, and with real people efficient networks won	Barkoczi & Galesic (2016); Mason & Watts (2012)	The rugged test (6.5); limits (8)
Where a landscape has several peaks, the starting point shapes which one is found	Levinthal (1997)	The rugged test (6.5)
AI assistance helped less skilled support agents most, but in a field experiment hurt weaker entrepreneurs and may have helped stronger ones; individual performance may be highly skewed, though that is disputed	Brynjolfsson, Li & Raymond (2025); Otis et al. (2026); O'Boyle & Aguinis (2012); Beck, Beatty & Sackett (2014)	Who builds (6.2)
A large swarm following the weighted voter rule moves like a single reinforcement learner	Valentini, Hamann & Dorigo (2014); Soma et al. (2024); Atkinson (2026a)	The model (4)

The failure of the rollout with no checks has a name. In an informational cascade, a person who has watched the choices of those before them follows them without regard to their own information, and mass behavior built that way is both locally uniform and fragile (Bikhchandani, Hirshleifer & Welch, 1992). In the same year Banerjee (1992) showed that sequential decision makers who rationally take account of each other's choices end up doing what others do rather than using their own information, and that the resulting equilibrium is inefficient. An organization in which everyone copies, and in which some voices count more than others, is a setting in which a cascade on the wrong option is easy to start.

The success of checking also has precedents, in research on social learning. Rogers (1988) presented a model of the evolution of social learning whose conclusion, that learning by copying does not increase a population's fitness, became known as Rogers's paradox (we take this statement of it from Enquist, Eriksson & Ghirlanda, 2007, and read Rogers's article in abstract only). Enquist and colleagues resolved the paradox with "critical social learning": an individual first copies, then learns for itself if the copied behavior proves unsatisfactory. In their model that strategy always did better than pure copying and typically better than pure individual learning, and it was evolutionarily stable unless copying was very unfaithful, the environment very variable or copying much more costly than learning alone. The check in our ladder is a version of it: copy a candidate, try it against your own, and keep your own if the candidate does worse. A computer tournament of social learning strategies found that strategies relying heavily on copying did remarkably well, because individuals tend to show the highest-paying behavior they know, which filters information for the people copying them (Rendell et al., 2010). That is the condition our flat organization meets and our hierarchical one breaks (Section 5.8).

That who adopts first shapes what spreads is the core of research on the diffusion of innovations, with its early adopters and opinion leaders (Rogers, 2003, cited here for that standard framing and not re-read). Using social network data to accelerate a behavior change is now a recognized family of interventions, with four broad strategies and many tactics within each (Valente, 2012).

Work on automation adds a warning that applies to the people who check. Automation complacency and automation bias, the tendency to accept an automated aid's output and to miss what it misses, appear in experts as well as novices; bias is not prevented by training or instructions, and it affects teams as well as individuals (Parasuraman & Manzey, 2010). A ladder whose human checks degrade into approvals has, in effect, fewer checks than it appears to, which is what our skipped-check runs model (Section 5.5).

Research on organizational learning and search bears on building. March (1991) argued that adaptive processes refine exploitation more rapidly than exploration, and so are likely to become effective in the short run but self-destructive in the long run. Levinthal (1997) modeled organizational change on a rugged fitness landscape, using Kauffman's NK model (Kauffman, 1993, cited here for that standard framing and not read), and found that where the landscape has several peaks, an organization's starting position shapes which peak it finds. In agent simulations in which the less successful copy the more successful, networks that spread information efficiently did better in the short run and worse in the long run on complex problems (Lazer & Friedman, 2007), and an organization divided into semi-isolated subgroups with moderate links between them reached the highest equilibrium performance (Fang, Lee & Schilling, 2010). Which network does better depends on how people copy: efficient networks did better when members copied the most frequent solution among their contacts, and inefficient ones when they copied the member with the highest payoff, a strategy that did best on simple tasks (Barkoczi & Galesic, 2016). People need not behave like these agents. In 256 online experiments with groups of 16 people, efficient networks did better than inefficient ones even on a problem thought to favor inefficient networks, which the authors explain by the participants' own decisions about when to explore (Mason & Watts, 2012).

On who gains from generative AI, the evidence is mixed. In the abstract of their study of 5,172 customer-support agents, Brynjolfsson, Li and Raymond (2025) report that AI assistance raised issues resolved per hour by 15% on average; less experienced and lower-skilled agents improved in both speed and quality, while the most experienced and highest-skilled saw small gains in speed and small declines in quality. In a field experiment with Kenyan entrepreneurs given an AI business assistant, Otis et al. (2026) found no detectable average effect on revenues and profits, but those who were low performers at the start did worse with the assistant and the high performers may have done better. Separately, and not about AI, O'Boyle and Aguinis (2012) report in their abstract that across 198 samples of 633,263 researchers, entertainers, politicians and athletes, individual performance followed a power law rather than a normal distribution, so a small number of people account for much of what gets done; others attribute much of that skew to how performance was measured (Beck, Beatty & Sackett, 2014). None of these studies tests what happens when a few people use AI to build for everyone else.

The model itself comes from swarm robotics. The weighted voter model was introduced as a simplified account of how honey bees choose a nest, studied for a choice between two sites, and found to choose more accurately in bigger swarms (Valentini, Hamann & Dorigo, 2014). Soma et al. (2024) showed that a very large swarm following it moves, on average, like a single reinforcement learner. Our earlier note (Atkinson, 2026a) reproduced that result, then showed that smaller swarms settle on worse options, that a little random scouting undoes it, that a few members who never change their minds can take over a close contest, and that checking a claim before copying it is the defense against exaggerated reports. That note ended with a hypothesis about teams that this one moves toward testing, in simulation.

Against this background, what is new here, as far as our search found, is putting both ways of staging a rollout into one model of an organization and pricing them in the same units: wrong calls, member-rounds spent waiting or working the wrong way, and checks performed. We also show that, within each plan, outcomes track how accurately the checkers read results, and that after a change no plan follows it unless someone keeps trying alternatives. In a second model we compare concentrating the building of new ways of working in a few strong members with spreading it across everyone, at equal and at unequal capacity, and an audience for the few with an order to follow them, on a smooth and on a rugged landscape. Our search was not systematic, and we may have missed relevant work, in particular in organizational behavior and the management literature on pilots and rollouts.

4 Methods

The organization. The model is the organization version of our swarm simulator (Atkinson, 2026a). Two hundred members are arranged in 25 teams of eight; the first member of each team is its lead, and member zero, a lead, is also the executive. The organization chooses among five ways of working with the tool. Each way has a true quality, and every time a member uses one they read its payoff with noise: the reading is uniform around the true quality with a half-width drawn for each member, once, between 0.1 and 0.5, so some members read results far more accurately than others. Each member keeps a running estimate of the way they are on, weighting each new reading by 0.3.

In each round every member revises once (or, in the one-at-a-time cells of Table 2, one member at a time). A revising member hears eight others and adopts one of the ways they hold, with probability proportional to how highly each of them rates their own way (their running estimate), which is the weighted voter rule. In the hierarchical organization, which we call the chart organization, each of those eight is a teammate 60% of the time, the member's boss 15% (a member's boss is the team lead; a lead's boss is the executive), the executive 5% and anyone in the organization 20%; a lead's vote counts three times; and 15% of the time a member simply adopts their boss's way. In the flat organization everyone is heard equally and nobody defers.

Table 1. The two choices. Five ways of working each; the reading noise is set per member (Section 4).

Name	Qualities of the five ways	Gap
Close call (the main case)	0.60, 0.62, 0.64, 0.66, 0.68	0.02
Wider gap	0.50, 0.55, 0.60, 0.65, 0.70	0.05

The plans. Everyone at once: every member has the tool from round 0 and nobody checks. The ladder: every member has the tool from round 0, and a member about to switch first checks, with five tries of each way in rounds 0 to 99, two in rounds 100 to 199, one in rounds 200 to 299 and none from round 300 on. A check uses the member's own noise; a switch made by deferring to the boss is checked too. The people-first plan: the ten members with the least reading noise have the tool alone for the first 100 rounds, as described in Section 2, and at round 100 everyone else gets it, each starting from the way chosen by hearing eight of the ten. Both: the people-first plan, with everyone then checking five, two and one times until rounds 175, 250 and 325. Each run is read at round 400. A run's call is the way most members hold; a wrong call is any way but the best.

Variants. The ladder with a share of its checks skipped at random (half, 80%, 90%, 95%, 98%), with switches by deference exempt from the check, with one try for 300 rounds, with five tries for 300 rounds, with 20 tries in the first stage, and with 20% of members re-picking a way at random at each stage change. The people-first plan with four or twenty members in the first group, with one or twenty tries per check, with no exploration, with 25, 50 or 200 rounds alone, with the rest starting from one handed-over answer or at random, with the first group stopping its checks after rollout, with everyone else taught to check once, with the first group filling 10% of everyone's hearing, with no head start, with ten members chosen at random instead of the sharpest, and with the 25 leads (the executive among them) as the first group with no checks.

After a change. In a further set of runs the two best ways trade qualities at round 400, after every plan has finished rolling out, and the runs continue to round 700. A run "followed" if it was not on the new best way at round 400 and was on it 300 rounds later.

What was counted. Besides wrong calls we counted, for every run, the member-rounds (200 members for 400 rounds, 80,000 in all) spent waiting for the tool, on the best way and on a worse way, and the member-rounds on a worse way with no check in force for that member; the check trials performed, where one trial is one try of one way; every switch to a better or worse way, made with or without a

check, and every switch a check stopped; and the first round at which 90% of the whole organization held the best way.

Runs. Each configuration ran 400 times on seeds 0 to 399, with 95% Wilson intervals on shares of runs and normal intervals over runs on shares of member-rounds. The simulator's update is copied from the earlier note's organization model with switches added, and before any run the copy had to return exactly what the original returns on settings where both apply; the counting version had to return every earlier record exactly, and did for all 21 configurations it re-ran. In all, 34,400 distinct runs. Predictions for every part were written into the scripts before the runs; most of the numerical ones failed, all in the same direction, and Section 8 says which.

5 Results

5.1 Both plans fix the close call

Both plans cut wrong calls on the close call from 68.8% to under 3%, and the two could not be told apart.

On the close call in the chart organization, with everyone revising together, the organization settled on a worse way in 68.8% of runs when everyone got the tool at once and nobody checked (95% interval 64.0 to 73.1). The ladder left 2.8% (1.5 to 4.9), the people-first plan 1.8% (0.9 to 3.6) and both 1.0% (0.4 to 2.5). The ladder and the people-first plan are inside each other's intervals, and a second set of 400 runs of each, from the change runs read at round 400, gave 2.2% for both; pooled, 20 of 800 and 16 of 800 (Fisher exact test, two-sided, 0.61), a difference of 0.5 points (95% interval -1.0 to 2.0). The runs cannot tell the plans apart, which is not the same as showing them equal. The second sets gave different results, so they did not repeat the first sets' random draws, but their seeds are not given here. Figure 3 and Table 2 give every organization, gap and pace.

	Everyone at once, nobody checks	The ladder	Ten sharpest first	Both
Org chart, close call	68.8%	2.8%	1.8%	1.0%
Org chart, close call, one member at a time	72.5%	2.2%	4.8%	2.0%
Org chart, wider gap	50.5%	0	0.2%	0
Org chart, wider gap, one at a time	59.2%	0	0	0
No org chart, close call	3.5%	0	0	0
No org chart, close call, one at a time	18.8%	0	0	0
No org chart, wider gap	0	0	0	0
No org chart, wider gap, one at a time	0.8%	0	0	0

wrong calls: share of 400 runs that ended on a worse way; darker is worse

Figure 3. Wrong calls at round 400 under each plan, in every organization, gap and pace we ran. **Simulated; 400 runs per cell.**

Table 2. Wrong calls at round 400, percent of 400 runs.

Organization, gap, who revises	Everyone at once	Ladder	Ten sharpest first	Both
Chart, 0.02, everyone each round	68.8	2.8	1.8	1.0
Chart, 0.02, one at a time	72.5	2.2	4.8	2.0
Chart, 0.05, everyone each round	50.5	0	0.2	0
Chart, 0.05, one at a time	59.2	0	0	0
Flat, 0.02, everyone each round	3.5	0	0	0
Flat, 0.02, one at a time	18.8	0	0	0
Flat, 0.05, everyone each round	0	0	0	0
Flat, 0.05, one at a time	0.8	0	0	0

The hierarchy is what makes the uncontrolled rollout fail. With everyone at once, the organization ended on the way the executive happened to start on in 41.8% of runs, where chance would give 20%; under the ladder, 20.0%. Once people check, the chart stops deciding.

5.2 What each plan costs

The ladder pays in checking and people first pays in waiting.

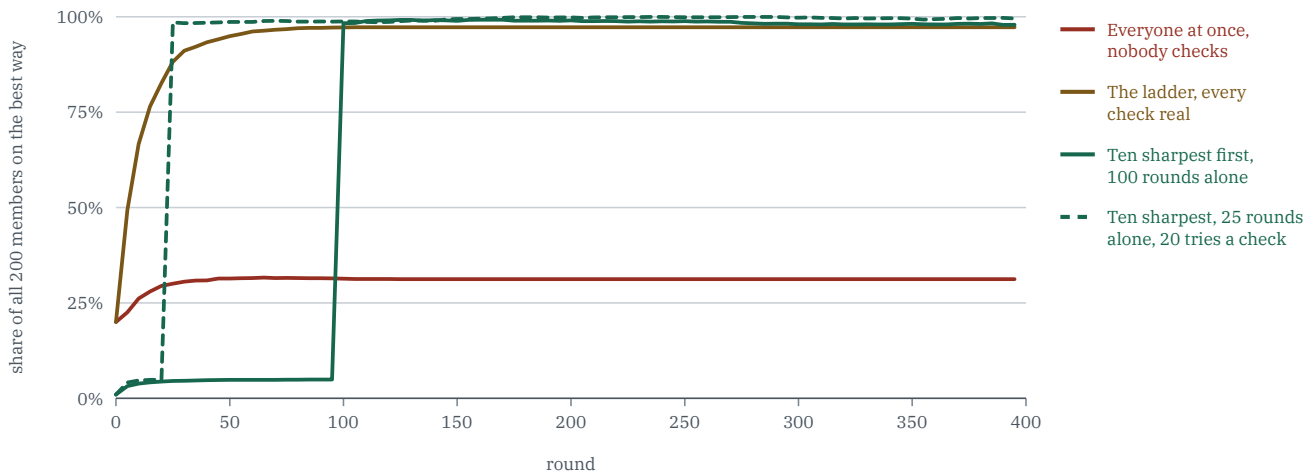


Figure 4. Share of all 200 members on the best way at each round, chart organization, close call, mean of 400 runs. The ladder converges within a few dozen rounds; the people-first plans jump when the first group hands over, at round 100 or round 25. **Simulated; seeds 0 to 399.**

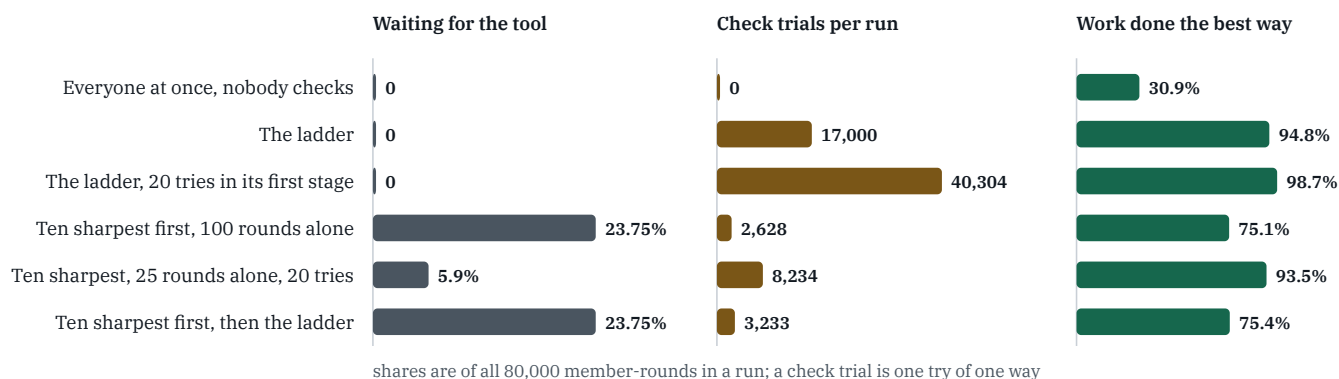


Figure 5. What each plan spends: time waiting for the tool, check trials, and the share of all work done the best way. Chart organization, close call. **Simulated; 400 runs per row.**

The two plans spend different things. The ladder has everyone on the tool from the first round and the organization converges quickly (Figure 4): 90% of all members held the best way by round 14 at the median (8 to 33 for the middle 80% of runs), and 94.8% of all member-rounds were worked the best way. The people-first plan made 190 members wait 100 rounds, so 23.75% of member-rounds were spent waiting and it could not pass 76.25% on the best way. The ladder paid in checking: 17,000 check trials per run, 85 per member, against 2,628 for the people-first plan, nearly all of them by the ten (Figure 5).

On the measure a ladder exists for, work done on a worse way with no check in force, the two are level. By the time the ladder's checks stop, the organization is already on the best way in all but the runs it lost in its first stage, so 0.69% of its member-rounds (0.29 to 1.09) were worked a worse way with no check in force. The people-first plan read 1.01% (0.47 to 1.55). Everyone at once read 69.1%.

Table 3. What each plan costs, chart organization, close call, everyone revising each round, 400 runs per row. Shares are of all 80,000 member-rounds per run.

Plan	Wrong calls	Waiting	On the best way	On a worse way	Of that, no check in force	Check trials per run	90% on the best way	Followed a later change
Everyone at once	68.8%	0	30.9%	69.1%	69.1%	0	33% of runs, round 19	0
Ladder	2.8%	0	94.8%	5.2%	0.69%	17,000	98% of runs, round 14	0
Ladder, 20 tries in stage one	0 of 400	0	98.7%	1.3%	0	40,304	every run, round 9	0
Ten sharpest first, 100 rounds alone	1.8%	23.75%	75.1%	1.2%	1.01%	2,628	every run, round 100	37.8%

Plan	Wrong calls	Waiting	On the best way	On a worse way	Of that, no check in force	Check trials per run	90% on the best way	Followed a later change
Ten sharpest, 25 rounds alone, 20 tries	0.2%	5.9%	93.5%	0.6%	0.5%	8,234	every run, round 25	78.0%
Both	1.0%	23.75%	75.4%	0.9%	0.22%	3,233	99% of runs, round 100	40.4%

The model does not price a waiting round, a round on a worse way or a check trial. With linear costs, which is our assumption and not a result, the counts give a break-even instead. If a round on a worse way loses L , a waiting round loses W (what the tool adds) and a check trial costs c , the people-first plan with 100 rounds alone beats the ladder only when c is above $1.32 W - 0.22 L$, and the 25-round plan with 20 tries does when c is above $0.54 W - 0.42 L$. Where checking is cheap, the ladder wins on these counts; where a careful check is expensive, the people-first plan does.

5.3 A check is only as good as the checker

The people-first plan could afford a sixth of the checking because its checkers were better: with the same five tries, ten random members let a worse way through 33.1% of the time, the ten sharpest 16.6%.

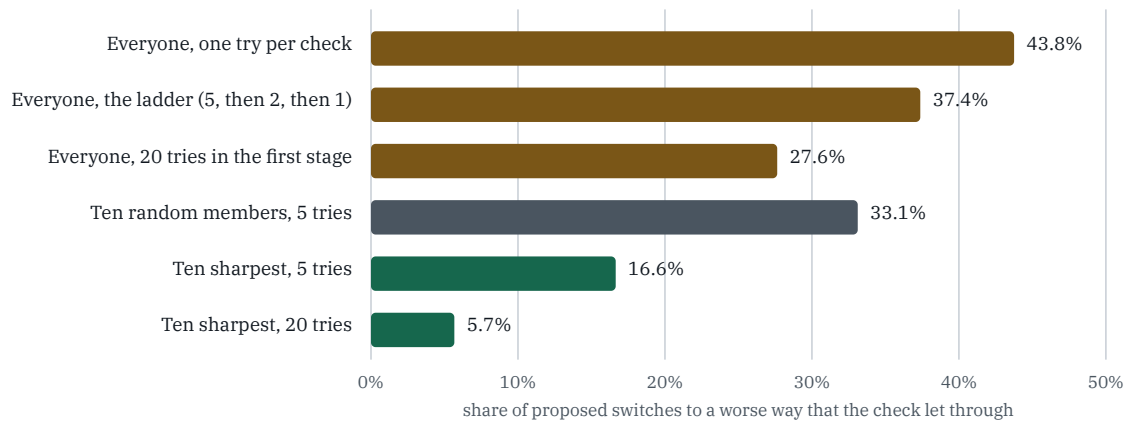


Figure 6. Share of proposed switches to a worse way that a check let through, chart organization, close call, 400 runs per bar. "Everyone" rows are ladders; the other rows are the first group of a people-first plan. The ladder row pools its five-, two- and one-try checks; at five tries each, the like-for-like comparison is ten random members against the ten sharpest. **Simulated; seeds 0 to 399.**

The people-first plan uses a sixth of the checking because only ten members check, and each of them checks more than a ladder member does (about 260 trials each against 85). Why it can afford so few checkers is in Figure 6. Across the ladder's checks, five tries of each way in its first stage and two or one later, each read with the member's own noise, 37.4% of proposed switches to a worse way got through.

With five tries each, ten members chosen at random let 33.1% through and the ten who read results most accurately 16.6%, about half as many, and 20 tries by the ten let 5.7% through. The ladder makes up for weak checks with volume: two hundred people checking, and copying amplifying the small bias each check adds. The cost of a weak check is not only letting worse ways through; the ladder's checks also stopped a switch to a better way 38.6% of the time, the ten's 25.3%.

5.4 What the people-first plan depends on

Who goes first matters: ten random people with the same checks left seven times as many wrong calls as the ten sharpest.

Who went first moved the result about as much as any other part of the plan. With the same checks and exploration, the ten sharpest left 1.8% wrong and ten members chosen at random 12.8%, seven times as many. That gap assumes the ten are picked on their true reading noise and that the sharpest read about three times as accurately as the average member (Section 8). The 25 leads going first with no checks left 57.8%, close to everyone at once; that row changes two things, the group and the checking, so it shows how far a leaders-first rollout without checks can go wrong rather than the effect of who goes first alone. The mechanism is a founder effect: the organization ended on the first group's own choice in 97.8% of runs under the default plan and 99.5% with the leads first, and where the leads' choice was wrong, in 58% of runs, the organization stayed wrong in 99.6% of those. The plan also needs the head start (16.8% wrong with none), the exploration (16.5% without), real checks (11.5% with one try instead of five), enough people (9.2% with a group of four, 1.8% with twenty) and the rest starting from what the group found (14.5% if they start at random instead). Handing over one answer instead of letting newcomers hear the group gave 2.2%.

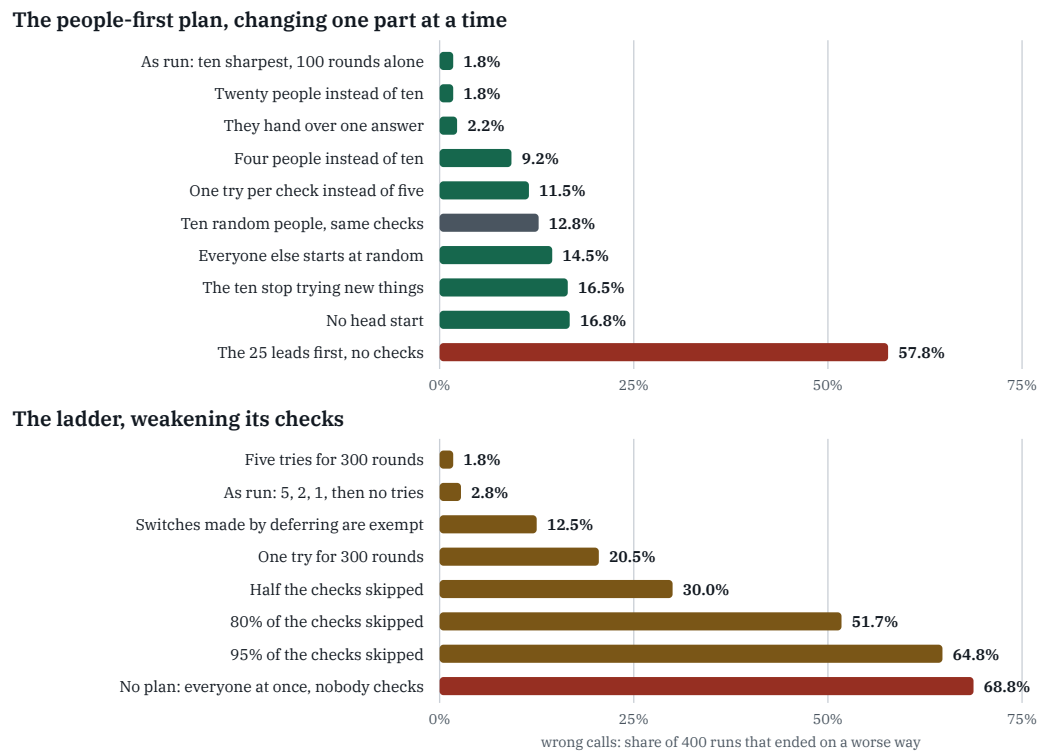


Figure 7. What each plan depends on: wrong calls at round 400 when one part of the plan changes. Chart organization, close call. **Simulated; 400 runs per bar.**

Table 4. What each plan depends on. Wrong calls at round 400, chart organization, close call, everyone revising each round, 400 runs per row.

The ladder	Wrong calls	The people-first plan	Wrong calls
Every check real (5, 2, 1, 0 tries)	2.8%	Ten sharpest, 100 rounds alone	1.8%
Five tries for 300 rounds	1.8%	Ten sharpest, 20 tries a check	0.2%
One try for 300 rounds	20.5%	Ten sharpest, one try a check	11.5%
Switches by deference exempt	12.5%	Ten sharpest, no exploration	16.5%
Half the checks skipped	30.0%	Ten random members	12.8%
80% skipped	51.7%	The 25 leads, no checks	57.8%
95% skipped	64.8%	The same ten, no head start	16.8%

5.5 What the ladder depends on

The ladder is only as good as its checks: with half of them skipped it was wrong 30.0% of the time (Figure 7, lower panel).

The ladder is only as good as its checks. With half of them skipped at random it was wrong 30.0% of the time, with 80% skipped 51.7%, and from 95% skipped up it was no better than no plan. Skipped checks also show up on the ladder's own measure: with half skipped, 7.5% of member-rounds were worked a worse way with no check in force, ten times the honest ladder's, and each run let 491 switches to a worse way through unchecked. Exempting switches made by deferring to the boss left 12.5%. The ladder's later stages were never put to the test: its wrong calls were 10.0% at round 25, 4.8% at round 50 and 2.8% at round 100, and then the same 11 runs at every later reading. Each stage lasts 100 rounds and 90% of members held the best way by round 14 at the median, so the call is made in the first stage, before the checking steps down; shorter stages were not run. How heavily the first stage checks is what matters: one try for 300 rounds left 20.5%, five tries for 300 rounds 1.8%, and twenty tries in the first stage alone 0 of 400.

5.6 The head start can be short if the checking is hard

The wait is a setting: 25 rounds alone with thorough checks did as well as 100 rounds alone, with a quarter of the waiting, at about three times the checking.

The people-first plan's cost is the wait, and the wait is a setting. Figure 8 shows the trade. With no head start the ten sharpest left 16.8% wrong; with 25 rounds alone 5.8%, 50 rounds 4.0%, 100 rounds 1.8% and 200 rounds 1.2%, while the share of member-rounds spent waiting rose from 0 to 47.5%. Harder checking by the same ten moved the curve: 25 rounds alone with 20 tries per check left 1 wrong run in 400, and 4 in a second set of 400 (0.6% pooled), with 5.9% of member-rounds waiting, 93.5% worked the best way against the ladder's 94.8%, and 8,234 check trials against the ladder's 17,000 and the 100-round plan's 2,628. A first group of twenty instead of ten, with 25 rounds alone, left 2.5%; ten with 25 rounds alone and then everyone checking once, 2.0%, with nothing worked a worse way unchecked.



Figure 8. Wrong calls against the share of member-rounds spent waiting, chart organization, close call, 400 runs per point. The green curve is the ten sharpest by length of head start; the separate points change how they check or what everyone else does afterward. **Simulated; seeds 0 to 399.**

5.7 After the rollout: following a change, and a slow leak

A finished ladder could not follow a change. People who kept checking and trying things could.

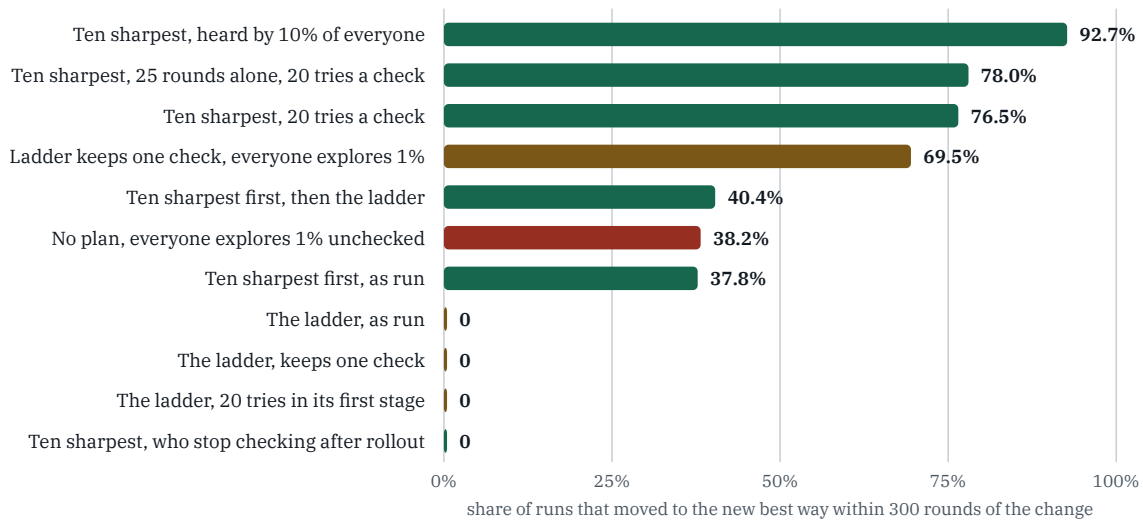


Figure 9. After the two best ways traded places at round 400: the share of runs that moved to the new best way within 300 rounds. Chart organization, close call. The ladder rows have no exploration by construction; the people-first rows keep the ten exploring. **Simulated; 400 runs per bar.**

When the two best ways traded qualities at round 400 (Figure 9), the ladder followed in 0 runs, a ladder that kept one check followed in 0 as well, and so did the ladder with 20 tries in its first stage (0 of 400), because a check only tests what somebody proposes and in a settled organization nobody proposes anything. The ladder as we built it has no exploration in any stage, so this failure is built into it rather than caused by its last stage removing the checks, as the ladder that kept one check shows. The people-first plan followed in 37.8% of runs within 300 rounds, 76.5% when the ten checked 20 times, 78.0% with the 25-round, 20-try plan, and 92.7% when the ten filled 10% of everyone's hearing; if the ten stopped checking and exploring after rollout, 0. Exploration is what does it, and it works best when it is checked: a ladder that kept one check and had everyone try a random way 1% of the time followed in 69.5% of runs and knocked none of the runs already on the best way off it, while everyone at once with 1% unchecked exploration followed in 38.2% and knocked half of them off.

The same exploration costs something when nothing changes. Under the default people-first plan the ten's own choice was wrong in 2 runs at rollout and the organization's in 7 runs at round 400, all seven of them runs where the ten had been right. The ten keep trying things, their checks let a worse way through about one time in six, and the hierarchy sometimes carries one of those choices into the organization. If the ten stop exploring after rollout, the count holds at 4 after round 200; if everyone else checks once, it holds at 2; with ten random members it rose from 24 to 51. With 20 tries per check the direction reversed: 5 wrong runs at rollout and 1 at round 400.

5.8 Where voices are equal there is little to fix

In the flat organization, with equal voices and no deference, everyone at once left 3.5% wrong on the close call when everyone revised each round, and every plan left none (Table 2). The organization still took longer to get there without checks: 90% of members held the best way at a median round of 70, against 13 under the ladder. This is the condition under which copying alone works well, because what people show is filtered by how well it pays them (Rendell et al., 2010). The hierarchy is what breaks the filter.

6 When the tool builds

6.1 A second model

The model in Sections 4 and 5 lets an organization choose among five ways of working that already exist. Nothing is built in it, so it cannot test the stronger form of the position: that a few strong people using AI to build for the organization change it more than everyone using AI in their own work. We added building to the same chart organization (Figure 10). Everyone starts on an incumbent way of working with true quality 0.50. In each generation each member who holds the tool builds with probability 0.01; a build is a new way of working whose true quality is the builder's current one plus a draw from a normal distribution with mean 0 and standard deviation 0.02, clipped to the range 0 to 1, so a build is as likely to be worse as better, whoever makes it. The builder tries the new way and their current one a set number of times each, with their own reading noise, and keeps the new one only if it read at least as well; a kept build can then spread like any other way. Every member tries a candidate and their own way once before switching, and a member who holds the tool checks a switch with as many tries as they build with. The only difference between the ten sharpest and anyone else is the one in Section 4: they read results with the least noise. Nobody is assumed to build better things.

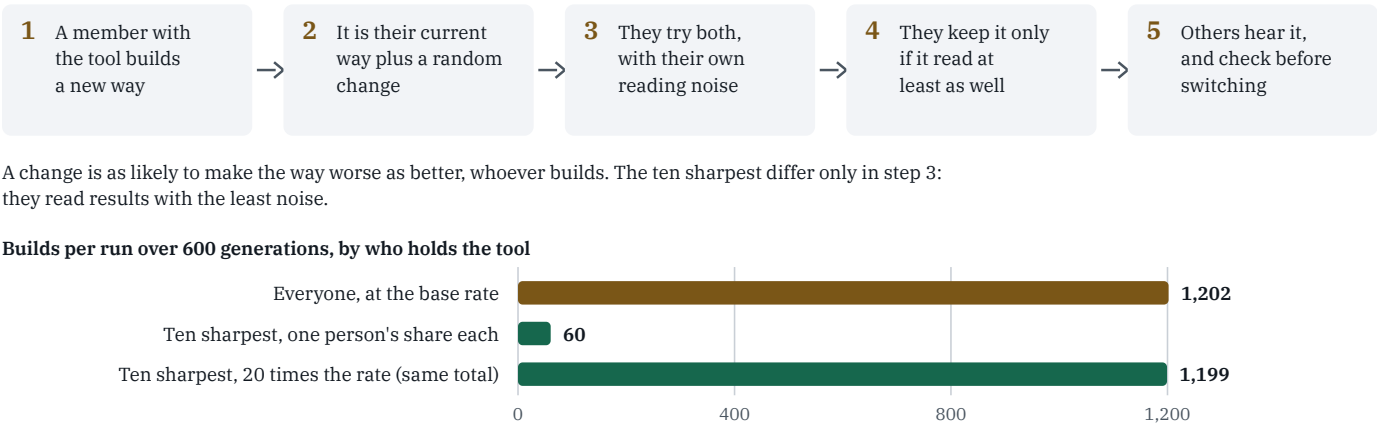


Figure 10. How a build happens in the second model, and how much building each plan does. A diagram of the model; the builds per run are simulated, 400 runs per bar.

The revision is the chart organization's, with one change of method: a member draws a heard member with probability proportional to that member's report times its weight, which gives the same distribution as the weighted vote over options in Section 4 and needs no fixed list of options. Before any run, the new code had to reproduce two configurations of Section 5 with building switched off, everyone at once on the close call (267 wrong calls in 400 runs, against the recorded 275) and the ladder with one check for 300 rounds (75 against 82), each inside the other's 95% interval; it did, and did again before each later set of runs. An exact match, as in Section 4, could not be expected after that change of method, so the gate is statistical, and weak: differences of about four points would also have passed it.

Plans. Everyone builds: all 200 members hold the tool and build at the base rate, trying each build once, or carefully, five times each. The ten sharpest: the ten members with the least reading noise build and check five times, either at the base rate per person, which gives them a twentieth of the organization's builds, or at 20 times the base rate, which gives them the same total building as everyone together. Ten members chosen at random, and the 25 leads (the executive among them), with the same total building and the same checks. Both: the ten sharpest at 20 times the rate, with everyone else at the base rate trying once. Heard by all: the ten sharpest with the same total building and 10% of every member's hearing given to them, as in Section 5.7; each member still checks before switching. Told: every 25 or every 100 generations everyone, the ten included, is switched without a check to the way the ten rate highest. Told with a check: at each order every member first tries the ordered way and their own, once each (a builder five times), and switches only if the ordered way read at least as well. Told with the ten exempt: the order moves everyone but the ten. **Variants.** The flat organization; the tool as a leveler, with a holder's reading noise moved halfway to the best member's; builders who read their own builds 0.02 better than they are; rare big wins (2% of builds add 0.10); no check before adopting; most builds worse (a mean change of -0.01); and bigger changes (standard deviation 0.05).

Readings. The gain is the organization's mean true quality in use minus 0.50. A way of working cannot pass quality 1.0, so by generation 600 most plans meet at the ceiling, a gain of 0.50, which we did not foresee when we chose 600; we read the plans at generation 200, where they are still apart, and give 100 and 600 as well. Each configuration ran 400 times on seeds 0 to 399, and the 95% interval on every mean gain at generation 200 is within ± 0.011 of it. A run is harmed if its gain is below zero. In all, 40 configurations and 16,000 runs. Predictions were written into the script before any run.

6.2 The few do more only with the capacity

With the same total building, the ten sharpest did 1.7 times what everyone did building carefully. With one person's share each and nobody else building, they made a twentieth of the builds and did little.

With the same total building, the ten sharpest raised the organization's quality by 0.279 by generation 200, against 0.162 with everyone building carefully, 0.115 with ten random members and 0.056 with everyone trying each build once: 1.7, 2.4 and 5.0 times as much (Table 5, Figure 11); the last comparison also gives the ten five tries a check against one. At the base rate per person they made about 60 builds a run over 600 generations against everyone's 1,200, and raised quality by 0.011, a fifth

of what everyone trying once achieved. That arm is a budget control, not a test of the few, since nobody else builds in it; the test of the few is at equal budget and care, 0.279 against 0.162. The few did more only when they had the capacity. Sharpness and care worked together. Care alone, everyone checking five times, closed much of the gap, and ten random members checking as carefully as the ten sharpest did less than half as well. The 25 leads with the same total building reached 0.223, ahead of the random ten although they do not read results better; our reading is that they are heard more and count three times (we did not run the leads without those advantages, and they are 25 people, not ten). Adding everyone at the base rate to the ten sharpest reached 0.317. Checking kept harm rare: no plan in Table 5 left more than 2.2% of runs below where they started.

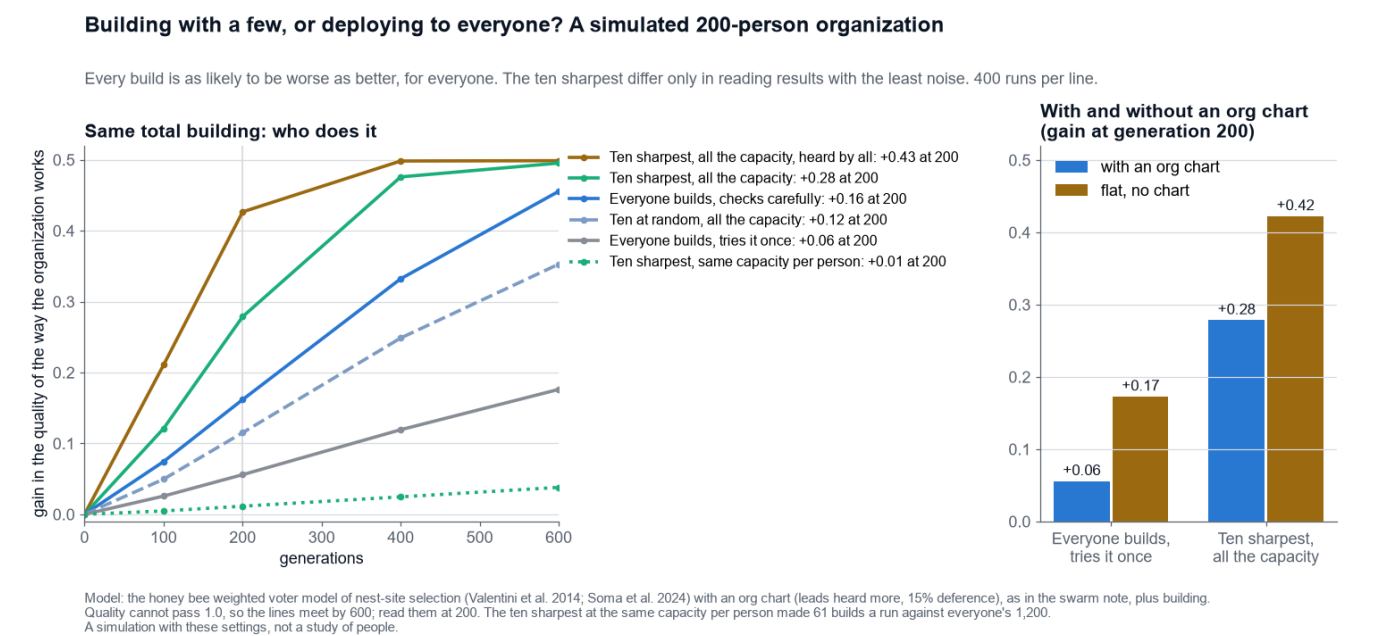


Figure 11. The building model, chart organization: gain in the quality of the way the organization works, mean of 400 runs per line (left), and the gain at generation 200 with and without the org chart (right). The ten sharpest differ from everyone else only in reading results with the least noise. **Simulated; seeds 0 to 399.**

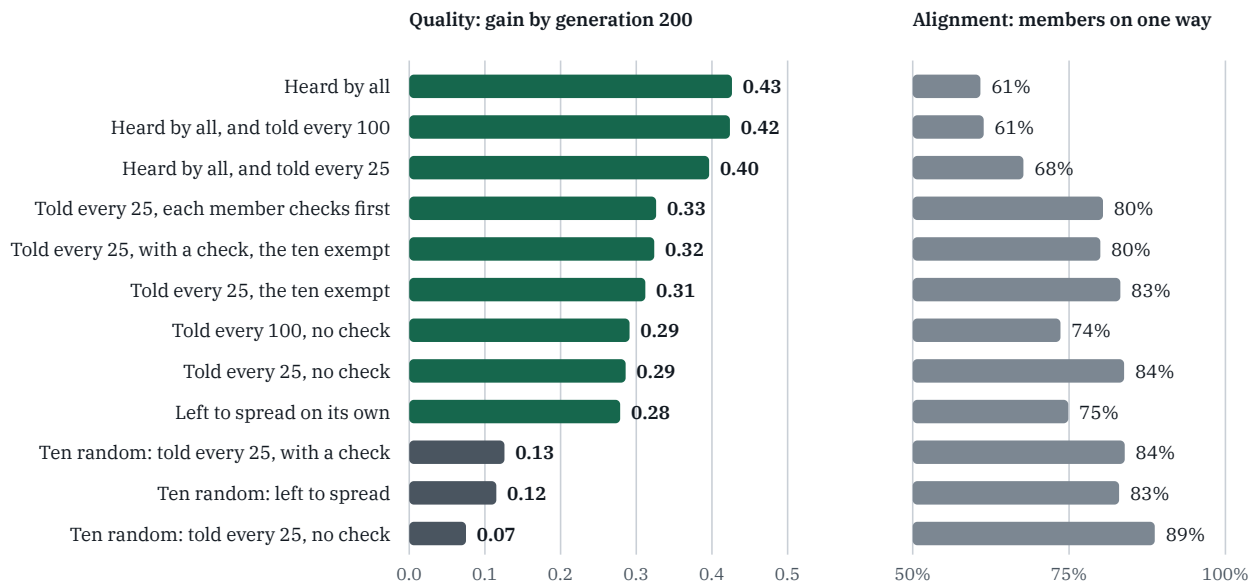
Table 5. The building model, chart organization, base settings. Gain in the organization's mean quality, mean of 400 runs per row.

Plan	Builds per run (over 600 generations)	Gain by generation 100	By 200	10th and 90th percentile at 200	By 600
No tool	0	0	0	0 and 0	0
Everyone builds, tries each once	1,202	0.026	0.056	0.007 and 0.108	0.176
Everyone builds, carefully	1,198	0.074	0.162	0.091 and 0.240	0.455

Plan	Builds per run (over 600 generations)	Gain by generation 100	By 200	10th and 90th percentile at 200	By 600
Ten sharpest, base rate per person	60.5	0.005	0.011	0.000 and 0.045	0.038
Ten sharpest, same total building	1,199	0.121	0.279	0.171 and 0.393	0.495
Ten random members, same total building	1,200	0.050	0.115	0.027 and 0.219	0.353
The 25 leads, same total building	1,200	0.104	0.223	0.137 and 0.312	0.492
Both	2,340	0.138	0.317	0.196 and 0.431	0.496
Ten sharpest, heard by all	1,200	0.211	0.426	0.326 and 0.497	0.499

6.3 Heard or told

An audience did more than an order, and an order worked better when members could check it first.



Ten random, told with no check: 8.2% of runs ended worse than they started; with a check, 0.8%. No other row was harmed.

Figure 12. Heard or told: how the builders' work reaches everyone else, by the quality it produced and by how aligned it left the organization, at generation 200. Green rows are the ten sharpest building; gray rows are ten random members. "Heard by all" is the arm in which the ten make up 10% of everyone's hearing. The alignment axis starts at 50%, which makes its differences look larger than they are. **Simulated; 400 runs per row.**

Giving the ten an audience did more than anything else we ran: with the ten making up 10% of everyone's hearing, 0.426 by generation 200. In the model an audience is a permanent increase in the ten's influence. It is not an order. Each member still tried what they heard against their own way before switching. An order with no check did no better than letting the ten's work spread on its own: told every 25 generations, 0.286, and every 100, 0.291, against 0.279 heard as usual. Telling everyone to use what ten random members built did worse than letting it spread, 0.075 against 0.115, and left 8.2% of runs below where they started (Table 6, Figure 12). Telling bought alignment instead: 83.8% of members held the most common way when told every 25 generations, against 74.9% heard as usual and 60.8% heard by all. Adding an order to the audience traded one for the other, 0.424 with 61.4% on one way when told every 100 generations and 0.396 with 67.7% when told every 25; neither reached the order's alignment or the audience's quality.

Two further sets of runs account for part of why the order did worse than the audience. It skips the check that makes copying safe: when every member first tried the ordered way against their own and switched only if it read at least as well, the order reached 0.326, with 80.4% of members on one way, and with ten random builders the same check cut the runs harmed from 8.2% to 0.8% and reached 0.126, above letting their work spread. And it moves the ten themselves onto a single line of work each time: with the ten exempt from their own order, it reached 0.312. The two together gave 0.324, no more than the check alone. An order that members could check did better than letting the work spread and kept most of the order's alignment, but it stayed well short of the audience on quality. Together these runs close about 0.04 of the 0.14 gap between the order and the audience (0.286 to 0.326, against 0.426). Conjecture: much of the rest is variety; the audience left the organization more varied, 60.8% on one way against 74.9% heard as usual, keeping several lines of improvement alive (we did not measure how many ways stayed in use).

Table 6. Heard or told, the building model, chart organization, at generation 200, 400 runs per row.

Plan	Gain	Runs harmed	Members on the most common way
Ten sharpest, heard as usual	0.279	0	74.9%
Ten sharpest, heard by all	0.426	0	60.8%
Ten sharpest, everyone told every 25 generations	0.286	0	83.8%
Ten sharpest, everyone told every 100 generations	0.291	0	73.6%
Ten sharpest, told every 25, with a check	0.326	0	80.4%
Ten sharpest, told every 25, the ten exempt	0.312	0	83.2%
Ten sharpest, told every 25, with a check, the ten exempt	0.324	0	80.0%
Heard by all and told every 100	0.424	0	61.4%
Heard by all and told every 25	0.396	0	67.7%
Ten random members, heard as usual	0.115	0	83.0%
Ten random members, everyone told every 25	0.075	8.2%	88.7%
Ten random members, told every 25, with a check	0.126	0.8%	83.9%

6.4 The org chart holds back everyone building

Without the org chart, everyone building did three times better, and the ten sharpest stayed ahead under every setting we tried.

In the flat organization, with equal voices and no deference, everyone trying each build once reached 0.173 by generation 200, 3.1 times its 0.056 in the chart, while the ten sharpest with the same building reached 0.423, 1.5 times their 0.279. Deference and the leads' extra weight pull members toward the ways their leads and the executive hold, and when everyone builds, those are rarely the best builds. The flat organization paid in fragmentation instead: 57.4% of members on the most common way when everyone built and 46.9% under the ten, against 83.3% and 74.9% in the chart. The other variants moved the levels without changing the order (Figure 13). Two readings are not in the figure: with rare big wins the ten at the base rate reached only 0.042, and with most builds worse everyone trying once was harmed in 19.8% of runs. With bigger changes every plan was near the ceiling by generation 200; at 100, everyone trying once had gained 0.226, everyone careful 0.418, the ten with the same building 0.482 and the ten at the base rate 0.067.

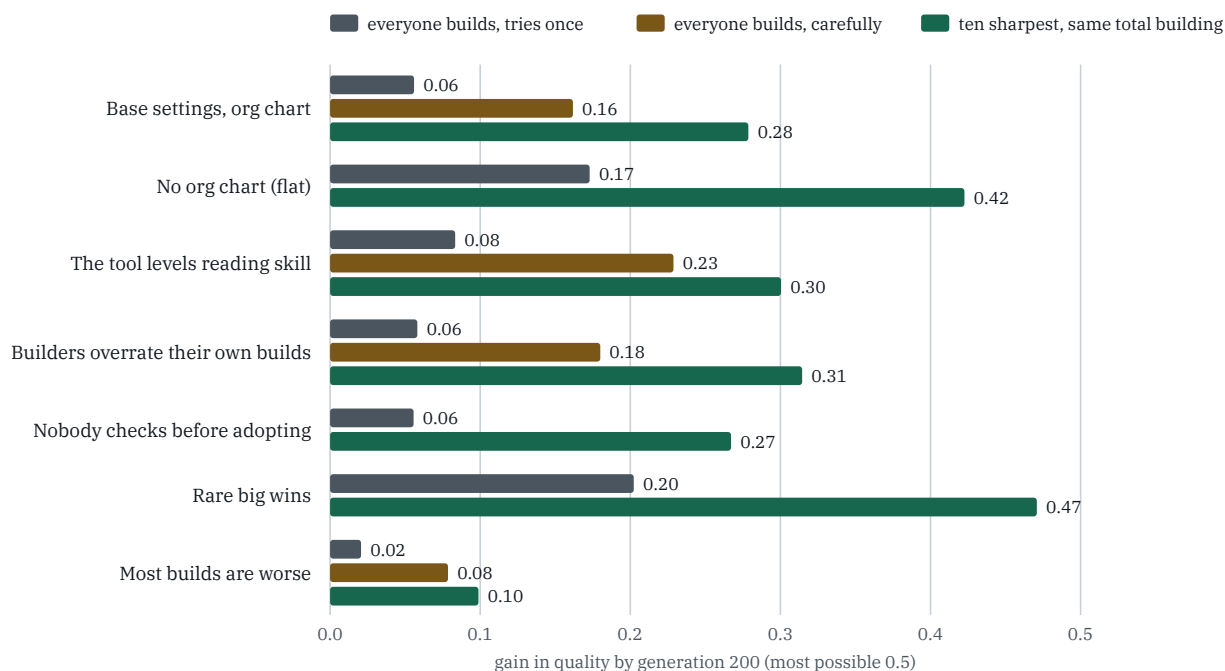


Figure 13. The same comparison under other settings: gain by generation 200 for everyone trying each build once, everyone building carefully (where run) and the ten sharpest with the same total building. **Simulated; 400 runs per bar.**

6.5 On a rugged problem, who builds matters little

On a rugged problem every plan climbed to a nearby peak and stopped, and who built made little difference. Bigger changes helped more.

The building model's landscape is smooth: a build adds an amount to its builder's way that does not depend on anything else. That flatters concentrating the building, because the work cited in Section 3 finds that on complex problems, processes in which people copy the most successful can settle early on local peaks. So we ran the same organization and plans on an NK landscape (Kauffman, 1993; Levinthal, 1997). A way of working is a string of 12 bits; each bit's contribution depends on itself and the next K bits, drawn at random; and quality is the mean contribution divided by the best possible in that landscape, so the best way has quality 1. With $K = 0$ the landscape is smooth with one peak, and with $K = 5$ it is rugged with many. A build flips one bit of the builder's current way, or three in the bigger-changes runs, and is checked and kept as before. Every plan saw the same landscape and the same starting way in each run, so plans are compared run by run, and the gain is measured from the starting way's quality. The replica gate passed again (271 and 90 wrong calls in 400 runs, against 275 and 82). In all, 18 configurations and 7,200 runs.

The same test on a rugged problem (many local peaks)

Every plan sees the same landscapes and starting points; 400 runs per line. The organization ends on the best possible way in only 3.0 to 8.8% of runs, by plan.

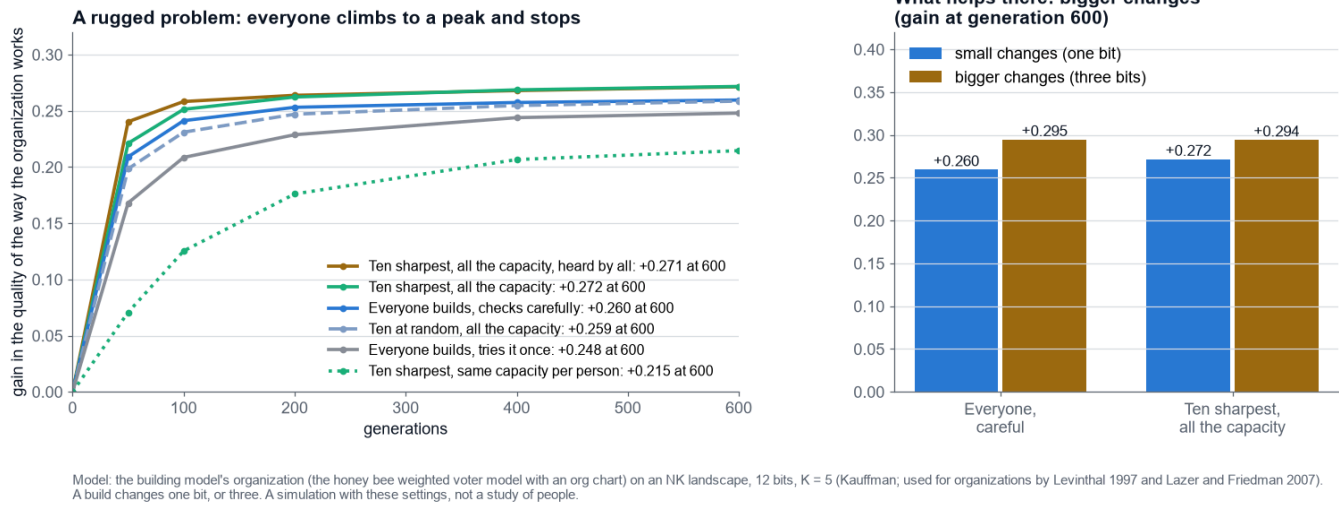


Figure 14. The same plans on a rugged NK landscape (12 bits, $K = 5$): gain from the starting way, mean of 400 runs per line (left), and the gain at generation 600 with one-bit and three-bit builds (right). **Simulated; seeds 0 to 399.**

On the rugged landscape every plan climbed quickly to a peak and stopped (Figure 14, Table 7). With single-bit builds the organization's most common way was the best possible way in only 3.0% to 8.8% of runs, by plan. Who built barely mattered there. By generation 600 the ten sharpest with all the capacity had gained 0.272, everyone building carefully 0.260, ten random members 0.259 and everyone trying once 0.248, and the plans had closed, on average, between 57% and 74% of the distance from the starting way to the best possible one. The ten sharpest's early lead did not shrink over time, contrary to what the objection predicts, but it was small from the start. What helped was the size of the change: with three-bit builds everyone careful gained 0.295 and the ten sharpest 0.294, and the share of runs ending on a local peak that was not the best fell from 81.0% and 79.8% to 49.5% and 55.5%. On the smooth landscape the order of Section 6.2 appeared again with smaller gaps, because twelve bits leave a short climb: by generation 200 the ten sharpest had gained 0.234, everyone careful 0.231, ten random members 0.212, everyone trying once 0.182 and the ten at the base rate 0.111.

Table 7. The same plans on an NK landscape of 12 bits, chart organization. Gain from the starting way, mean of 400 runs per row; the last two K = 5 columns are shares of runs at generation 600.

Plan	K = 5, by generation 50	By 100	By 600	Most common way is the best	On another local peak	K = 0, by generation 200
Everyone builds, tries each once	0.168	0.209	0.248	4.5%	70.5%	0.182
Everyone builds, carefully	0.209	0.241	0.260	6.0%	81.0%	0.231
Ten sharpest, base rate per person	0.071	0.125	0.215	3.0%	44.0%	0.111
Ten sharpest, same total building	0.221	0.251	0.272	8.8%	79.8%	0.234
Ten random members, same total building	0.198	0.231	0.259	4.8%	72.5%	0.212
Ten sharpest, heard by all	0.240	0.258	0.271	7.5%	86.8%	
Everyone careful, three-bit builds	0.219	0.249	0.295	11.0%	49.5%	
Ten sharpest, three-bit builds	0.225	0.258	0.294	8.5%	55.5%	

7 What it means

7.1 For the position

The position held up, in a narrower form than it is usually stated. Who goes first matters as much as what the tool is allowed to do, not more. The model has no tool permissions, so what it tests is staged checking by everyone against concentrated checking by a few. Both move a close call from wrong about two times in three to a few percent, and both fail the same way when the people involved do not really check. What the people-first plan offers is efficiency and adaptability. It reaches the ladder's accuracy with a fraction of the checking, because it concentrates the checking in the people who are good at it, and its first group, which keeps checking and trying things, keeps the organization able to follow a change that the finished ladder cannot. What it costs is time, and the time can be short if the first group checks hard, though that triples its checking.

The two plans also combine naturally. Anthropic's program stages its people the way a ladder stages a tool: its engineers train on a simulated deployment, are graded before they lead, then lead a real project for 12 weeks and are graded again (Anthropic, 2026). It is a ladder for the people, and it says nothing

about staging the tool. In the interview the author described both halves, starting with the A-team and withholding the keys until trust is earned (York, 2026). Our best cells are the same combination: a short head start for the sharpest people with thorough checking, then a light check for everyone, or the ten followed by the ladder.

The building model sharpens the position and narrows it. Giving the A-team a team is a claim about capacity. With the organization's building and the same care in checking, the ten sharpest did 1.7 times as much as everyone (0.279 against 0.162); given one member's share each and nobody else building, they made a twentieth of the builds and did little. Our position: that separates the two kinds of program in Section 1. A seat for everyone, measured by use, gives the few one person's share of the capacity, where they barely moved the organization; the builder programs give a nominated few the time and support to build. The few did best with an audience rather than an order, which is the interview's other instruction, to let the rest of the organization see the results; where there was an order, it worked better when each member could check it first. On a rugged problem the plan mattered little and the size of the change mattered more, so concentrating the building is not a general answer; it pays where improvement comes in steps that build on each other.

7.2 For anyone rolling out a tool

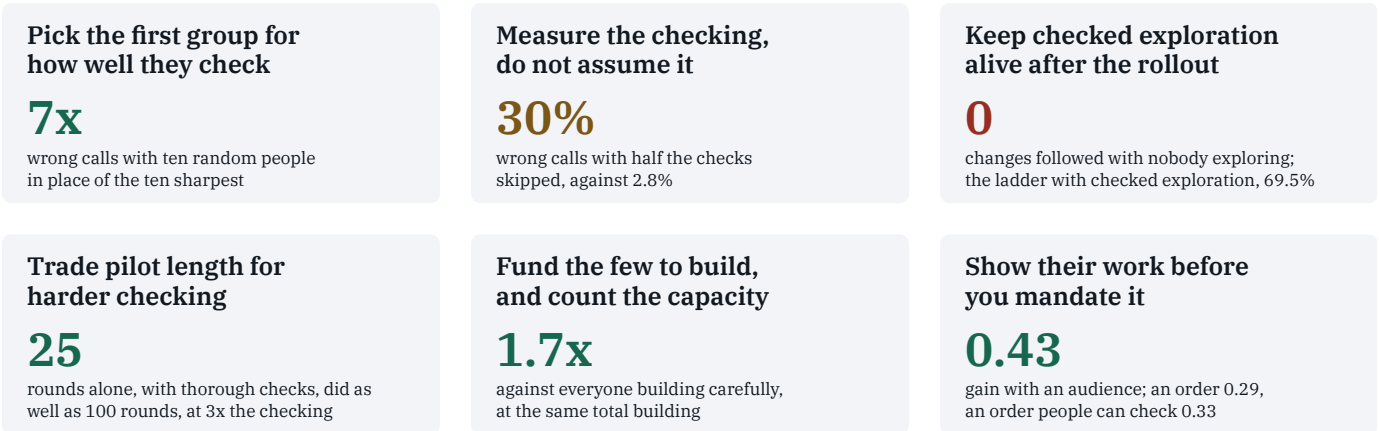


Figure 15. Six readings for anyone rolling out a tool, each with the number behind it. "Audience" and "order" are model arms, the ten as a fixed share of everyone's hearing and a periodic switch onto their best way, read here by analogy. **Simulated; hypotheses for real organizations, not findings about them.**

Four practical readings follow, each as a hypothesis for real organizations rather than a finding about them. Choose the first group for how well its members check, not for rank: in the model ten random members in place of the ten sharpest left seven times as many wrong calls. The model picked the ten on their true noise and ran the leads first only without checks, so the rank half is our position, not a result. Measure the checking rather than assume it: a ladder whose checks become approvals is the ladder with half or most of its checks skipped, and automation bias makes that likely even for experts (Parasuraman & Manzey, 2010). In the interview's terms, the question is whether the human in the loop is checking or watching. Keep some checked exploration alive after the rollout, by the first group or by everyone: in the model no plan without it followed a change, and the ladder with it followed in 69.5% of

runs. And treat the length of a pilot as a setting to be traded against how hard the pilot group checks, not as a fixed phase.

Two more come from the building model, with the same standing (Figure 15 shows all six). Fund the few to build rather than giving everyone a seat and counting use, and count the capacity honestly: the model assumes one member can build twenty times as often as an average member, and a real program has to make that true. And make their work visible before making it mandatory: an order nobody could check did no better than letting the work spread, and with builders chosen at random it did harm, while an order that each member could check first kept most of the alignment and removed nearly all of the harm.

7.3 A hypothesis for teams of AI agents

The same structure may describe a team of AI agents that reuse each other's work, as our earlier note suggested (Atkinson, 2026a). There, the analogue of the first group would be the agents or the operator whose checks are most reliable, and the analogue of the ladder would be a review rule applied to every agent. If the model carries over, concentrating verification in the most reliable checkers would buy most of the accuracy for a fraction of the review, and a team that stops checking once it has settled would stop noticing change. We did not test any of this on agents.

Every check in this model is external: a member tries both ways themselves and cannot change the reading without changing the way of working. The background paper *Never let your AI grade its own homework* argues that an instrument the measured party operates will rot; this model adds that, given an external check, how well the checker tells better from worse decides what the check is worth.

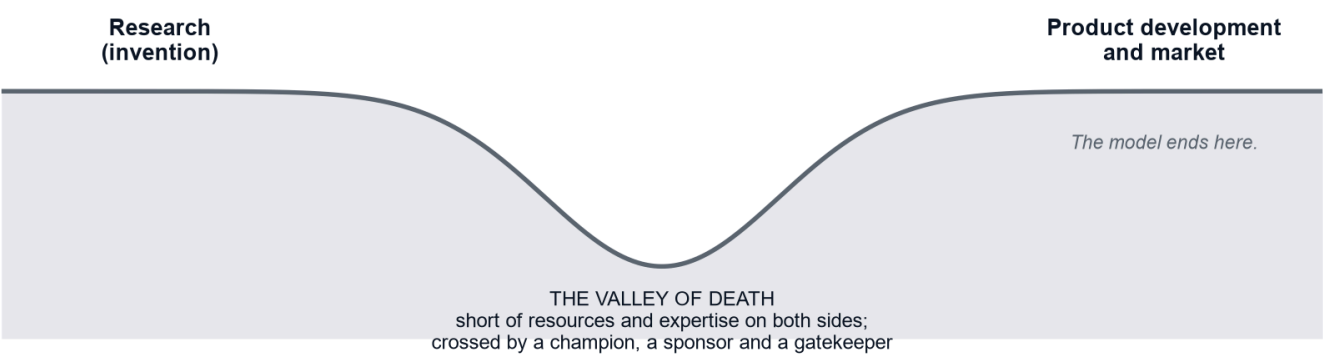
7.4 Three valleys instead of one

The valley of death is a picture of the move from invention to a commercial product. Markham, Ward, Aiman-Smith and Kingon (2010) use it for the front end of product innovation, the stretch between research and product development where resources and expertise are relatively scarce, and they find three roles that carry a project across: a champion who conceptualizes the idea and prepares the business case, a sponsor who provides resources to demonstrate that it is viable, and a gatekeeper who sets criteria and decides. Auerswald and Branscomb (2003) describe the same stretch as a financing gap that comes from asymmetries of information and motivation and from institutional gaps, but they question the image itself: the stretch is not barren, they argue, and they propose a Darwinian sea, in which many business and technical ideas contend, in its place. In their study for NIST a workshop participant defined the valley inside a company as the point where the bill adds up and management asks what it will get for it, while the project is still too early to say (Branscomb & Auerswald, 2002). Nor is one gap the only picture in use: clinical research has long described two translational blocks, from laboratory to clinical study and from clinical study to practice (Sung et al., 2003).

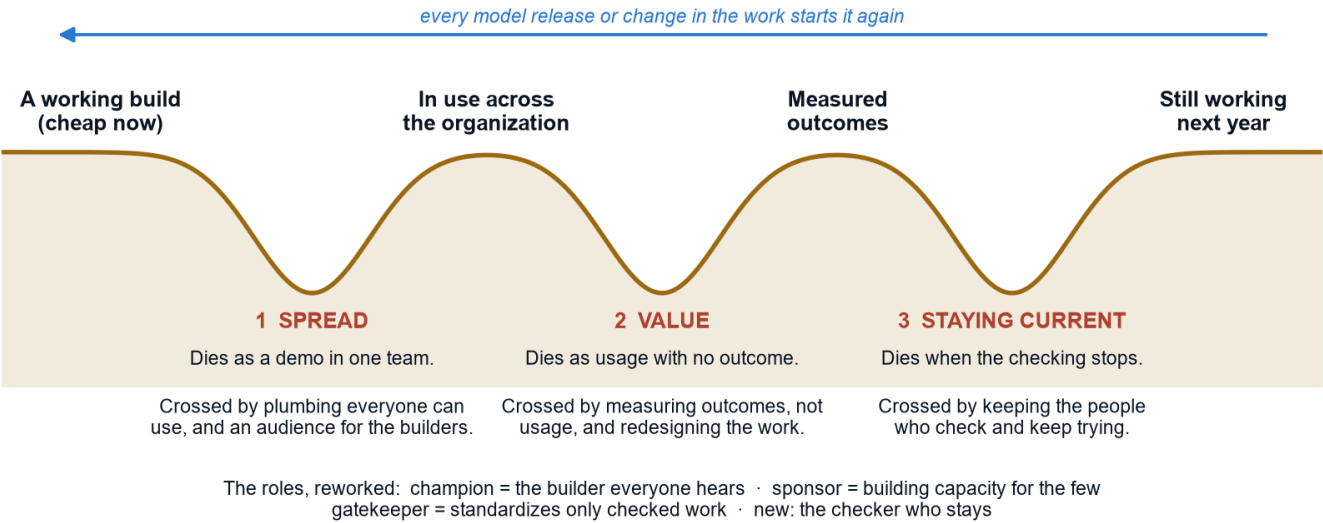
The author's proposal adapts that picture to the adoption of an AI tool inside one organization: there are three valleys, and the last one repeats (Figure 16). A working build is cheap now, so the first valley is no longer between research and a product. It is spread: something that works stays a demonstration in one

team. The second is value: the tool is in use across the organization and no gain in output can be shown, as in the United Kingdom's evaluation of its Copilot pilot, which found time saved but no robust evidence of improved productivity (Department for Business and Trade, 2025). The third is staying current: it works, people stop trying alternatives and checking them, and when the tool or the work changes the organization cannot follow. The single-valley picture ends when the product ships. This one does not end, because every model release or change in the work starts it again.

The standard model: one valley, between research and product



AI adoption in an organization: three valleys, and the last one loops



In this paper: valley 1, Section 6.3 (the builders heard by everyone gained 0.43, against 0.28 when their work was left to spread).
 Valley 2, the UK trade department's Copilot evaluation in Section 1 (time saved, no productivity gain shown).
 Valley 3, Section 5.7 (after its last stage the ladder followed a change in 0 runs).
 Sections 5 and 6 are simulations, not studies of people. The three valleys are a proposal.

Figure 16. The single valley of death as Markham et al. (2010) use it, an image Auerswald and Branscomb (2003) questioned, and the three valleys proposed here for AI adoption inside an organization. **A proposal; its last line says where this paper bears on it.**

Our results sit in the first and third valleys. On spread, who went first mattered (Section 5.4), and giving the builders' work an audience raised the organization's gain to 0.426, against 0.279 when it was left to

spread and 0.286 when everyone was ordered onto it (Section 6.3). On staying current, the finished ladder followed a change in 0 of 400 runs and the group that kept checking and trying things followed in 37.8%; what carries an organization across is checked exploration, which the ladder as built lacks, and given it the ladder followed in 69.5% (Section 5.7). The model says nothing about the second valley: quality in it is given, so it cannot show use without value.

The fuller argument, including the author's observation from practice that the first valley is where the money stops, what the NIST study reports about the companies that kept early-stage funding alive, and how the roles of champion, sponsor and gatekeeper change, is in the companion note *Three Valleys* (Atkinson, 2026c). The change runs suggest a fourth role, the checker who stays. This is a proposal and not a finding: the simulations bear on two of the three valleys, and we did not study how organizations fund the carry.

8 Limits

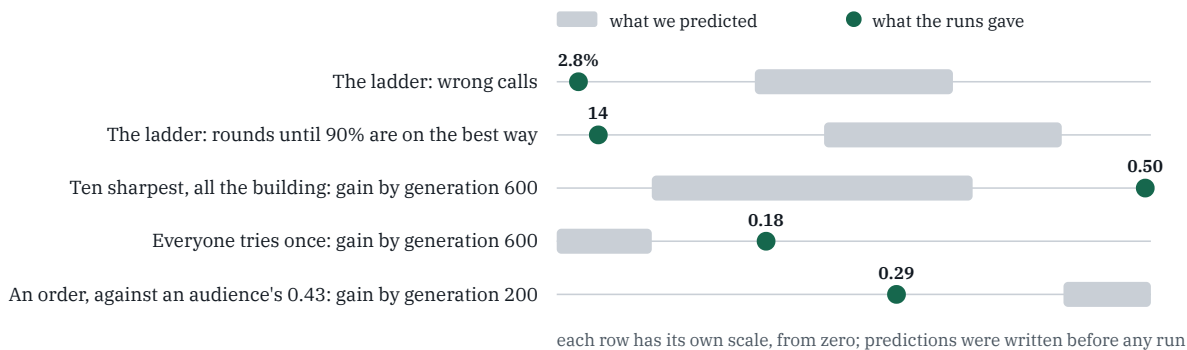


Figure 17. Five of the predictions written into the scripts before the runs, against what the runs gave. **Simulated; predictions as written in the scripts.**

These are simulations of a simplified model, not of people or of any organization. The mapping from two rollout philosophies to rules about who checks, how well and when is ours, and the advocates of either might draw it differently; in particular, a real capability ladder governs harmful actions by an automated system, which the model does not contain, and the model's "checks" are a member trying both ways themselves, which is stronger than reviewing an output a tool shows them. The ten sharpest are picked on their true reading noise, which a real organization would have to estimate, and the sevenfold gap between them and ten random members depends on the assumed spread of noise (half-widths from 0.1 to 0.5). The organization's settings (15% deference, a lead weight of three, the 60/15/5/20 hearing mix, the 0.3 weight on new readings, five options, 200 members) were each run at one value, with the flat organization the only contrast, so how far the pattern carries is untested. Members differ only in how accurately they read results; they are honest, they do not learn to check better, they have no incentives, and they are not subject to automation bias, which would make real checks weaker than modeled ones for everyone, including the sharpest. There are five options with fixed qualities, except in the change runs. Waiting, working the wrong way and checking are counted

and not priced, and the break-even in Section 5.2 assumes linear costs. Rounds are rounds of the model and map to no calendar. Each cell has 400 runs, so a reading of 0 has a 95% upper bound of about 1%. Most of our written predictions failed, all in the same direction (Figure 17): the organization settled faster and more accurately, once anybody checked, than we expected; for example, we predicted 25% to 50% wrong calls under the honest ladder and 90 to 170 rounds to reach 90% on the best way, and measured 2.8% and 14. In the building model, building costs nothing and takes no time, and the ten sharpest with the same total building are assumed able to build twenty times as often as an average member, which is the real-world question the position rests on. Everyone's builds come from the same distribution by design; strong builders who also build better things would widen the gap. There is no cost of coordination or fragmentation in its dynamics, and no correlated error, a build that looks good to everyone and is not. The quality ceiling compresses its readings at generation 600, and its NK landscape has 12 bits, which leaves a short climb. Agents in models of this kind may copy more than people do (Mason & Watts, 2012). Of the building model's first 18 written predictions, 6 held, 8 held in part and 4 failed, again mostly toward more improvement than we expected; we predicted that an order would do at least as well as an audience, and it did not; the four predictions written for the checked order held; of the rugged model's 8 predictions, 4 held and 4 held in part. Several details needed to re-implement the model are not given here: each member's starting way and estimate, what waiting members hold and whether they count in the call, whether deference and lead weights apply inside the first group during its head start, what exploration draws and whether it is checked, and tie rules. Our code is not public, so no one else has reproduced these results, and none of this work has been externally peer reviewed.

Disclosure. Drafting and literature synthesis were assisted by AI models (Claude, Anthropic) working under the author's direction; the author is responsible for the content. The experiments were designed under the author's direction and written and run by an AI agent (Claude, Anthropic, in Claude Code). The works in the References were read by AI reading agents working under the author's direction, at the depth stated below; where a number is taken from a work read in abstract only, the text says so. Every number in the paper was read from the run logs, and every chart was drawn from them by a script; the plotted values are recorded beside the scripts for all but Figures 11 and 14, which read the building model's run logs directly. Figure 1 shows numbers reported by the organizations named, Figures 2 and 10 are diagrams of the models, and Figure 16 is a drawing of the proposal in Section 7.4. The articles by Enquist, Eriksson and Ghirlanda (2007), Rogers (1988), Bikhchandani, Hirshleifer and Welch (1992), Banerjee (1992), Rendell et al. (2010), Valente (2012), Parasuraman and Manzey (2010), Parasuraman, Sheridan and Wickens (2000), March (1991), Levinthal (1997), Lazer and Friedman (2007), Fang, Lee and Schilling (2010), O'Boyle and Aguinis (2012) and Brynjolfsson, Li and Raymond (2025) were read in abstract only, from their records at OpenAlex, with titles, authors and venues matched; Otis et al. (2026), Beck, Beatty and Sackett (2014) and Sung et al. (2003), added on 7 October 2026 after the review, were read in abstract or summary only, at publisher and index pages; Barkoczi and Galesic (2016) and Mason and Watts (2012) were read at PubMed Central. Microsoft (2026), The Next Web (2026) and the Department for Business and Trade (2025) were read at their own pages for the passages cited, the last in its published PDF. Anthropic (2026) was read in full; the interview (York, 2026) was read in full in YouTube's auto-generated transcript, which garbles some names; Valentini, Hamann and Dorigo (2014) and Soma et al. (2024) were read as described in our earlier note. Markham et al. (2010)

and Auerswald and Branscomb (2003) were read in abstract only, the first at OpenAlex and the second on its publisher's page, with the latter's proposal of a Darwinian sea read as quoted in a secondary source; Branscomb and Auerswald (2002) was read at its public PDF for the passage cited. Rogers (2003) and Kauffman (1993) are cited for standard framings and were not read for this note. **Competing interests.** The author owns Wolfberg LLC, and the interview cited is his own. The work was not funded. **Code and data.** The simulation code, the run log (one record per configuration, with parameters, seeds, metrics and the script's hash) and the figure data are available from the author on request.

Open questions

Readings of our own runs that cut against a position this note takes. They are collected here rather than beside the positions they concern, and left open.

- **Section 7.1, adaptability.** From our own runs: the adaptability comes from the checked exploration, not the plan, since a ladder given the same habit followed more changes, 69.5% (Section 5.7).

References

- Anthropic. (2026, October 2). *Anthropic invests \$100 million to train 10,000 engineers and tackle the enterprise AI talent gap* (Claude Frontier Academy). anthropic.com/news/claude-frontier-academy
- Atkinson, B. (2026a). *When the hive gets it wrong: Where a swarm of imitators stops acting like one reinforcement learner*. Wolfberg research note, 30 September 2026. wolfberg.ai/papers/when-the-hive-gets-it-wrong
- Atkinson, B. (2026c). *Three valleys: Spread, value and staying current in AI adoption*. Wolfberg research note, 3 October 2026.
- Auerswald, P. E., & Branscomb, L. M. (2003). Valleys of death and Darwinian seas: Financing the invention to innovation transition in the United States. *The Journal of Technology Transfer*, 28(3-4), 227-239. <https://doi.org/10.1023/A:1024980525678>
- Banerjee, A. V. (1992). A simple model of herd behavior. *The Quarterly Journal of Economics*, 107(3), 797-817. <https://doi.org/10.2307/2118364>
- Barkoczi, D., & Galesic, M. (2016). Social learning strategies modify the effect of network structure on group performance. *Nature Communications*, 7, 13109. <https://doi.org/10.1038/ncomms13109>

- Beck, J. W., Beatty, A. S., & Sackett, P. R. (2014). On the distribution of job performance: The role of measurement characteristics in observed departures from normality. *Personnel Psychology*, 67(3), 531-566. <https://doi.org/10.1111/peps.12060>
- Bikhchandani, S., Hirshleifer, D., & Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy*, 100(5), 992-1026. <https://doi.org/10.1086/261849>
- Branscomb, L. M., & Auerswald, P. E. (2002). *Between invention and innovation: An analysis of funding for early-stage technology development* (NIST GCR 02-841). National Institute of Standards and Technology. nist.gov/system/files/documents/2017/05/09/gcr02-841.pdf
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*, 140(2), 889-942. <https://doi.org/10.1093/qje/qjae044>
- Department for Business and Trade. (2025, August 28). *Microsoft 365 Copilot pilot: DBT evaluation report*. GOV.UK. gov.uk/government/publications/microsoft-365-copilot-pilot-dbt-evaluation-report
- Enquist, M., Eriksson, K., & Ghirlanda, S. (2007). Critical social learning: A solution to Rogers's paradox of nonadaptive culture. *American Anthropologist*, 109(4), 727-734. <https://doi.org/10.1525/aa.2007.109.4.727>
- Fang, C., Lee, J., & Schilling, M. A. (2010). Balancing exploration and exploitation through structural design: The isolation of subgroups and organizational learning. *Organization Science*, 21(3), 625-642. <https://doi.org/10.1287/orsc.1090.0468>
- Kauffman, S. A. (1993). *The origins of order: Self-organization and selection in evolution*. Oxford University Press.
- Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, 52(4), 667-694. <https://doi.org/10.2189/asqu.52.4.667>
- Levinthal, D. A. (1997). Adaptation on rugged landscapes. *Management Science*, 43(7), 934-950. <https://doi.org/10.1287/mnsc.43.7.934>
- March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, 2(1), 71-87. <https://doi.org/10.1287/orsc.2.1.71>
- Markham, S. K., Ward, S. J., Aiman-Smith, L., & Kingon, A. I. (2010). The valley of death as context for role theory in product innovation. *Journal of Product Innovation Management*, 27(3), 402-417. <https://doi.org/10.1111/j.1540-5885.2010.00724.x>
- Mason, W., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3), 764-769. <https://doi.org/10.1073/pnas.1110069108>

- Microsoft. (2026, July 29). *Microsoft fiscal year 2026 fourth quarter earnings conference call*.
microsoft.com/en-us/investor/events/fy-2026/earnings-fy-2026-q4
- The Next Web. (2026, September 23). *OpenAI's Colin Jarvis says enterprise AI is stuck on deployment, not models*. thenextweb.com/news/openai-colin-jarvis-enterprise-ai-deployment-fde-humanx
- O'Boyle, E. H., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology*, 65(1), 79-119. <https://doi.org/10.1111/j.1744-6570.2011.01239.x>
- Otis, N. G., Clarke, R., Delecourt, S., Holtz, D., & Koning, R. (2026). The uneven impact of generative artificial intelligence on entrepreneurial performance: Evidence from a field experiment in Kenya. *Management Science*, advance online publication.
<https://doi.org/10.1287/mnsc.2024.06909>
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.
<https://doi.org/10.1177/0018720810376055>
- Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, 30(3), 286-297. <https://doi.org/10.1109/3468.844354>
- Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., Fogarty, L., Ghirlanda, S., Lillicrap, T., & Laland, K. N. (2010). Why copy others? Insights from the social learning strategies tournament. *Science*, 328(5975), 208-213.
<https://doi.org/10.1126/science.1184719>
- Rogers, A. R. (1988). Does biology constrain culture? *American Anthropologist*, 90(4), 819-831.
<https://doi.org/10.1525/aa.1988.90.4.02a00030>
- Rogers, E. M. (2003). *Diffusion of innovations* (5th ed.). Free Press.
- Soma, K., Bouteiller, Y., Hamann, H., & Beltrame, G. (2024). *The hive mind is a single reinforcement learning agent*. arXiv:2410.17517. arxiv.org/abs/2410.17517
- Sung, N. S., Crowley, W. F., Jr., Genel, M., et al. (2003). Central challenges facing the national clinical research enterprise. *JAMA*, 289(10), 1278-1287.
<https://doi.org/10.1001/jama.289.10.1278>
- Valente, T. W. (2012). Network interventions. *Science*, 337(6090), 49-53.
<https://doi.org/10.1126/science.1217330>
- Valentini, G., Hamann, H., & Dorigo, M. (2014). Self-organized collective decision making: The weighted voter model. In *Proceedings of AAMAS 2014* (pp. 45-52).

York, R. (2026, September 27). *Video interview on AI and work: Berg Atkinson* [Video]. YouTube.
youtu.be/Ver3Mjj6JWs