



Augmented Intelligence for Public Good

A DISCUSSION PAPER

The Use of What We Know

Building Decision Intelligence for Institutions That
Learn, Adapt, and Deliver Better Results

Rolf Sartorius

Founder & CEO, Saipion, Inc.
Chairman & Co-Founder, Social Impact, Inc.

DRAFT FOR DISCUSSION
August • 2026

A NOTE ON THIS DOCUMENT

This paper draws on nearly four decades of experience advancing public sector outcomes around the world: designing public programs, evaluating them, building the performance management systems meant to steer them, and strengthening the skills of public officials.

The case has two parts. We can now bring what we already know to bear at the moment a program is designed. We can also now watch a program closely enough during implementation to correct it while correction still matters. Artificial intelligence is what makes both newly practical, in an assisting role rather than a deciding one.

I argue that the current constraint is not the quality or the volume of evidence. The constraint is the ability to identify evidence, judge it, and use it in time. I also argue that the knowledge held by the people who use public services belongs alongside evaluations and statistics, and matters as much as either for public outcomes.

Contents

The question, and when we ask it	4
Where the difficulty lies	5
One decision, and what stands behind it	5
Twenty years after the evaluation gap	6
The voice that is missing	8
Five convictions	9
One: better public outcomes are the point	9
Two: the binding constraint is use, and it has become solvable	9
Three: the people who use services belong in the design and the delivery	10
Four: decision intelligence is the bridge, not the destination	10
Five: a shared reservoir of public knowledge	11
Two moments	12
The first moment: design	12
The second moment: the long middle	13
From efficiency to outcomes	14
What has to be built for better public outcomes	15
What this asks of us	16
Of governments	16
Of development institutions	16
Of the AI field	17
What I am not saying	17
Ten years out	18
How we would know	19
The measure	19

PREFACE

The question, and when we ask it

My aim in this paper is to describe two things that have recently become possible in public decision-making, and to say what each would ask of governments, development institutions¹, and the people building artificial intelligence, myself included.

I begin with an observation from practice. Public institutions ask whether something worked far more rigorously than they ask, during program design, what is likely to work here.

For nearly forty years I have helped public institutions in Washington and in ministries from Nairobi to Jakarta deliver better outcomes. I have designed public programs, evaluated them, built the performance management systems meant to steer them, and strengthened the skills of the officials who run them. Across all of that work, the question asked with the most care is whether the program worked. That is a good question. By the time it is answered, the money is spent, the design is set, the team has moved on, and the people the program was meant to serve have already lived through whatever it turned out to be.

The reason is not a lack of effort or investment. Governments and development institutions have built an immense evidence infrastructure. It includes evaluation offices with real independence, learning agendas and evaluation plans required by statute, national statistical systems, RCT-specialist organizations, household survey programs, administrative data platforms, clearinghouses, and systematic review bodies. Hundreds of thousands of studies have been produced, many of them excellent. The production problem is broadly solved.

The use problem is not. Four decades of literature describe why. Evaluations address the wrong questions, arrive too late, lack stakeholder ownership, are not actionable, meet organizational or political resistance, are poorly communicated, or never say whether the findings apply to the case at hand. Officials do not ignore evidence. In my experience they look for it, ask for it, and are often frustrated by what they find. A team on deadline to prepare a design decision faces a body of relevant knowledge that is enormous, scattered across incompatible systems, written in several professional dialects, uneven in quality, and mostly about somewhere else. Identifying the relevant fraction, judging it, and reconciling the parts that disagree takes a skilled team many months.

This is not a new observation for me. Use has been the standing question of my career, starting with my graduate work with Michael Quinn Patton at the University of Minnesota, where the subject was evaluation utilization. Patton's argument was that an evaluation should be judged

¹ Although this paper is primarily meant to address governments at all levels and development institutions, much of the argument applies to foundations, nongovernmental organizations, and private funds with social or environmental objectives.

by whether it is used. Forty years on, the argument still holds and the problem has grown rather than shrunk. What has changed is that we now have a new way of addressing it.

That way is artificial intelligence and two things have changed. Identifying and synthesizing knowledge across large and heterogeneous bodies of evidence has become a tractable engineering problem rather than a limit on how fast people can read. And a program under implementation can now be watched continuously against what was expected of it, by systems that flag divergence early enough for someone to act. Both rest on machine reading, retrieval, and inference at a scale no team could reach before.

What I am not claiming is autonomy. These systems should assemble, reconcile, forecast, and flag. People should judge, decide, and answer for the decision. AI is what makes the argument in this paper newly practical, and it is the smaller half of the story. The larger half is whether institutions change what they count as a well-made decision.

The two changes matter because they land at the two moments when knowledge can still alter an outcome: the moment of design, and the long middle of implementation. Most of a program's life is spent in the second, and as Part Three describes, current performance and adaptive management practice serves it poorly.

Twenty years ago the binding constraint was the opposite one. What was scarce then was reliable knowledge about whether social programs worked. That case was made and it was won. What follows is a sequel to that argument.

A third thread runs through this paper, and it is often left out. Evidence tells you what has worked somewhere. Statistics tell you what conditions look like here. Neither tells you what people actually need, or will use, tolerate, own, and sustain once the funding ends. That knowledge exists. It sits with service users, and public institutions have never had a reliable way to hear it at the speed decisions are made.

PART ONE

Where the difficulty lies

The problem is not a shortage of evidence, and it is not a shortage of will. The problem is that there is no practical route from what is known to the question in front of someone.

One decision, and what stands behind it

Take a concrete case. A ministry of education, or a state education agency, has run a structured early reading program in a few dozen schools for three years. Teachers received scripted lesson materials. Children were grouped by reading level rather than by grade. Coaches visited classrooms monthly. The pilot results look encouraging.

The decision now is whether to take the program to every school in the system. If so, at what pace, with what coaching intensity, in which languages of instruction, under what rules for districts and providers, and at what cost per child who ends up reading.

This is a fairly ordinary public decision, and it is being made somewhere this month. Consider what is already known that bears on it. There are randomized trials and quasi-experimental studies of structured pedagogy, level-based instruction, teacher coaching, and materials provision, spread across many countries, school systems, and languages. There are process and performance evaluations, implementation studies, completion reports, mid-term reviews, and the pilot's own monitoring data. There are official statistics: learning assessments, enrollment and dropout series, teacher deployment records, school censuses, and household survey data on the language spoken at home. There is operational knowledge about whether the books arrived, whether coaches visited, and what the binding constraint on supervision turned out to be. And there is the experience of the people involved: what teachers found workable in a class of seventy, what parents noticed, what children could and could not do.

Almost none of it will reach the decision in usable form. The trials report average effects in the systems that were studied, and those may or may not be relevant here. The completion reports were written for accountability and are filed where nobody searches. The statistics sit in four agencies with different identifiers and different release calendars. The operational knowledge is in the heads of people who have rotated out. The teachers and parents were never asked in any systematic way, and the children were not asked at all.

So the decision gets made the way most public decisions are made. A small team has a few weeks. It assembles a briefing package from what it can find and from what the institution already believes. Its members will have read a great deal and pulled the studies they knew about. What defeats them is the calendar. Identification, retrieval, appraisal, and synthesis take longer than they have, and little machinery exists to do that work.

Twenty years after the evaluation gap

The current situation is in large part the result of a reform that worked. That deserves saying before I criticize where it left us.

In May 2006 the Evaluation Gap Working Group at the Center for Global Development, co-chaired by William Savedoff, Ruth Levine, and Nancy Birdsall, published *When Will We Ever Learn? Improving Lives Through Impact Evaluation*. The diagnosis was that we knew far less than we needed to about whether social programs worked, that rigorous impact evaluations were rare, and that the shortage was a collective action problem, because everyone benefited from knowing what worked and no single agency had sufficient reason to pay for finding out. The recommendation had two parts. Individual agencies should reinforce what they were already doing. And a group of them should jointly create and fund one new collective entity.

That entity became the International Initiative for Impact Evaluation (3ie), founded in 2008. It joined organizations that already existed. Innovations for Poverty Action was founded in 2002, J-PAL in 2003, and the World Bank's Development Impact Evaluation initiative in 2005. The report cites DIME by name as an example of what an institution could already do. What the report changed was the scale of the field and its claim on public money. The decade that followed brought agency evaluation policies, statutory evaluation requirements, well-funded research networks, trial registries, and reporting standards that had no equivalent in 2006. Rigorous impact evidence is no longer scarce. In the sectors that received the most attention it is abundant, well curated, and still growing.

That is why the problem has changed shape. Causal impact evidence remains the best way to learn what a program caused, and it is now a mature and well-supplied category. It is also one category among several, and that is the part the field has been slow to absorb. The movement's success established a single well-defined standard of rigor, and in practice that standard came to define what counted as good evidence. Evidence outside it was often treated as less sound.

A more accurate picture takes in several distinct kinds of knowledge. Each has its own logic, its own standards of quality, its own proponents and critics, and its own legitimate claim on a decision.

- **Causal impact evidence.** Trials and strong quasi-experimental designs. Strongest on what a program caused. Weakest on whether the result transfers to a different setting.
- **Implementation and process knowledge.** Process and performance evaluations, completion reports, supervision records, mid-term reviews. Strongest on whether a design survives contact with a real delivery system.
- **The statistical record.** Censuses, household and labor force surveys, facility censuses, geospatial data. Strongest on who and where, on starting conditions, and on how much places differ from the places that were studied.
- **Professional judgment.** The accumulated knowledge of people who have run these programs. Currently unrecorded, unpooled, and lost at every rotation.
- **Service user and local knowledge.** What people need, will use, tolerate, own, and sustain. Strongest on whether a program fits the place it lands in. Least often collected at a resolution that could influence anything.

Then there are new sources of data such as satellite and other remote observation. What matters is that each of these bodies of knowledge is deep, each is kept up by serious people, and each is walled off from the rest by different units, identifiers, vocabularies, quality standards, access rules, and owners. Connecting them is partly a technical problem and mainly a problem of incentives, standards, and institutional design.

None of these substitutes for another, and none of them is soft. Each is difficult to produce well and each is difficult to mine. The difference is that one of them received twenty years of deliberate infrastructure, including standards, registries, reporting conventions, synthesis bodies, and dedicated funding. The others received less. They are more abundant than ever and less usable than the category that got the investment, though there are now serious efforts to change that.

Two are underused for reasons worth naming. The profession spent thirty years building a hierarchy of evidence. It is a hierarchy of internal validity, and it was never a hierarchy of usefulness, though it has been read as one. A rigorous trial of a reading program in another country may tell a ministry less about its own system than a workmanlike process evaluation from the next province. The statistical record is underused for a different reason. It can answer much of what program designs currently assert rather than establish: who and where the affected population is, what the starting conditions are, and how much variation there is across districts. The material is public. It is also fragmented across agencies, inconsistently coded, and costly to assemble, so design teams may not be able to mine it, or they commission another baseline survey instead.

The evaluation gap has largely been closed. What remains open is the utilization gap, and it runs across every one of these bodies of knowledge, each with its own problem of identification, appraisal, and synthesis. Until recently, the synthesis work on top of them was beyond what any team could do inside a decision calendar. That last part is what has changed.

The voice that is missing

One body of knowledge is the one public institutions are worst at reaching. It is what the people a program is meant to serve know about their own circumstances.

This is usually treated as a matter of fairness or legitimacy. The practical case is stronger. Evidence and statistics cannot tell you whether a clinic's opening hours are the hours people can come, whether a training course runs in the planting season, whether the materials are in a language children speak at home, whether the enrollment process assumes a document half the target population does not have, or whether anyone locally will maintain the thing once the project closes. Those are the failure modes that account for much of the gap between programs that work in evaluations and programs that work.

Anyone who has spent time in the field has a version of the same story. The water point installed where the hydrology was right and the walking route was wrong. The youth training program with the right curriculum and the wrong schedule. The reading materials in the official language of instruction, in a district where children speak something else at home. These are not evidence failures, they are engagement failures.

Public institutions do have channels for this: consultations, grievance mechanisms, complaints, surveys, and elections. Each is real. Each is also slow, thin, backward-looking, and heavily filtered. Consultations may happen once, at design, when the design is largely settled. Complaints capture only those willing and able to complain, which excludes many of the people with the most to report. Evaluations mostly never reach service users. The result is an institution that answers upward to ministries, legislatures, auditors, and funders through strong and well-enforced mechanisms, and answers downward to the people it serves weakly or not at all.

PART TWO

Five convictions

These are the claims I think anyone working in this field should be willing to defend.

One: better public outcomes are the point

Governments do not exist to write reports, process forms, hold meetings, satisfy auditors, or deploy artificial intelligence. They exist to make people's lives better. Everything else is a means to that.

That sounds obvious and it gets ignored constantly. An agency that cuts benefits claim processing from forty days to four has produced an output. Whether it produced an outcome depends on whether the right people got the right benefit, whether fewer households fell into hardship while waiting, and whether improper payments went down. Those questions often go unasked, because the output was easier to measure and good enough for the announcement.

The discipline is easy to state and hard to keep. Every system and every reform should trace back to a change in the lives of real people, not to a modernization milestone or a maturity model, but to whether more children finish school able to read, more farmers earn more, more patients recover, and fewer households fall into poverty.

Two: the binding constraint is use, and it has become solvable

The common way of talking about AI in government is that it adds intelligence to institutions that lack it. That is close to backwards. These institutions are surrounded by intelligence they cannot reach and synthesize in time.

This matters because it changes what to build. If the problem is missing intelligence, the answer is better models. If the problem is knowledge that cannot be identified, retrieved, and reconciled inside a decision window, the answer is standards, identifiers, shared vocabularies, quality grading, provenance, and routes from evidence to decision. Most of the value is in the second. Most of the effort goes to the first.

It also changes what counts as evidence worth reaching for. A synthesis that carries the implementation record, the relevant statistics, and what service users reported will be messier and more contested than one built from the rigorous tier alone, and far more likely to change a decision for the better. The discipline is not to flatten these into a single confidence score. It is to keep the sources distinguishable, say plainly where they disagree, and let the person deciding see the disagreement.

And it changes who matters. Nobody produces a good synthesis of what is known about community health worker programs from a model alone. It takes the epidemiologist who knows why two similar-looking trials are not comparable. It takes the implementer who knows the supervision ratio was the binding constraint in both. It takes the district officer who knows neither design survives the local labor market. What has changed is that these people can now be put in front of the whole relevant body of knowledge rather than the fraction someone had time to find. AI makes the work possible at scale. It does not do the work.

Three: the people who use services belong in the design and the delivery

This is the conviction I would most like to see taken seriously, and the one most institutions struggle with. Part One made the case that the knowledge is real and that the existing channels do not carry it. The question here is what to do instead.

The answer is not more surveys. It is to treat service users as active participants in design, and to change what we take a service interaction to be during delivery. At design that means careful stakeholder analysis and systematic, not token, consultation or co-design. During delivery, when someone collects a benefit, visits a clinic, enrolls a child, or applies for a permit, the institution records a transaction. It could also record something about how well the system is working: whether the service arrived, whether it was adequate, whether it came on time, whether the person was treated decently, and whether it did what it was meant to do. At the scale of a national program that is a continuous, disaggregated signal about delivery quality, available in weeks rather than years, and capable of being broken down by district, facility, and group in a way no evaluation sample can match. It works only if the institution holds itself accountable to its service users, makes improvements, and shares results with them.

This is meant to be learning, not surveillance. It is legitimate only under four conditions: taking part is voluntary and informed; the data goes toward improving services, not toward policing the people who use them; nobody is identified in analysis without consent; and people can see what was reported and what changed because of it. A system that breaks those conditions is a control system in the guise of learning.

Done properly, this is the social accountability that Part One found missing, built into how the system works rather than bolted onto a project.

Four: decision intelligence is the bridge, not the destination

I use the term decision intelligence to mean the things that turn knowledge into better decisions. That includes identifying and synthesizing evidence, reasoning about context, forecasting, being honest about uncertainty and evidence gaps, structuring deliberation, and learning from what actually happens. It matters, and it is not the point. It is a bridge between two things that are.

Knowledge → better decisions → better outcomes

Keeping that order straight has practical consequences. A decision intelligence system does not succeed by being accurate. It succeeds when decisions change and outcomes improve. The test is not how the model scores on held-out data. The test is whether the institution made better decisions than it otherwise would have. If something simpler gets the same result, such as a checklist, a better template, or a monthly conversation with the people using the service, the simpler thing wins.

Five: a shared reservoir of public knowledge

This is ambitious in scope, and some progress is being made.

Almost every government solves problems that have already been solved. A state agency builds an eligibility system without access to what a dozen other states learned building theirs. A federal program office funds an approach whose central assumption was disproved by an evaluation a sister agency commissioned three years earlier. A ministry designs a cash transfer targeting system without knowing what seven other countries learned building one. The waste is not mainly financial. It is the waste of lessons paid for with public money and human cost, and then not used. Every failed program cost real people something. The least we owe them is that the lesson survives.

A shared reservoir of public knowledge would carry more than published findings. It would hold the conditions under which findings held and where they did not; the statistical base linked to the programs it describes, so a designer can see the starting conditions of comparable places rather than a national average; the implementation detail that decides whether a design survives; forecasts and how well past forecasts held up, so others can judge how far to trust them; what was changed partway through, why, and what happened next; and what service users said, in their own words.

Consider a state health agency learning from what four other states found about care management. A federal program office reading a sister agency's findings before it writes the funding notice. A ministry in Kenya learning from Tanzania or Indonesia. None of this requires a new institution with authority over anyone. It requires shared standards, careful curation, sensible protections, and a norm that lessons bought with public money stay in public

circulation. That knowledge was paid for by the public and by people's own experience. It should not end up belonging to whoever commissioned the study.

PART THREE

Two moments

This part sets out where the new capability actually matters. These are the two moments: the design decision, and the long middle of implementation where a program can still be changed. Predictive governance is the new input, meaning institutions that see problems while there is still time to fix them and treat looking ahead as routine. Decision intelligence is what makes that possible.

The first moment: design

Return to the reading program. What would be different if the identification and synthesis problem were solved?

The team would not begin with a blank page and three weeks. It would begin with an assembled, sourced picture of what is known about this kind of program: the rigorous studies and the conditions they held under; the implementation studies and completion reports from systems structurally similar to this one; the relevant statistics for the districts under discussion rather than the national average; what previous programs in this system tried, changed, and concluded; and what teachers, parents, and children in comparable settings reported. Assembled in hours, not months.

Three properties matter more than speed. Provenance, so that every claim is traceable to its source, with the quality and design of that source visible, and a reader can discount it appropriately rather than take a summary on faith. Disagreement preserved, so that where studies conflict the conflict is shown and characterized rather than averaged into a number. And honesty about thinness, so that where the evidence for a sub-question is weak or absent, that is stated as a gap. Knowing you are about to make an assumption is most of what makes the assumption safe.

This is not a literature review at speed. It is the reconciliation work that nobody has ever had time to do: reading across sources that use different outcome definitions, different populations, and different professional vocabularies, and saying where they actually agree. It is a genuine research and engineering problem, and it is now approachable.

What it does not do is exercise judgment. Which trials are comparable, which local constraint is binding, which finding will not survive the labor market here: those remain expert questions that demand time and attention. The change is that the expert is looking at the whole relevant body of knowledge rather than the fraction someone could find on deadline.

The second moment: the long middle

Most of a program's life is during implementation, and implementation is where correctable failure happens. The coaching visits that stopped in three districts. The materials that never arrived in the fourth. The enrollment that fell off in one grade. The facility where reported treatment quality dropped after a staffing change. All of these are knowable within weeks. Most are missed.

It is not that governments lack machinery for the middle. In the United States the machinery is statutory. The Government Performance and Results Act of 1993 required agencies to publish strategic plans, annual performance plans, and annual performance reports. The GPRA Modernization Act of 2010 went further, requiring agency priority goals set every two years, a named goal leader for each, quarterly data-driven progress reviews, and a standing judgment about which goals are at risk. Outside the United States the same idea travels under results-based management, and most ministries I have worked in have some version of it.

These systems do something real. They set a cadence, they name an owner, and they put numbers in front of senior officials four times a year. What they do not do is explain. An indicator tells you that a number moved or did not. It does not tell you which assumption broke, whether the cause sits in delivery or in design, or what to change before the next quarter. So the review becomes a status meeting. The goal leader reports the number, explains the variance from experience rather than from evidence, and commits to attention rather than to a change.

The Evidence Act sits on the other side of the same gap. The Foundations for Evidence-Based Policymaking Act of 2018 codified the evaluation side: learning agendas, annual evaluation plans, capacity assessments, and named Evaluation Officers, Chief Data Officers, and Statistical Officials. The statute itself defines evidence and evaluation broadly. OMB's implementing guidance is what named randomized trials and quasi-experimental designs as the designs for impact evaluation in federal practice. But an evaluation commissioned under an annual evaluation plan reports after the period it studied. It is rigorous and it is late.

So the long middle is bracketed by two statutory systems that do not meet. Performance management is fast, frequent, and shallow. Evaluation is deep, credible, and slow. Neither was built to say, in month fourteen of a five-year program, that something is going wrong and here is the likely reason. Adaptive management is the practice meant to fill that space, and in most agencies it does not. There is no forecast for the program to deviate from, so a divergence has nothing to be measured against. The assumptions underlying the program may have been analyzed at design but are not monitored during implementation closely enough to support course correction. And changing course reads as an admission that the original design was wrong, which nobody is rewarded for saying.

What changes this is not a better dashboard. It is a design in which the program carries an explicit forecast, meaning what is expected to change, by how much, by when, and for whom, with uncertainty attached and the load-bearing assumptions named, together with systems that continuously compare what is happening against it. When delivery data, statistical indicators, and user reports stop matching what was expected, that divergence is raised, along with what changed, which assumptions it bears on, what the plausible explanations are, and what evidence would distinguish between them.

This is where agentic systems earn their place. They do not decide. They do the continuous work of watching, reconciling, and asking. The properties that make them safe here are the same ones that make them useful. Every alert traces to the data that produced it. Every recommendation names the assumption it rests on. The system proposes, and a named official decides and is accountable for deciding. Because the program was forecast with uncertainty in the first place, a deviation is an expected event rather than an admission of failure. That is what makes correcting course a normal act rather than a career risk.

This is what the quarterly review was meant to be. GPRA established the cadence. The Evidence Act established the standards. What has never existed is a working signal layer between the quarterly number and the retrospective evaluation. That layer is now buildable, and building it requires no new statute.

User feedback is not a supplement to this. It is one of the two signals that make it work. Delivery data tells you what the system did. User reports tell you what it was like to be on the receiving end, and they frequently move first. Attendance can hold up for a school term after the reason for attending has quietly disappeared.

From efficiency to outcomes

Almost all public sector AI today makes government more efficient: faster processing, lower cost per transaction, fewer staff hours, shorter queues. Those are real gains. But efficiency measures how the institution runs inside. It says nothing about whether the institution is doing the right things. What I am proposing is a shift from making government run better to making government produce better outcomes.

Efficiency asks	Outcomes ask
Can we process claims faster?	Are people getting the benefit they are owed, sooner?
Did we obligate the funds on schedule?	Did the intended households actually get help, on time?
Are our systems integrated?	Are more students finishing school having actually learned?

Efficiency asks	Outcomes ask
Did we hit the output targets?	Are participants earning more a year later, and which ones?
Is the caseload falling?	Are patients recovering, and is homelessness falling?
Are we compliant?	Do citizens have reason to think government works for them?

The right-hand column is harder to measure, slower to move, and less flattering in a report. It is also the only column that matters to the people paying for all of it. An institution that gets very good at the left-hand column while the right-hand column stays flat has become efficient at doing something that does not work.

What has to be built for better public outcomes

None of what follows is a big reveal. All of it is underbuilt.

Identifying and synthesizing what is known. Finding everything relevant to a decision across the full range of sources, on the three properties set out above. This is not summarizing. It is the harder job of reconciling sources that disagree and saying why.

Forecasting and continuous course correction. Estimating what is likely to happen, before the commitment and continuously after it, with uncertainty attached, with assumptions stated rather than buried, and with early warning when a program drifts off the expected path. A forecast without uncertainty is not a forecast.

Continuous feedback from service users. Knowing what the people using a service are actually experiencing, at the speed of operations rather than the speed of evaluations, disaggregated enough to act on, and under the conditions set out in Part Two. This is what most separates what I am describing from ordinary analytics, and it is the piece most often left out.

Shared learning. Contributing to and drawing on the shared reservoir, so that every project, evaluation, implementation lesson, and user report leaves the next decision better informed. This is a network capability, so the early years mean giving before getting. That is a familiar public goods problem and it will not solve itself.

Artificial intelligence sits underneath all of it. It makes identification and synthesis workable at scale, forecasting continuous rather than occasional, and user feedback analyzable rather than merely voluminous. It is decisive, and it is not the point.

What this asks of us

Three groups have three different obligations.

Of governments

The main thing to ask of public institutions is not that they buy technology. It is that they change what counts as a well-made decision. A decision made without reference to available evidence, without stated assumptions, without a forecast, and without a plan for spotting error should be treated as a weak decision, the way an agency would treat a budget submission with no numbers in it.

In practice that means recording decisions and the reasoning behind them, so someone can learn from them later. Saying in advance what you expect to happen, so you can be surprised. Building substantive and frequent feedback from service users into design and delivery rather than into reports about delivery. And protecting analysts from pressure to confirm what has already been announced. An institution that learns has to be able to find out it was wrong, which means leaders who can survive finding out.

It also means valuing and using the evidence you already own. Making that material findable, structured, and reusable may be cheaper than any new study and may change more decisions. And it means capacity. None of this works if the people who design and deliver programs have not been trained to read evidence, build a theory of change based on evidence or question a forecast. I have spent much of my career on that training, and the lesson is that it does not stick as a workshop at the start of a project. It sticks when it is part of how an institution hires, promotes, and supervises, and when the institution's underlying norms and incentives are aligned with the skill being built.

In the United States, most of the structure for this already exists. GPRA and the Evidence Act between them require the plans, goals, reviews, and officials described in Part Three. What is missing is the connection between them. The learning agenda rarely shapes the design decision taken six months later. The evaluation plan rarely determines what gets funded. The performance review rarely changes course while there is still time. I am not asking for new authority. I am asking that the existing authorities do more work. The same is true in states and cities, and in ministries with outcome budgeting frameworks that gave people the vocabulary of a learning government without the capability.

Of development institutions

Multilateral banks, bilateral agencies, UN organizations, foundations, and larger NGOs are in an odd position. They hold the largest collection of evaluation and implementation knowledge

anywhere, and they are structurally bad at using it. Lessons from the portfolio rarely reach individual task teams in time. Completion reports get written for accountability rather than for learning.

I have four asks. Make the knowledge usable, by investing in the standards, structure, and retrieval that turn archives into inputs. Value the whole portfolio rather than only its rigorous tier, because the implementation record holds transferable operational knowledge that is routinely undervalued. Make it shared, by contributing to the common store rather than to internal repositories only insiders can search. And tie incentives to outcomes rather than commitments, so that the institution is aligned with the advice it gives its clients.

Of the AI field

Model developers, funders of AI for public benefit, and researchers face a choice about where to put their effort. Right now most of it goes to general capability and commercial uses, with government treated as a deployment problem downstream.

Treat evidence identification and synthesis as a real research and engineering problem. Tracking where a claim came from, grading source quality, reasoning across sources that disagree, and checking whether forecasts hold up are hard problems worth serious attention. Build agentic systems for the conditions public programs actually operate in: intermittent and dirty data, causal lags measured in months, and a requirement that every alert be traceable and every recommendation contestable. Build for what public decisions require, which is not what commercial ones require, since a public decision has to be explainable to the person it affects, auditable by an oversight body, and defensible in a legislature. And resist selling autonomy where augmentation is what is needed and what is right. Public authority should stay with people who can be held accountable, and systems should be built so that is real rather than claimed.

What I am not saying

Arguments like this get misread in predictable ways, so let me be explicit about the limits.

- **Public servants are not ignoring evidence.** The failure is infrastructural, not attitudinal. Any version of this argument that reads as a complaint about officials has misunderstood the problem.
- **None of this displaces rigorous evaluation.** Trials and strong quasi-experimental designs remain the best way to learn what a program caused. We need more of them, and better timed. My claim is that rigorous impact evidence is one kind of knowledge among several, not that it matters less.
- **Algorithms should not make public decisions.** They should inform them. Authority comes from accountability, and no system now or soon can be held accountable in the

way that matters. Every consequential recommendation needs a named person who owns the decision and can explain it.

- **Evidence does not settle questions of value.** It can tell you what a program is likely to do. It cannot tell you whose interests should win when they conflict. Those are political and moral questions and they belong to political and moral processes.
- **More data is not the answer.** Much of what gets collected now is unused, unusable, or misleading. Integration, quality, and usability come first, not volume.
- **Technology will not fix institutions.** Capacity, incentives, politics, norms, and leadership decide whether any of this takes hold. Technology can make the right thing cheaper to do. It cannot supply the will to do it.
- **This is not risk-free.** Systems built on pooled data about citizens can be turned into surveillance. Forecasting systems can lock in past inequities. Automation can hollow out the judgment it was meant to support. These risks are real, not hypothetical, and the people building these systems have to address them.

PART FIVE

Ten years out

It helps to be concrete about what success would look like, because a paper on public outcomes would be hypocritical without it. Ten years from now, in an institution that took the argument seriously, several things would be different.

A federal program officer, or a country team task manager, preparing a design decision would start from an assembled picture rather than a blank page, and would be able to question it: which findings come from places like this one, how much do the studies disagree, and what would have to be true for this to work here.

The design would carry a forecast. What is expected to change, by how much, by when, and for whom. Uncertainty would be stated honestly. The assumptions doing the most work would be identified, so that someone can watch them.

During implementation, the institution would know within weeks, not years, that something had gone off track. Not because a supervisor reported it, but because the delivery data and user feedback stopped matching what was expected. Correcting course would be normal and unembarrassing.

People using services would be able to see what was reported about them and what was done in response. That closes a social accountability loop that has been open for a long time.

And the lessons from all of it, the successes, the failures, and the much more common ambiguous middle, would go into the shared reservoir, where the next official facing a similar decision could find them. The reservoirs would no longer be walled off from each other. Together they would form an ocean of usable knowledge.

How we would know

First and foremost, we need to measure changes in the outcomes: movement in the outcomes themselves, in completion, income, health, safety, inclusion, and trust, attributable in large part to better decisions at design and during delivery.

Second, we track the leading indicators of better outcomes:

- Decisions that visibly changed because of evidence that would not otherwise have reached them. Measured, not assumed.
- Less time between a problem appearing in delivery and the institution addressing it.
- Better forecasts. Predictions that turn out right about as often as they say they would.
- Programming corrected during implementation rather than finished and then judged ineffective.
- Findings from service users reaching decision-makers and visibly changing something.
- Lessons reused across institutions and countries, at a rate you can count.

Third, we employ some well-placed evaluations, with a strong utilization focus:

- An eventual RCT or quasi-experimental evaluation to test causality of this hypothesis.
- A performance evaluation or some case studies to see how this is unfolding from the perspective of public officials who use the new systems and techniques suggested above.

This should all sound familiar now. Again, the bottom line: if public outcomes don't move, this did not work.

CLOSING

The measure

We have never known more about what improves people's lives. We have never had better tools to find, analyze, forecast, and share what we know.

And millions of public decisions still get made without the best available evidence, and without the voices of the people they affect. The evidence exists and does not get reached. The statistics exist and do not get used. The voices exist and do not get heard. The lessons exist and do not

get carried forward. That is not a failure of intention. It is a failure of infrastructure and institutional incentives. The former can be built. The latter, with great effort, can be changed.

So the task ahead is not only to build more capable artificial intelligence. It is to build institutions that learn faster, decide better, and improve lives more effectively. It is a difficult problem, and the solution is finally within reach.

The future of AI for public good will not be measured by the intelligence of machines. It will be measured by the wellbeing of people.

A note on what this is

This is written from one vantage point, my own, and that of the programs I have worked on directly and through my sister enterprise, Social Impact, Inc. While it is a good vantage for seeing some things it's weak for seeing others and I am sure parts of this paper are wrong.

Let me know what you think, and where you may be headed with this.

Rolf Sartorius

Founder & CEO, Saipion, Inc.

Chairman & Co-Founder, Social Impact, Inc.

rolf.sartorius@saipion.ai · saipion.ai/insights/public-outcomes

A NOTE ON PREPARATION

I drafted this paper with AI assistance, Claude (Anthropic) and ChatGPT (OpenAI), working from Saipion's methodological work and from my own practice in international public management. The tools helped with drafting, structuring, and revision. The argument, the judgments, the examples, and any errors are mine.

A paper arguing for knowing where things came from, and for augmented rather than autonomous intelligence, should be willing to say how it was made.

