ADAM RUSSELL

How I Learned to Stop Worrying and Love Intelligible Failure

With the growth of "ARPA-everything," it's time to start testing predictions about what makes the model succeed. That means being able to understand when and why it doesn't.

When I was appointed by the US health secretary to serve as the very temporary, de facto leader of the newly created Advanced Research Projects Agency for Health (ARPA-H) in May 2022, I knew I didn't know enough. Or at least not enough to ensure that my small team and I could set up an ARPA to actually do what it was supposed to, rather than end up as a cargo cult with all the ARPA window dressing and none of the substance.

Allow me to explain.

ARPAs are built to produce innovation. They are charged to bring about world-changing advances through technologies that do not yet exist. The examples everyone trots out are ARPAnet (the predecessor to the internet), GPS, and self-driving vehicles. Delivering such technologies requires taking risky bets on ideas that could be big—if they work. The question, as I see it, is whether we can make them big even if they don't work.

Since the inaugural DARPA (for defense) debuted in 1958, there have been a stream of follow-ons. An ARPA for intelligence (IARPA), for energy (ARPA-E), for health (ARPA-H), and for infrastructure (ARPA-I). With growing worries that the return on investment in science and research is falling, there's an ever louder call for more ARPAs both in the United States and abroad. So there is now the United Kingdom's Advanced Research and Invention Agency (ARIA), Germany's Federal Agency for Disruptive Innovation (SPRIN-D), as well as nonprofits, which in the United States include the Advanced Education Research and Development Fund (better known as AERDF) and innovation incubators such as Convergent Research and Speculative Technologies. And those are just the ones that I'm aware of.

Illustration by Shonagh Rae

I had the privilege to serve as a program manager at DARPA and, before that, IARPA (during its very earliest days, in fact). I am a sociocultural anthropologist—a.k.a. the DARPAnthropologist—who has spent a lot of time focusing on promoting innovation in the social sciences, which is helpful for scaling cooperation, building better institutions, and designing complex, effective human systems.

So when I was asked to help take ARPA-H from an idea to an agency, how could I say no? Which may have been another unspoken qualification: I was just crazy enough to take the job, which came with no real authority, an uncertain budget, a skeleton crew of great people but with relatively little ARPA experience, and the intense attention of legislators with strong (and divergent) opinions about ARPA-H's mandate. I didn't even have the authority, at that point, to hire program managers, or PMs, who are the lifeblood of any ARPA.

After 11 years of working at ARPAs, I knew how hard it would be to get a brand-new ARPA ready to start changing what people think is possible. What I didn't appreciate at first was how little was actually known about what makes an ARPA tick.

Let me be clear. While we don't have enough data or observations, and certainly not a big enough sample size to know what, really, is key to ARPA-ness, we *do* have plenty of opinions. Which, in the spirit of work I'd been doing on forecasting, I came to see as predictions. All these opinions could ultimately be reduced to an implicit prediction: "If you do (or don't do) *x*, then you will (or won't) get *y*."

Now, these opinions (read: predictions) are far from uninformed. This is clear even from just a little digging into the smorgasbord of "what makes DARPA work" predictions of innovation scholars and former ARPA leaders. Some of them predict ARPA's success depends on the trio of ambitious goals, temporary project teams, and independence, to which others add trust and a sense of mission. There are predictions about the importance of having a portfolio big enough to take "many shots on goal" across different technology sectors. There are predictions sewn into a book called *The DARPA Model for Transformative Technologies*; in former DARPA head Regina Dugan's podcast interviews; and in innovation entrepreneur Ben Reinhardt's blog, "Why Does DARPA Work?" Academic articles predict that ARPAs succeed by being "activist organizations" and "public sector intermediaries between science and industry," argue that success depends on understanding the ecosystem in which an ARPA operates, and point to factors that might stand in the way of future successful ARPAs. And more.

Collectively, these predictions are insightful but paradoxical, and difficult to operationalize. How does an ARPA create processes to "remain agile" but also "build continuity"? What mechanisms are best to "avoid the 'false fail,'" while making sure that the agency "kills lots of early ideas"? Which organizational structure best enables the agency to "always keep looking for new opportunities," yet "maintain focus at all times"? What kinds of management and support staff is required to "empower your people to take big risks," while making sure that "nobody is bigger than the agency"?

An ARPA is about something even more abstract than "innovation." Its real job is to motivate progress by grappling with huge, intractable problems. So I tried to approach this what-makes-an-ARPA-tick task as an ARPA-hard problem in itself. After all, many important problems are ARPA-hard precisely because there's no immediately credible way to measure whether a solution will work or not, which means an ARPA needs to be distinguished by being able to use failure to find a path forward. I call this idea "intelligible failure."

To put it bluntly, the number of ARPAs and ARPA-like agencies is growing, while best practices lag. How can any ARPA know whether its processes are effective, or even whether the assumptions behind them are valid? My own prediction is that learning to use failure as a way to test predictions will build that knowledge base. But to do this, an ARPA needs a culture that prioritizes intelligible failure.

What does that take? A first step is to realize that a technical failure is different from a mistake; in fact, a lack of technical failures might indicate the mistake of playing it too safe. A second step is collecting data, which could include everything from getting PMs' own predictions of a program's success to postmortem analyses of what went right and what went wrong. And a third step is learning how to apply these data to assessing the plethora of predictions about what's key for an ARPA. That's super tough since we're not even sure how to measure an ARPA's impact. But emerging scholarship in metascience promises to make these problems more tractable.

The number of ARPAs and ARPA-like agencies is growing, while best practices lag. How can any ARPA know whether its processes are effective, or even whether the assumptions behind them are valid?

I predict that the bigger difficulty in prioritizing intelligible failure will be in the willingness to do so. For all the emphasis on failure as inherent to an ARPA's success, it is understandable that "failure" remains a dirty word. Even DARPA tends to say it is a place that "tolerates" failure, rather than valorizing failure as a principal teacher.

There's a perfectly good reason ARPAs don't glorify failure or prioritize intelligibility: doing so invites all kinds of criticism, and that can be tough for organizations that, by design, lack careerists who can defend their institution. Especially in conditions of low trust, it's far more comfortable to avoid scrutiny; an agency can't be attacked for what it isn't set up to know.

But that avoidance ultimately does ARPAs a disservice. If an organization can't learn from failure, then it can't quantify—much less communicate—how failure contributes to its mission. And if the organization doesn't build in feedback loops to reward turning failure into insight, expect people to make safer—if still flashy—bets. I found it informative that, among all the predictions out there, not a single one posits that taking safer bets is the way to find those rare, powerful technologies that expand the scope of the possible and make for a safer, healthier, happier world.

Six predictions to think with

Time for me to ante up: my time at the ARPAs has left me with some predictions about what an ARPA needs to be effective, along with baby steps toward testing those predictions.

To be clear—and because I expect at least a few of these predictions to be wrong—I don't speak for anyone else about what an ARPA needs to succeed. I'm even inventing a few words to go along with these predictions, in a nod to linguists Edward Sapir and Benjamin Lee Whorf's suggestion that new words can facilitate new ideas. I offer these predictions as tools that are "good to think with," if only to better prove my intuitions incorrect.

My hope is not that you accept my predictions, but that you start to consider how to go about proving that they're wrong—all the better to build knowledge of what really makes ARPAs succeed and fail. Given the complexity of the problems humans are facing, we all have an interest in having our ARPAs use intelligible failure to boost success.

Prediction 1: A compelling origin story feeds "endurgency"—an enduring, mission-driven sense of urgency that promotes an ARPA's success. It is rare to find anyone at—or even familiar with—DARPA who is unacquainted with its Sputnik origin story, when the Soviet Union appeared to threaten US technological superiority by becoming the first nation to launch a satellite into orbit. Similarly, in the post-9/11 era, everyone at IARPA is aware of what can happen if the United States loses its intelligence advantage. Sputnik and 9/11 have become the basis of unofficial origin stories for those ARPAs, which, like most origin stories, infuse history with mythology.

The value of origin stories to technocentric ARPAs may seem trivial, but I predict that they are essential for fueling an ARPA's sense of "endurgency," or enduring urgency. Yes, term limits add to PMs' drive to get things done, but endurgency creates a collective drive, an institutional hyperawareness of what happened—and what could happen again—if that ARPA doesn't take big, principled bets. Hence ARPAs gain that sense of "what you stand to lose," which, according to prospect theory, motivates more risk-taking than "what you might gain."

Endurgency should spur an ARPA to make intelligible failure a priority, since it means that every bet it takes will contribute to its mission. Endurgency also keeps an agency from overlearning from failure, in particular getting hung up on past failures and falling into the WCSYC—We Couldn't So You Can't—mindset.

How could we disprove this prediction? One idea might be to survey whether staff share a common mental model of what their Sputnik moment is, why their ARPA is essential to preventing another one, and whether that predicts impact—or not.

Prediction 2: Being "catechommitted," or sticking to a clear framework for saying "yes" or "no" to programs, will be key for making an ARPA successful. I often joke that if you want to make a former ARPA PM sweat, just say "Heilmeier Catechism." It's only partly in jest. Those deceptively simple questions formulated by the 1970s DARPA director George Heilmeier range from "What are you trying to do?" to "Who cares?" The task of formulating coherent answers to these questions is notorious for leaving PMs (and those who apply for funding) with equal parts emotional scar tissue and hard-earned experience. It's not uncommon for PMs developing their programs to take months to credibly answer these questions. The catechism imposes a kind of ruthlessness on all the decisions that are made in an ARPA, from picking projects to enforcing PMs' term limits to ending programs. It is the opposite of fun to have to decide—or be told—that a program is winding down. It is also crucial.

The Heilmeier Catechism institutionalizes a principled ruthlessness into an ARPA's business model. It serves as an oft-needed counterweight to government tendencies to never end anything once started, while avoiding the Silicon Valley "hopium" of tech bubbles and investment stampedes. In other words, DARPA employees use the catechism because it brings clarity to making hard decisions. Being hard on themselves as decisionmakers is a feature, not a bug, of the ARPA model.

But it can be tough to maintain this ruthlessness, or to be "catechommited." I've heard too many people say that the Heilmeier Catechism is "suggestive" and shouldn't be taken too seriously. Worse is seeing the catechism applied retrospectively, as a post-hoc justification for decisions made by instinct instead of principle. The result is that a schism develops between an ARPA's declared operational model and how decisions are actually made. Thus, hard decisions can end up being arbitrary, self-serving, and purely tactical, generating distrust and inviting real danger for any ARPA. Ultimately, everyone inside and outside the agency, and especially those told "no," need to feel that they understand the principles behind a decision.

How could we test the prediction that being catechommited means being a more effective ARPA? The famous opacity of ARPAs, along with regular personnel turnover, complicates the collection of data, but I think it can be done if an ARPA operationalizes intelligible failure as a disciplined process, rather than simply focusing on outcomes (or what poker champion and decision science pundit Annie Duke calls "resulting"). Can an ARPA point to both failures and successes that came out of the same principled process? And does that predict big achievements?

Prediction 3: "Empathineering," or engineering organizations for empathy, is important for an ARPA's success. ARPAs are largely thought of as technology shops. This stereotype can promote the mistaken notion that ARPAs have managed to dispense with human irrationality—

emotions, incentives, need for belonging, psychological safety, and the like. The reasoning is that an all-encompassing focus on the mission erases interpersonal competition and political dynamics, which are miraculously replaced by technology-focused transactions such as awarding project funds, setting goals, signing contracts, and transitioning results. But no one who's worked inside an ARPA actually believes this.

As the late computer scientist Gerald Weinberg said, "No matter how it looks at first, it's always a people problem." ARPAs are very much human endeavors, and successful ARPAs both acknowledge and commit to managing humans as much as technology—not in spite of, but in pursuit of, the highest standards. That means managing the incentives, egos, identities, personalities, subcultures, ambitions, and other human elements that shape the timbre of an ARPA. By actively engineering the human aspects—for employees, advocates, transition partners, and performer communities—effective ARPAs will embrace what I call "empathineering."

Empathineering also helps an ARPA better serve us humans who will ultimately benefit from the outcomes of ARPA programs. Site visits, "immersion trips" (where ARPA personnel spend time in the field—or the desert, swamp, or submarine), and cross-office engagements are all investments in building what network theorists call "weak ties," as much as they are in knowledge-gathering. These weak ties, far from being from purely transactional, form the sinew to anchor ideas and ARPA-worthy problems as different minds collide and commingle.

A real danger, however, is that these human dynamics may be dismissed as distractions by an ARPA that focuses exclusively on building technology. If conflicts are ignored or allowed to fester, a poisonous brew of misaligned incentives, personality struggles, and office politics will erode an ARPA's ability to build weak ties and psychological safety. Without empathineering, ARPA staff will be afraid to risk failure, let alone make it intelligible, because taking risks will be seen as expanding an attack surface rather than a way to achieve better outcomes.

How could we disprove my prediction that empathineering matters to the most tech-y of places? One way may be as simple as getting feedback about whether staff—and not just PMs—are unafraid of failure and willing to try new things because they believe even their failure can contribute to the ARPA's success. (They should also believe they will be supported accordingly.) It would be interesting to see whether such beliefs predict impact. Of course, getting good feedback requires spending time learning how to ask the right questions, which, I note, is one currency of anthropology. (I pause here briefly to appreciate the not-insignificant irony of social science potentially being a key to an ARPA's success.)

Prediction 4: "Solvationism" will reduce an ARPA's impact by causing it to seek out ready solutions at the expense of finding ARPA-worthy problems. To succeed, ARPAs need the resources and willingness to fail more often than succeed. (The unofficial consensus is that the "hit rate" of successes should be somewhere around 5%–30% of programs). Therefore the agencies need a portfolio of investments that reflect their somewhat absurd, ARPA-level ambitions—in part by being brave enough to define and tackle problems whose solutions are currently "in the realm of the barely feasible," in the words of former DARPA head Arati Prabhakar.

But there's a conflict here because many of an ARPA's own advocates will be clamoring for game-changing technology that they want within months, not years. That kind of pressure can drive an ARPA toward seeking out projects with ready solutions, problems which might better be left to other kinds of agencies. I call this "solvationism."

Solvationism is the tendency to protect the organization's reputation (and budget) by looking less for "ARPA-worthy" problems, which rarely have obvious solutions, and more for problems that make the ARPA look worthy—that is, they have straightforward, if not simple, solutions. This results in window-dressing successes that might provide short-term achievements but can signal that an ARPA has stopped existing for its mission, and like many bureaucracies has now made its mission to exist.

How could we test my prediction about solvationism? I think making it routine practice to estimate risk for each project and track failure rates could go a very long way. Yes, that would expose an ARPA to criticism, but it would also reveal an important insight: how well the ARPA forecasts risk. This would lend confidence (and data!) to an ARPA's assertion that it's tackling risky problems, as well as help quantify the impacts if an ARPA edges towards solvationism. Maybe such a thing as solvationism doesn't exist, or maybe it doesn't matter. But how would we know one way or the other?

Prediction 5: "Badvocacy," or heavy-handed influence from powerful advocates, will erode an ARPA's degrees of freedom and thus its long-term impact. While ARPAs are much more than "100 geniuses connected by a travel agent," as DARPA has memorably described itself, they are nonetheless bureaucratically lean. Without an army of career federal employees, ARPAs have to rely on advocates—senior executives, legislators, performers (both in industry and elsewhere), and of course customers—to do much of the organizational knife-fighting that's required to keep an ARPA sufficiently funded and protected from interference. That means having powerful champions who protect the ARPA but keep their hands off it (which is why I think they should be called "champioffs").

But often an ARPA's strongest advocates also have equally strong opinions about what an ARPA should fund, how an ARPA should operate, and what kinds of people should work there. Accordingly, ARPAs are always susceptible to

ARPAs are very much human endeavors, and successful ARPAs both acknowledge and commit to managing humans as much as technology—not in spite of, but in pursuit of, the highest standards.

what I call "badvocacy" (bad plus advocacy), when powerful advocates try to pressure, legislate, or otherwise influence an ARPA's decision-making, whether about resources, personnel, or even what is and is not within scope of the ARPA's mission. Whatever the mechanism, the result is an erosion of degrees of freedom that can be asphyxiating for an organization that works because of those degrees of freedom in the first place.

How could we test this prediction? While it's informative, if a bit anachronistic, to juxtapose the two-page 1958 directive that became DARPA to analogous documents for more recent ARPA-like organizations, we might test it by quantifying the degrees of freedom an ARPA has for making unpopular decisions, and whether those decisions predict the agency's impact. Another approach might be to track whether badvocates praise and protect an ARPA's intelligible failure or focus instead on advertising successes, a focus which may in turn lead to solvationism.

Prediction 6: "Alienabling," or recruiting, empowering, and protecting boundary-crossing "aliens" to define and tackle ARPA-worthy problems, will improve an ARPA's long-term impact. Let's accept that an ARPA ultimately depends on its people. Besides PMs empowered to take high-risk, high-reward bets, an ARPA also depends on having extremely competent people who make the organization work: human resources, contracting, IT systems, and more. (It is a fact too infrequently acknowledged that the ratio of ARPA support staff to those vaunted PMs can be as high as five to one.)

For both PMs and support staff, I predict that success depends on an ARPA attracting a type of individual I call an "alien" (and then "alienabling" them). I derive these terms from work by sociologist James Evans and others into how breakthroughs often occur when scientists leave their home worlds for entirely new spaces in the research universe. These aliens are different from "colonizers," or those experts who coopt another field without taking time to understand it. Instead, aliens travel into new spaces out of a complex mix of genuine curiosity, dissatisfaction with their own disciplines, and a drive to tackle problems that cannot be solved with existing thinking.

While expertise has real value, experts may often be at a disadvantage with ARPA-worthy problems; they tend to be hedgehogs, who know one big thing. And their history of success in a few domains can mean they overgeneralize their competence, assuming they already know all they need. Aliens, however, often leave footprints in many places, without settling into any conventionally defined space or role. They tend to remain more "learner" than "knower," which some evidence suggests may set them up better to learn from failure.

Aliens are hard to spot and even more difficult to characterize, whereas experts, with their publications and other evidence of conventional success, are easy to identify. And experts' acclaim makes their recruitment and promotion equally easy to justify. This can lead an ARPA to exhibit "expertilection" (a predilection for experts) that hampers the very alienabling necessary for finding and tackling uniquely ARPA-worthy problems.

How could we test this prediction? Emerging methods of visualizing interactions, social networks, funding sources, teams, and even influence hold some promise in identifying talented boundary spanners. These tools could assess where "ARPAliens" may be found, whether enough are being produced, whether they're being sufficiently recruited and supported, and where they can be most useful. Such work may also encourage efforts to deliberately develop more aliens. A useful array of tools, data, and methods are emerging to help identify underexplored areas of the research universe, and perhaps enough are coming online to test whether my prediction about alienabling holds any weight.

Onward to intelligible failure

These may be the same tools and methods that can help make failure intelligible. Just as (I predict) ARPAs need aliens to find the best problems, they also need intelligible failure to learn from the bets they take. And that means evaluating risks taken (or not) and understanding—not merely observing—failures achieved, which requires both brains and guts.

That brings me back to the hardest problem in making failure intelligible: ourselves. Perhaps the neologism we really need going forward is for intelligible failure itself—to distinguish it, as a virtue, from the kind of failure that we never want to celebrate: the unintelligible failure, immeasurable, born of sloppiness, carelessness, expediency, low standards, or incompetence, with no way to know how or even if it contributed to real progress. Until we have that perfect neologism, I predict that promoting intelligible failure requires a word that has characterized the best ARPAs to date, and that I hope all ARPAs will keep as a lodestar: courage.

***Adam Russell**, a veteran of three ARPAs, directs the artificial intelligence division at the Information Sciences Institute at the University of Southern California.*