After the Manifesto: Learning to Think Honestly About Outcomes

Public Essay · Protopia Garden

Tetiana Honcharenko Alexey Konstantinov

Protopia Garden · June 2026

Where to begin this conversation

In the manifesto we described an attempt to assemble an organization of a new type — one able to work on hard shared problems not worse, but better, than the old institutions manage. We said back then that we were half a step from the first live launch, and that we were assembling what such an organization needs to stand on: the tools, the rhythms of work, a way to hold memory, to check ourselves, and to act.

This text is an account of the first piece in our current build. We have put together the first element of what could be called the future organization's architecture of thinking — not the part that acts in the field, but the part that thinks. This tool helps us work on the hardest and most neglected question in all of social work. The question sounds simple: **how do we know that our work actually leads to anything?**

To explain what we built and why, we first have to work calmly through why this simple question is in fact so hard, and why the sector that exists to help society answers it worse and worse.

How organizations usually think about outcomes

Picture an ordinary charitable or civic organization. It wants to do something good: help children learn better, support people after a disaster, breathe life back into a fading town. To get money for this, it writes a grant proposal. Sometimes the proposal even has a section that, in professional language, is called a "theory of change."

A theory of change, put very simply, is a story about how our actions will lead to a good outcome. It runs roughly like this: "if we run training for teachers, the teachers will teach better, and then the children will learn better." A logical chain: do this — get that. A budget is built around it, then activities are carried out, then a report is written saying the activities were carried out.

There is nothing stupid in this scheme. For many years it worked as well as it could. The problem is that a beautiful story about a future outcome and the outcome itself are very different things. The story is smooth, logical, pleasant for a donor to read. Reality is complicated and uneven. Teachers can go through training and change nothing in the classroom. Children can start learning worse for reasons that have nothing to do with school. A town can fade not because it has too few jobs, but because people have stopped believing this place has any future — and then no number of jobs will help.

Here is the most unpleasant part. Writing a convincing theory of change is now very easy. Any language model will compose a smooth, logical chain for you in a minute. And so the smoothness of the text no longer tells you anything about whether the organization actually understands what it is doing. A well-written proposal used to be at least some sign that a living thought stood behind it. Today that is no longer so. We wrote about this in detail in the manifesto: the document has stopped being a reliable carrier of understanding.

And here the thing we consider central and most neglected in this whole story comes into view.

The hardest place, which everyone skips past

When accessible artificial intelligence appeared, the sector rushed to automate exactly what was weak to begin with: writing proposals, reports, presentations. That is, to speed up the production of those same smooth documents that were already detached from reality. This is humanly understandable — it is the most visible, most tiring routine. But it is treating the symptom.

The genuinely hard place was left untouched. It is not the writing of documents. It is **honest thinking about the outcome before you have started acting**. Thinking about cause and effect — about what actually

leads to what — turns out to be very hard. It demands admitting that you do not know a great deal. It demands holding several explanations of what is happening in your head at once and not clutching at the first convenient one. It demands checking your guess against what humankind has already worked out on the matter. And it demands the courage to say out loud: here we are not sure, and that we are simply not entitled to claim yet.

Most organizations feel deeply uncomfortable in this place. And, frankly, they have almost neither the time nor the tools for it. A small team under operational pressure physically cannot sit down and compare its experience with the world's body of research. So this place gets skipped past. It gets filled with boilerplate phrases, and everyone involved, out of inertia, pretends that understanding stands behind the phrases.

From this comes our main thought, and it is simple. An ordinary organization starts with the proposal, the budget, and a beautiful story about the future outcome. We believe one should start earlier, and from the other end. First — work through, seriously and in detail, the very thing the initiative is trying to change. Understand how it is built, who acts within it, what forces are at play, what can change in it at all and what cannot. And only then, on the basis of this understanding, talk about the intervention, the plan, and the outcome. We are not speeding up the old machine. We are building a different first step.

What exactly we assembled — and why it is not just a program

Before describing how it works, it is important to say what it even is. We did not assemble a program that does people's work for them, nor yet another "smart assistant" writing texts for you. We assembled the architecture of what the manifesto called joining a human and an artificial intelligence into a single shared act of thinking.

It can be explained like this. A human and a machine have different strengths. The machine can hold an enormous amount in memory, find its way quickly through a large body of research, hold several versions of what is happening at once and compare them, notice contradictions, never tire. A human can do something quite different: understand living context, sense where words diverge from deeds, carry responsibility, make a decision where there is no ready right answer, see a person in the other rather than a row of data. Join them wisely and you get thinking stronger than either alone. Not a human instead of a machine, not a machine instead of a human, but a single working circuit that simply has more memory, more attention, and more honesty.

So when, further on in this text, we say "machine" or "system," we mean precisely this joining, not a self-standing program acting apart from people. The machine part here is not the master but a supporting participant: it prepares, searches, compares, prompts and, importantly, does not let the human quietly cut a corner. But it is always the human who chooses, decides, and answers for it. The machine proposes — the human decides. We hold this boundary firmly: there is work where the machine's help is fitting, and there are places where the disappearance of the human destroys the very heart of the matter — the encounter with another's pain, a promise, a political decision, responsibility for a mistake.

And within this joining the machine part is built in a particular way. Not to make the human's life easier, but the opposite — to keep the human from deceiving themselves. What follows is about exactly how.

What "working out what you are changing" means

Here we need to explain one concept, because it is central for us. We call what an organization is setting out to change the "object of intervention." It sounds dry, but there is no help for it — this is an academic approach, the methodology of building scientific knowledge.

The usual NGO practice: we look at some already-formulated problem through the lens of possible activities, run through them, and try to make the logic of action hang together.

We insist that one must first look at reality itself: what this community is, what this situation is, what this trouble is — in all its complexity, with its history, its conflicts, with its own life that ran before us and will run after us. A community is not an empty object that we "process." It has its own understanding of itself, and it has the right to argue with our understanding.

Take a simple, recognizable example. Picture a small town that young people have been leaving for years: enterprises close, schools empty out, and people little by little lose faith that the place has any future at all. This is exactly the kind of hard, years-in-the-making problem for which social work exists. (We work through this same example in detail in the manifesto — but here it is needed simply as a vivid illustration, and you do not have to read the manifesto for that.)

The old project arrives in such a town with a ready-made frame: "we will develop employment." It gathers data to fit that frame, runs activities, files reports. Perhaps there will be some benefit. But it may not touch the knot of the problem at all, because it was looking in the wrong direction: the real trouble might have been not a lack of jobs but, for example, that people long ago stopped believing any promises.

This is where our joining of human and machine works differently from the usual project. It does not let you pick a single frame straight away. Instead the machine part helps the human lay several different views of what is even happening here side by side. Maybe it is about employment. Or maybe it is about accumulated distrust: people have stopped believing promises, figures, strangers who come, measure, promise, and leave. Or maybe it is that the young see no future here, and that already feeds on itself. These are different explanations, and they lead to entirely different actions. The system keeps them all alive and does not let them collapse to one convenient version — until the human consciously chooses a working version and writes down why exactly this one. To return to the dry language of scientific methodology, we are building the object and the frames of inquiry through which we study it.

Then, for each explanation, the machine part asks the human an uncomfortable question: what is it even based on? Is this our guess, or is it confirmed by serious research, or did it work somewhere but in a completely different context, so it is unknown whether it will work here? To help answer, it assists in turning to the world's body of scholarly work and prepares, for a team that has no scholar of its own, a rough marking-up: this idea seems strong and tested; this one is plausible but contested; this one is fashionable but weak; and this one worked in Kenya but most likely does not apply to our town. The last word in this marking-up stays with the human.

And, perhaps most important: the system requires keeping **the map of the unknown**. This is an honest list of what we do not know about this situation, where our model is weak or empty. In an ordinary proposal the zones of not-knowing are absent — one simply does not know of them, because not-knowing itself looks like a weakness. With us it is the reverse: a model that has no such map simply does not pass further. Admitting "here we do not understand" is not a disgrace, it is the condition for being able to move on at all in the layer of thinking about the problem and the options for impact.

Where the boundary runs that everyone blurs

There is one more distinction, and it is essentially what another part of our system is assembled for. The sector constantly merges into one two completely different things: "we have a good model" and "the outcome actually happened." This is the gap between theory and practice: the theory in itself may be correct or mistaken, but its connection to practice, to reality, is a separate object of care.

These are not the same. You can build a beautiful, well-grounded model of how training for teachers will improve children's learning — and in reality the children may still not learn any better. A good model is about the quality of thinking. An outcome that happened is about what actually changed in the world. Between them lies all the unpredictability of real life.

So our machine keeps these two questions strictly separate and never lets one pretend to be the other. There is separate proof that the model is good. And there is separate — far harder — proof that the change actually happened in the field. And until the second is there, the organization is not allowed to declare success loudly. A weak, unverified claim can be kept for internal learning or for a frank interim conversation with a donor — but it must not be turned into a public victory, still less into a basis for money. With us this rule is not a wish but a strictly checkable prohibition.

Why entrust this discipline to a machine

One might ask: if everything rests on honesty of thinking and on human judgment, what does a computer have to do with it at all?

This: it is almost impossible to hold on willpower alone — especially when the team is tired and the donor needs a report by Friday. The temptation to cut a corner is too great: to skip the map of the unknown, to pass off a guess as proven, to call a chance coincidence the result of your work. Of their own good will people promise themselves not to do this — and under pressure they do it. So we turned this discipline into a set of rules that the machine checks.

You can picture this as a built-in "checker." It does not check whether the organization is right on the merits — that is not its business and not within its competence. It checks something else: whether corners have been cut. Whether the map of the unknown has been filled in. Whether the alternative explanations are still being kept alive. Whether a coincidence is being passed off as an effect. Whether more has been claimed than proven. If corners are cut, it simply does not pass the work further — not as a boss, but as a rule, the same for everyone. This way honesty stops depending only on the good will of a tired human and becomes a built-in property of the process itself.

But the "checker" is only one side. To put the main thing plainly, the machine here does two things for us. The first — it helps discipline thinking and hold on to its complexity. The human mind cannot hold a dozen explanations, hundreds of studies, and the whole history of past decisions all at once — it inevitably simplifies, boils everything down to one convenient picture. The machine takes this weight onto itself: it remembers, compares, does not lose the thread — and because of that we can afford to think about the problem in a more complex way than alone, without collapsing into a flat version just because it is easier to hold. The second — it helps to see and get around the cognitive fallacies that all of us, without exception, are prone to: the pull to clutch at the very first version, to take what we wish for as proven, to notice only what confirms our being right and not notice the rest. The machine is no smarter than us in this — but it is dispassionate and does not tire, and so it is able to point out in time: look, here you seem to be giving in to the familiar mistake. It does not decide for the human — it gives the human back the chance to see themselves from the outside.

How it happens: a conversation between human and machine

So that none of this sounds frightening, let us describe how it looks in practice. There is no console with toggles and no complicated questionnaires here.

A person comes with ordinary words — the way a normal person describes a situation that has been hurting: "young people in our district are drinking their lives away, and we don't know what to do about it"; "after

the flood, people aren't coming back to the village." Not in professional language, not as a theory of change — just a living account.

Then the conversation begins. You can speak aloud or type, you can attach documents — reports, notes, interview transcripts, whatever is to hand. The machine listens and is in no hurry to hand over a ready answer. It asks clarifying questions — exactly the ones that in the rush nobody usually asks: and who exactly does this affect? and what have you already tried? and what do the people themselves say? Out of this conversation it gradually assembles structured records — we call them cards: a separate view of the problem, a separate causal link, a separate claim, with a note of where it is a guess and where there is a ground for it. When it needs to check something, it goes to the scholarly literature and the internet itself, brings back what it found, double-checks itself, and openly shows where it found no confirmation.

What you get is not an interrogation or an exam, but more like working with a very well-read, tireless, and meticulous assistant: it has read almost everything, forgets nothing, and gently keeps you from slipping past the hard place. But everything it has gathered stays a draft until the human has looked at it, argued with it, and approved it. So, step by step — from a living account to a tested model — that object we have been speaking of is born.

What it is built from: three linked nodes

If you look inside, all of this rests on three closely linked parts. Let us name them plainly — so it is clear we mean a concrete, assembled tool, not some hazy "artificial mind."

The first node is the shared knowledge base. Since human and machine work in tandem, they need a shared memory that both people and the machine part can access at the same time. This is not a folder of files or an archive of reports, but a living, organized store: past models, gathered evidence, signals from the field, decisions made, and the reasons they were later revised all go in here. The main property of such a memory is that it does not leave with a departing employee and does not start from a blank page every time. The organization stops reinventing itself with every change of team, and its experience becomes not a heap of folders but a working resource that both human and machine can look into at any moment.

The second node — the model of the object of intervention — lives inside this knowledge base. This is the heart of the whole construction — that very model of the object we spoke of: a **machine-readable object** made of structured records, the cards — a separate one for each view (frame) of the problem, for each causal mechanism, for each piece of evidence, for each hypothesis. Such records are read by both human and program. The model is versioned like program code: its whole history is visible — what changed, who added it, why. From it you can make different "views" for different tasks — a causal diagram, a map of the actors, the familiar theory of change for the donor's language — but no such view is passed off as the single truth. Attached to the model is **the academic layer**: a link to the world's open catalogues of scholarly work (for example, OpenAlex and Semantic Scholar). When you need to check what an explanation rests on, the machine helps you turn to this literature — it searches for works, where possible checks whether an article has been retracted, finds summarizing reviews, prepares a draft evidence card, and openly marks what it did not find. And the judgment — whether this research fits our case, how soundly it is done — stays with the human. And over all of it works that same "checker": a small separate program (in our language, a linter) that runs through the records and checks whether the rules are kept; if corners are cut, it does not pass the work further.

The third node is the impact-mandate layer. If the first part answers the question "do we understand the object well enough to have any right to act at all?", then this node answers the next one: "what are we entitled to claim about the outcome — and what has been promised to whom?" Put simply, an impact mandate is

an open and spelled-out-in-advance agreement on what exactly is to change in reality, over what horizon, under what constraints, who is entitled to set such a goal, and how we will know that the change has actually happened.

This agreement has several important properties, and they are not accidental. It can be born from two sides: from a donor who sets the goal and does not pick the doer in advance (with an open selection of those who will take it on to follow) — or from the field side, when an initiative comes with a real problem and a model. We deliberately permit the second path: if you allow goals to be set only from the side of money, the system builds injustice into itself from the very start. Defenses against such injustice are written directly into the mandate: what the community's role is (co-author, consulted, merely informed), whether it has given its consent, whether it has the right to challenge the mandate before it is even launched, and who controls the data and the rights to the outcome. In sensitive situations open consent may be unsafe — then the mandate goes in a "narrow" mode, without loud public claims.

And the main thing — here lives that very gap between "the model is good" and "the outcome happened" spoken of above. Out of it grows a hard rule: no claim about an outcome can become public, financially significant, or turned into any kind of asset without both proofs — the model proof and the field proof — and without a separate decision to verify. Weak or one-sided evidence is fit at most for internal learning and a calm interim conversation with a donor, but not for a loud victory. There are no exceptions.

Together these three nodes form that very architecture of thinking: a shared memory in which a disciplined model of reality lives, and a grammar that does not let you leap from a beautiful model to a significant claim about the outcome without proof and without the right to make it.

Let us return to the question we began with: how do we know that our work leads to anything? An honest answer begins not with a loud "we did it," but with this discipline — understand the object, hold on to doubt, separate the model from the outcome, and claim no more than has been proven.

The model is assembled — what to do with it next

The model is not in itself a goal or a one-off trinket. It is a working tool, and it is used in different ways depending on where you are — from the very first vague questions to a project already underway. And, importantly, it does not freeze: it lives and changes along with the work.

The earliest and most important moment is the strategic one, when there is no project at all yet. A team or initiative has only just come together and wants to take on some problem area or a whole tangle of one particular community's problems. In essence, it has strategic questions to decide: which goals to choose, what to focus on, which steps to take in what order, and where to direct the resources that are always limited and insufficient. It is precisely here, at the strategic level, that working with the model makes a critically important contribution — because everything else depends on these early choices. The architecture works as a space for shared inquiry: in conversation with the machine, the team begins to work out how these problems are actually built — it singles out the important contexts and structural nodes, untangles the snarled knot into its parts, looks for the points where the situation can be influenced at all — and not on a single level, but on different ones, from the individual person to the institution and the surrounding environment. Layer by layer, as the team's own understanding grows, a whole set of possible interventions takes shape — candidates for future projects.

And then, out of this set, you can assemble not a scatter of disconnected activities but a coherent strategy: where each action has a clear rationale, where you can see how it is linked to the other plans and points of impact, where possible side effects have been thought through, and where observation of outcomes, set up in advance, also serves the team's own learning. And all of this — with a sober understanding that there are

always fewer resources than one would like, and they must be directed consciously, to where they will do the most.

The same early work is needed from the other side too — from the side of a large, systemic donor, and especially before it rolls out its programs. Before allocating funds, such a donor can use the same discipline to work through the problem field, see where the points of impact lie at different levels, and assemble not a set of disconnected grants but a coherent program — with thought-out links between its parts, expected effects, and built-in observation of outcomes. Then the money is directed consciously from the start, rather than being fitted to already-announced priorities.

When a concrete project grows out of such inquiry, the model gives what is usually missing: clear, well-grounded hypotheses and a truthful causal picture of what you are acting upon. From it a project proposal is assembled easily and quickly — but no longer an invented, smooth theory of change; rather a proposal with a tested model, evidence, and an open map of the unknown standing behind it. The project comes out honest: to the donor you show not a guess but a disciplined picture — including what you do not yet know, and how you intend to find it out.

And what is more — since the object and the proposed interventions are drawn out without embellishment, the model also makes the scope of needed resources clearer: how much of what the work will really demand is visible not from fitting to a budget line but from the very picture of the problem. This is not a precise estimate, but a far more sober footing than the usual eyeball guess. And, unexpectedly, such a proposal is also more convincing than an ordinary one: you can see that work of thought stands behind the words, not just a well-turned text.

If you are already in implementation, the model works differently. It lets you check your versions against what is really happening in the field: what is confirmed and what is not; where the context needs to be widened; what new hypotheses life itself suggests. The indicators from the model become what you track in practice, and a divergence between the model and the data is not an inconvenience to be hidden in a report but a signal to revise. So action continually returns into thinking: the field teaches the model, and the refined model leads the next action.

And here is the main thing: this is not a one-off assembly at the start, after which the model is laid on a shelf. It is a living object that changes, is added to, and is corrected by feedback over the whole life of the project. It has versions and a history — you can see how the understanding changed and why. When reality argues with the model, the model is updated; when the team has learned something, it stays in the shared memory and not only in the head of a person who may leave tomorrow. The same discipline that assembled the model keeps it honest as it changes. And openly revising your own picture here is a sign of strength, not of failure.

What it guards against

The easiest way to explain the value is through the concrete traps that even clever and conscientious people fall into. Let us keep our example in front of us — the town that young people are leaving — and look at how the same mistakes repeat again and again. They sort conveniently into several families: distortions of thinking itself, errors of causal models and of hypothesis testing, the difficulty of measuring real effect, errors of documents and language models, and errors of how we relate to people. Our system is built so as to hold the team back from each one.

Let us start with **distortions of thinking itself** — with the makeup of our own minds. Each of us has built-in mental shortcuts: in ordinary life they help us out, but in hard work they quietly lead us astray. Under time pressure they fire unnoticed, and the trouble is that from the inside the mistake feels like clear understanding.

Anchoring on the first version. The commonest trap of the mind is to clutch at the first explanation that comes and never let it go again. The team decides: "the town is dying because there is no work," — and after that everything turns around jobs, while the real knot (for example, that people long ago stopped believing any newcomers and promises) stays unnoticed. The system gets in the way of this: it keeps several explanations alive side by side and demands that you consciously choose a working one, writing down why exactly it and what we are giving up.

Confirmation bias. We tend to notice what confirms our being right and to skip what gets in its way; to remember our successes and forget our failures; and to gladly chalk up to ourselves any good that happened nearby. This works especially strongly once we have already invested in an idea. Having come to believe the town needs jobs, we start marking every newly opened workshop as confirmation of our being right — and not noticing that young families are packing their things all the same. So the system requires, from the very start, keeping a map of the unknown — an honest list of what we do not understand — and keeping uncomfortable explanations alongside. A smooth picture in which everything adds up is most often the first sign that we have given in to a distortion.

The next family is **errors of causal models and of hypothesis testing**. Thinking about cause and effect is a separate and genuinely hard skill, and here even experienced people make mistakes.

Correlation passed off as causation. After our program several young families returned to the town — so it was us? Not necessarily: maybe a plant reopened nearby and they would have returned anyway. Many causes are always at work alongside, and a simple coincidence in time proves nothing in itself. The system requires you to mark plainly which rung you are on: did you merely notice a connection — or did you actually show that one thing caused another.

A hypothesis that cannot be tested. "We are reviving trust in the town" sounds good — but if you do not say in advance by which observations exactly we will know that trust is NOT growing, such a claim can be neither confirmed nor refuted. And an unfalsifiable hypothesis is no longer a working model but a faith. So for each causal link the system asks you to name in advance which observation would refute it, and to show what evidence it rests on.

Naively glued-together causal chains. In our town this looks like: "our meetings bring neighbors closer" → "closeness brings back trust" → "trust keeps the young from leaving" → therefore our meetings will keep the young from leaving. Each link on its own sounds reasonable, but each is taken from a different source of differing reliability, and the conclusion drawn from their splicing falls apart. The system does not splice such chains silently: it shows where each link is taken from and how far it can be trusted, and remembers that the conclusion is no stronger than the weakest link in it.

Next — **the difficulty of measuring real effect**. Suppose the model is good and the link is really causal. The hardest part remains — and the most often faked: to prove that the change really happened and that it is our doing.

"What would have happened without us?" Suppose the outflow of young people from the town has slowed. But maybe it would have slowed anyway — because of a general turn in the region? The central question of measuring impact is a question about the invisible: what would have happened if we had not been here? Without an answer to it, any "we did it" hangs in the air. And answering it conscientiously is almost always hard: there is nothing to compare with, a control group cannot always be assembled, and life sets no clean experiments. So the system requires you to distinguish "we contributed alongside others" from "we caused the outcome" — and not to pass off the first as the second.

False precision. Out of this difficulty is born the temptation to hide it behind a single beautiful number — to reduce the revival of a whole town to the line "we created so-many million worth of social impact."

The number looks solid, but enormous uncertainty usually hides behind it, and a single figure creates a false sense that we know for sure. More truthful is a conversation about which paths of events are possible and what each is worth. The system does not let you reduce the complex to a deceptively precise figure.

Activity mistaken for outcome. The easiest thing is to report what we did: "we held twelve meetings with the town's residents, gathered three hundred people." But activities carried out are not yet a changed reality. To speak of an outcome, you have to show that something observable and verifiable changed: for example, that young families really did start staying, rather than just sitting through our meetings. And even then the question remains whether it is our doing. The system makes you name plainly what exactly we are claiming: did we merely carry out an activity, notice a change nearby — or did we actually show that the observed change was caused by our intervention.

Another family — **errors of documents and language models**. This is that very ailment of the collapsed document layer with which the conversation began — only now amplified by artificial intelligence.

Smoothness instead of understanding. The habitual moves of reports and proposals are arranged so that a beautifully written text is taken as a sign of good work. And now, with language models, in a minute you can get a flawlessly fluent theory of change about our town — convincing and at the same time never once checked against what is actually happening there. So our system ties each claim not to the beauty of the wording but to a verifiable record: you can see what it rests on, and whether it rests on anything at all.

Invented scholarliness. Language models have a dangerous property: they can confidently make things up — invent a nonexistent citation, attribute to a real study a conclusion that is not in it, pass off a plausible phrase as an established fact. Lean on this and you get scholarly-looking emptiness — for example, a confident reference to "a study that proved employment programs revive small towns," which either does not exist, or is about completely different towns, or was retracted long ago. Our academic layer works differently: it turns to real scholarly databases, where it can it checks whether an article has been retracted, and it conscientiously records what it found and what it did not — so authority is harder to fake, and a gap harder to pass over in silence.

Fashion and outdated knowledge. Today the fashionable idea is that small towns are revived by co-working spaces and the "creative class" — but fashion does not prove rightness; and data on the causes of outflow that were true for our town ten years ago may by now be outdated. The fact that an idea is often cited helps you find it, but does not make it true. The system distinguishes "an idea is little mentioned because it is weak" from "because it is new," marks the freshness of evidence, and requires it to be reassessed.

And finally — **errors of power and of how we relate to people**. They are the most dangerous, because they often look like a virtue.

The community as a source of data, not as a co-author. The people whom the problem touches know things about their own life that no visiting expert sees. In our town the real cause of the decline — say, a long history of broken promises that all the locals remember — is truly known only to the residents themselves; without them the model simply will not see it. Their voice is not "data" that we collect but a source of understanding, without which the model will almost certainly be incomplete. And it is precisely this voice that is usually heard the least.

Most often this happens in one of two ways. Either the community is turned into a source of information: strangers come, gather data, leave, and return with a ready-made verdict about someone else's life. Or "consulting the community" turns out to be a ritual: people were heard out — and the decision was still made our own way, on our own terms. In both cases people remain an object that is thought about, not a participant who thinks with us.

There is a subtle point of power here too. Even when we sincerely offer the right to object, it is real only when it is also available to those who lack the right language, status, or courage to speak aloud. Otherwise the "chance to challenge" goes only to the most confident, and the most vulnerable voices — the very ones for whose sake it is often all undertaken — are again unheard.

So in our system the people whom the matter touches have the right to be co-authors of the model about themselves: to take part in how it is built and to challenge it before anything is launched, not only after the fact. The model is built not about people but together with people.

The hasty victory. And, finally, the main thing — the temptation to declare "we have revived the town!" already within a year and to raise new money and a loud reputation on it, while nothing has really been proven yet. This is exactly why the system keeps "the model is good" and "the outcome happened" separate and does not allow hope to be turned into a public victory or into an asset until the change has been independently confirmed.

None of these traps is exotic. This is the ordinary physics of how conscientious people under time pressure start describing themselves better and better and seeing the world worse and worse. The machine does not make us smarter or more honest — it simply does not let us quietly slip past these traps and each time returns our attention to them.

What is already assembled — and what we have not yet done

So as not to mix one thing with another, let us separate what we already have from what we do not.

Already assembled and working: the shared memory — the knowledge base; the working circuit that builds and checks the model of the object; the circuit that holds the grammar of honest claims about the outcome; the automatic checking-rules that catch cut corners; and trial runs on training examples. This is not yet proof that we have changed something in the field — but it is no longer a design on paper, it is a working, executable discipline of thinking.

And now about the limits. We have the discipline of building a good model and the grammar of well-checked claims, but as yet no working way to prove that a change really happened in the field. This is a separate, very hard task, and it is a load-bearing wall we have still to build. Until it is there, our system is a discipline of thinking and a groundwork for the future, not a confirmation of real outcomes.

Where this leads: what we are actually heading toward

This architecture of thinking is not a goal in itself. It makes sense only because the world around us has changed and the old form of help cannot keep up with it. The world has become faster, more conflict-ridden, more layered, and longer in its consequences. And the habitual way of working on shared troubles has stayed slow where speed is needed; opaque and vertical where trust is needed; tangled in approvals where one must act; and time and again it arrives late — often after a predictable trouble has already become a catastrophe. And with all that, it most often cannot clearly say whether it changed anything at all.

We spoke about these gaps in detail in the manifesto. Gathered together, the world places several demands that the old form answers poorly: the need to respond quickly (some troubles come not over years but over days); the need for transparency and horizontality (so that help is not obliged to pass through heavy donor structures and the long approvals of states and commissions); and the need to hold on to complexity (to think about a problem at the level it deserves, rather than simplifying it to fit the report's format). All that we are doing is an attempt, where it is possible, to join the efforts of human and machine so as to throw out the dead and the superfluous, not reproduce the old inefficiency, hold on to the real complexity of the task, and

at the same time report honestly on the outcome. The architecture of thinking described here is the piece of precisely this that we have assembled so far.

Where this will spill out further we honestly do not yet know. One of our hypotheses: perhaps over time it will grow into some platform solution — a place where those who want change and those ready to support it find each other directly, build together a verifiable picture of what is happening, and somehow gather resources for it — faster and more transparently than today. Maybe funds could be directed through it, maybe not; maybe new digital infrastructure will be useful for this, and maybe it will not be needed. We do not pretend to know the shape of this solution. It is a distant horizon and as yet only an outline.

To make the direction more vivid, let us picture a situation. Somewhere in Portugal — a forest fire; or a flood, or a sudden influx of refugees. Society is ready to respond, but usually everyone waits for the state, and it deploys slowly, and between the readiness to help and real help a chasm opens in which time and energy die. Now imagine a different course of events: those ready to support respond at once; the problem is quickly run through tested models — not for the sake of a beautiful report but to work out what is actually happening — and is broken down into verifiable tasks; a practitioner on the ground takes on a task, does it, has the outcome confirmed. And the whole chain — from readiness to help to a confirmed real change — does not break in that very chasm between diagnosis and action. We do not claim that we know for sure how to build this. But it is precisely in this direction that we would like to move.

And here you can see why we began with thinking and not with money rails. Any such fast and horizontal mechanism is exactly the place where it is most dangerous to admit a false or poorly grounded claim: speed and automation will only make the mistake bigger and longer-lived. So the first thing to build is not rails for resources but a layer that decides which claims about an outcome are even worthy of having anything move on them. This layer is what we are assembling.

And the digital tools that could carry all this — including smart contracts — are for us a distant and as yet sketch-stage horizon: they make sense only after a sound model, evidence, and verification, and on top of them, not instead of them. In other words, before an outcome becomes a line in a registry, an obligation under a smart contract, or a reason for a payout, it must travel the whole road: the model of the object, the evidentiary check, the confirmation that the change really happened, and a separate decision on verification. This missing piece of the large infrastructure is what we are assembling.

Let us be fully honest: we do not have this platform, its hardest piece — proving that a change really happened — is still ahead, and we are not even sure all of it will come off. But the direction is clear to us, and we have laid a necessary foundation stone — without which everything else would turn into a beautiful but dangerous superstructure over emptiness.

Why we are telling you this

We have assembled an architecture of thinking — and now we need the most important thing: interested potential colleagues and partners. And we will be glad of interlocutors from two different sides.

From the side of the field and implementation — those who work right now as a nonprofit organization or a social-change initiative, who see that the field around them has changed, and who want to rebuild their practice for it. If this architecture of thinking is close to you and you want to reassemble, with its help, your own way of thinking about the problem and about the outcome — we will gladly do it together with you.

From the side of resources and support — donor bodies, foundations, corporate social-responsibility programs — all who want to invest funds in transparent and verifiable ways of working, rather than in beautiful reports, so that funds go to verifiable models and a confirmable outcome.

In both cases we are interested not in those who are looking for “artificial intelligence to write proposals faster,” but in those who feel that very pain with which this text began: the sense that we invest our energy, write smooth reports — but honestly do not know whether we are changing anything at all. If this feeling is familiar to you, we will be glad to get to know you and to share what we have worked out.

Like the manifesto, this text is at once a signal and an invitation. Our tool itself we are ready to show, to take apart, and to discuss — and we will be glad if people try to challenge it and stress it to breaking point. Because a machine that confirms everything and resists nothing is worth nothing.