NAZA: When a Palestinian Becomes a Score — One Error in Every Ten



When we see a house in Gaza after it has been bombed, we usually see the end of the process: a collapsed building, ambulances, the wounded, body parts, families digging through rubble, and new names added to the lists of the dead.

But what happened before the bomb fell?

Who chose the person targeted? Who gathered the information about him? Was his phone analyzed, along with his calls and his movements? Did an algorithm give him a score? How likely was it that this classification was wrong? How many civilians were expected to be inside the house? And who finally made the decision, and in how many seconds?

This is the territory the documentary **NAZA** enters.

In this article, I will try to explain, in language that non-specialists can follow, how

this system works technically, and why its error rate is not just a figure in a performance report, but a question that goes to the heart of international humanitarian law.

What is NAZA?

NAZA is a documentary written and directed by Yuval Abraham and Rachel Szor, the directors of the Oscar-winning *No Other Land*. It premiered at the Venice Film Festival in September 2026, where it won the Special Jury Prize. It was produced by The Guardian and JW Films.

Between 2023 and 2025, the directors interviewed 24 soldiers and officers who worked inside Israel's military and intelligence establishment, concealing the identities of many of them for their protection.

The film does not focus on the moment the bomb falls, but on the system that comes before it: surveillance, artificial intelligence, identifying people, producing targets, calculating expected civilian casualties, and the human review before strikes.

The name **NAZA** itself, according to The Guardian, is a military intelligence term referring to the number of civilians expected to be killed in an airstrike.

That idea alone captures much of the film: civilians are not always a number counted after the bombing. In some cases, their expected number exists **before** it happens.

The Guardian plans to make the film available for free on its website after its theatrical run ends in late November.

Don't picture a robot pressing a button

When people hear "AI in war," some imagine a computer deciding on its own to fire a missile.

What the testimonies and investigations reveal is more complicated. We are looking at a chain of connected systems, what specialists call a **data pipeline**: a production line in which information passes from one stage to the next.

```
Data and surveillance
      ↓
Linking people to devices
      ↓
Pattern analysis (machine learning)
      ↓
Assigning a score
      ↓
Proposing targets
      ↓
Tracking the target
      ↓
Estimating civilians expected to be present
      ↓
Human review
      ↓
Strike
```

Keep this picture in mind, because we will come back to it. The most important question in this article is not only whether one of these stages can make mistakes. It is what happens when several stages in a row make mistakes.

Stage one: the human being as data

Any AI system needs data. In intelligence systems, that data can come from phones, calls, locations, images, relationships between people, movement patterns and existing databases.

Inside the system, a person might look like this:

```
Person
 ├── Name
 ├── Phone
 ├── SIM
 ├── Home
 ├── Location
 ├── Contacts
 ├── Calls
 ├── Family
 └── Movement History
```

Then these records are linked together:

```
Person A ── called ── Person B
   │                     │
 lives in            contacted
   │                     │
House X               Person C
```

This is known as **graph analysis**. In civilian life it is used to detect bank fraud, for example. In intelligence work it is used to map the relationships between people.

To a programmer, these are just entities and relationships in a database. But every node here is a real human being.

And this is where the sensitive questions begin. Does calling someone make you part of his group? Does using a phone that once belonged to someone else make you a suspect? In a small, besieged society like Gaza, where everyone knows everyone, how many people could be "connected" to someone?

Lavender: when a human being becomes a score

In April 2024, +972 Magazine and Local Call published an investigation in which six Israeli intelligence officers said that a system known as **Lavender** was used to rank Palestinians in Gaza by their likelihood of being linked to the military wings of

Hamas or Islamic Jihad.

According to the investigation, the system analyzed information on a large share of Gaza's population and gave each person a rating from 1 to 100 reflecting, according to the model, how likely that person was to be a member of one of the organizations.

Put simply:

```
A person's data
      ↓
Feature extraction
      ↓
Machine learning model
      ↓
Score from 1 to 100
      ↓
Threshold
      ↓
Target candidate
```

What is a feature?

A feature is any piece of information a model relies on to make its decision. According to the investigation, sources spoke of features such as communication patterns, links to known individuals, and changing phones or addresses.

The system learns from people that intelligence had already classified, then searches for other people who "resemble" them.

But think about this: in a war in which most of Gaza's population has been displaced again and again, changing your address is not suspicious behavior. It is everyone's life. And changing your phone may simply mean the old one was lost under the rubble.

> Similarity is not identity. That is one of the most basic truths of machine learning.

A 10% error rate: what does it actually mean?

According to the sources who spoke to +972 and Local Call, the army tested a sample of Lavender's output and found the classification was correct in about **90% of cases**.

That may sound like a good number. But we need to understand exactly what it measures.

It does not mean the system gets one in ten wrong whenever it checks random residents of Gaza. It means that **out of every ten people the system placed on the target list, at least one did not belong to the category that was supposed to be targeted**.

In technical terms, this is called **precision**: of everyone the system labeled a "target," how many actually were?

An example from a medical lab

Imagine a lab tells you your test came back positive, and that its positive results are wrong 10% of the time.

Would you agree to dangerous surgery based on that result alone, without a second test?

Any responsible doctor would order a confirmatory test. That is the bare minimum when the consequences cannot be undone.

In a targeting system, the "surgery" is a bomb. And the question is: was there a real second test? We will come back to this shortly.

Doing the math with real numbers

According to the same investigation, the number of Palestinians Lavender classified as potential targets reached about **37,000 people** at one stage of the war.

We don't need hypothetical examples. We only need to apply the error rate the army itself acknowledged to that figure:

```
37,000 people on the list of potential targets
× 10% error rate (according to the army's own test)
————————————————
≈ 3,700 human beings
```

In other words, about **3,700 people** may have been placed on the target list who were not who the system thought they were.

This does not necessarily mean 3,700 people were killed because of Lavender's errors; we have no data on how many people from the list were actually struck. But it does mean the system was operated with advance knowledge that thousands of innocent people were likely on its list.

And the real error may be larger than 10%

Here is a subtle technical point, but a very important one.

Ninety percent "correct" compared to what?

Every machine learning model is measured against what is called the **ground truth**: the "truth" we treat as correct. One source in the +972 investigation said the definition of a "Hamas operative" used in some of Lavender's training data was broad, and that data on people such as civil defense workers was included in the training.

If that is true, the system could be "90% accurate" by a definition that was loose to begin with. It could be right in labeling someone "linked to Hamas" by that

definition, while that person is not a combatant in the legal sense.

```
A broad, imprecise definition of a target
    ↓
Biased training data
    ↓
A model that learns the bias
    ↓
Wrong classifications at scale
    ↓
Yet the test says: 90% accurate!
```

Programmers call this **garbage in, garbage out**. If the input data is wrong, AI will not turn it into truth. It will make the error faster, more consistent, and larger in scale.

In other words, the error rate by the legal standard, meaning "is this person actually a combatant?", may be higher than 10%, because the ruler the system was measured with is itself crooked.

Who sets the threshold?

If the system gives every person a score from 1 to 100, there has to be a cutoff that decides when a person becomes a target:

```
Score ≥ 90 → Target
Score < 90 → Ignore
```

But why 90? Why not 95, or 99, or 70?

This is not a scientific fact. It is a decision made by people. And every time the threshold is lowered, the list gets longer and the error rate goes up.

According to one source in the +972 investigation, lowering the threshold produced more targets, and he described pressure to produce more of them.

Changing a single number in the system's settings can add thousands of people to the list of suspects.

When the human becomes the bottleneck

In any large technical system, we look for the **bottleneck**: the part that slows everything down.

In the target-production system, the bottleneck was the human being.

The investigation pointed to a book titled *The Human-Machine Team*, published under a pseudonym (Brigadier General Y.S.) and attributed to the commander of Israel's Unit 8200. It argues for using AI to produce large numbers of targets, because humans cannot process that volume of data.

The purpose of automation is usually to remove the bottleneck.

But if the bottleneck is **the person who reviews the evidence before another person is killed**, is removing it an improvement to the system? Or the removal of its last safety valve?

Human in the loop... for 20 seconds

One of the main responses to criticism of military AI is that a human remains the final decision-maker. This is called keeping a **human in the loop**.

Remember the lab example: the second test is what protects the patient from the 10% error. So what did the "second test" look like here?

According to one source in the +972 investigation, he spent about **20 seconds** reviewing some targets, and said that in some cases his role was limited to confirming that the person the system had selected was male. He described himself as little more than a rubber stamp.

The Zionist occupation army rejected this account, saying analysts are required to conduct an independent review and verify that a target is lawful under its

directives and international law.

But if the testimony is accurate, checking a person's sex reveals nothing about the 10% error. A man who was wrongly classified is still a man. That is not a second test. It is a formal stamp on the first one.

When we trust the computer too much

There is a well-known phenomenon in systems engineering called **automation bias**: the human tendency to trust a machine's output, especially when it looks precise and scientific.

Imagine this screen appearing in front of an analyst who has 20 seconds *(an illustrative example prepared by the author, not an image from the actual system)*:

```
Person ID: 182736
Affiliation Score: 94%
Location: Building 428
Civilian Estimate: 11
Status: Target Candidate
```

Everything looks orderly. There are numbers, a percentage, a location.

But the 94% is not a fact. It is the output of a model built on data, definitions and thresholds set by people. And in 20 seconds, nobody can ask: where did this 94% come from?

Where's Daddy?: when a man comes home

Once a target is identified, another problem arises: where is he now?

According to the investigation, a system called **Where's Daddy?** was used to track selected individuals and send an alert when they entered their family homes.

```
Target ID
   ↓
```

```
Location tracking
   ↓
Target entered home
   ↓
Event
   ↓
Strike process begins
```

Every app developer knows this pattern. It is called **event-driven architecture**: the application waits for a certain event, and when it occurs, it executes an action.

But here the event is: **a man came home**. And the action may be an airstrike.

According to the sources, targeting people in their homes was easier from an intelligence standpoint. The home, the place that should be the safest and most civilian of all, becomes the easiest location to pin down. And when a man enters his home, his wife, children, parents and neighbors are around him.

NAZA: how many civilians are expected to die?

This may be the most important point in the entire film.

The system does not only ask: is there a target here? It also asks: how many civilians will be killed if we strike this place?

```
Target + Location + Building + Weapon + Estimated occupants
                        ↓
      Expected number of civilians killed
                        ↓
                  Decision
```

In the database, the line might look like this:

```
Expected civilian casualties: 17
```

To the system, 17 is an integer. To a report, it is a statistic. To a Palestinian family,

it is a father, a mother, children, siblings and relatives.

According to two sources in the +972 investigation, during the first weeks of the war the army permitted, in some strikes on low-ranking operatives, the killing of up to **15 or 20 civilians** for each person targeted. The sources said the limit was far higher for senior commanders, reaching more than 100 civilians in some cases.

These are source testimonies, and they do not mean a single number was applied to every strike; the Israeli military rejected several of the investigation's characterizations. But if they are accurate, we are not talking about civilians who died in an unforeseen event. We are talking about deaths that were **calculated in advance**.

Errors don't happen once... they compound

So far we have talked about Lavender's error alone. But go back to the pipeline at the start of this article. **Every stage has its own margin of error**. According to the same investigation, sources described errors at other stages as well:

- **Identity errors:** the system tracks the phone, not the person. In Gaza people swap phones, and a phone may pass to another family member after its owner is killed.
- **Timing errors:** in some cases time passed between the "entered home" alert and the strike, so the targeted person had already left, and his family alone was killed.
- **Civilian estimate errors:** estimates of how many people were in a building relied on rough data, in a city where displaced people moved in with relatives and every household multiplied in size.
- **Weapon errors:** according to the sources, low-ranking targets were often struck with unguided bombs because they were cheaper. These are less precise and have a wider effect.

And the problem is that these errors don't add up. They multiply.

Take an illustrative example *(the numbers here are hypothetical to explain the*

idea, except for Lavender's rate): if each of four stages is correct 90% of the time:

```
Classification × Identity × Timing × Civilian estimate
     0.9       ×   0.9    ×  0.9   ×       0.9
                    ≈ 0.66
```

In other words, a chain of four "good" stages, each correct nine times out of ten, might produce a fully correct decision in **only about two-thirds of cases**.

This is a basic rule of systems engineering: a chain is never more reliable than its weakest link, and is usually less reliable than any single link on its own.

Scale: when error multiplies

Humans make mistakes too. That's true.

But a human analyst may review dozens of cases. An automated system processes tens of thousands.

```
Human error × dozens of decisions

versus

Automated error × 37,000 classifications
```

Automation doesn't only increase speed. It amplifies every error built into the design, the data or the policy, and turns it from an individual incident into a systematic pattern.

From technical error to war crime

This is where the discussion stops being only technical.

International humanitarian law rests on three core principles: **distinction** between civilians and combatants, **proportionality** between civilian harm and military advantage, and **precaution** to minimize harm to civilians.

The Zionist occupation is not a party to Additional Protocol I to the Geneva Conventions, which spells out these principles in detail, but the core principles are considered part of customary international law, binding on all states. Let's see how each technical figure collides with them:

1. In case of doubt, a civilian

Additional Protocol I (Article 50) states that in case of doubt whether a person is a civilian, **that person shall be considered a civilian**.

A system that operates with a known error rate of at least 10%, reviewed in 20 seconds, turns this rule on its head: a person is treated as a target until proven otherwise, and nobody has the time to prove otherwise.

2. The duty to verify

The same protocol (Article 57) requires those who plan an attack to do **everything feasible** to verify that the target is military.

The question is simple: is confirming that a person is male, in 20 seconds, "everything feasible"? Especially when the army itself knows that one in ten people on the list is wrong?

3. Proportionality collapses when the target is wrong

This may be the most important technical and legal point.

The principle of proportionality is an equation: expected civilian harm on one side, military advantage on the other. When the killing of 15 or 20 civilians is permitted to strike one person, that calculation rests on the assumption that the person really is a military target.

But what if he is among the 10% error?

```
Expected civilian harm:   20 people
Military advantage:       0  (the person is not a combatant)
```

```
Result:  21 civilians killed for nothing
```

In that case, everyone killed in the strike is a civilian, including the "targeted" person himself. The proportionality equation doesn't merely become skewed. It collapses completely.

And when a system operates knowing this will happen in at least about one in ten cases, this is not a theoretical possibility. At the scale of tens of thousands, it is a statistical certainty.

4. A known error is not an accident

In war, not every mistake is a crime. A soldier who errs in circumstances he could not have foreseen is different from an institution that knows the size of the error in advance and accepts it.

The Rome Statute of the International Criminal Court lists among war crimes **intentionally directing attacks against the civilian population**, and **intentionally launching an attack in the knowledge** that it will cause civilian harm clearly excessive in relation to the military advantage anticipated.

The key word here is **knowledge**.

When an institution tests its system and knows it is wrong one time in ten, then decides to use it on tens of thousands of people, with a formal human review, a pre-set allowance for killing dozens of civilians per strike, and deliberate targeting of homes at night when families are inside, the deaths of innocent people are no longer an "error." They have become a **known and pre-accepted outcome**, built into the design of the system itself.

Between accusation and verdict

It is important to be precise: proving that a particular strike was a war crime requires investigating its circumstances, its evidence and the person who decided

on it.

But the problem NAZA and the surrounding investigations reveal is not in a single strike. It is in **the system itself**, as described by people who worked inside it: a system designed in a way that makes the mistaken killing of civilians an expected, calculated and accepted outcome. That is exactly what makes it a matter of legal accountability, not just ethical debate.

And this accountability is not hypothetical. On 21 November 2024, the International Criminal Court issued arrest warrants for Israeli Prime Minister Benjamin Netanyahu and former Defense Minister Yoav Gallant.

The Pre-Trial Chamber found **reasonable grounds to believe** that they bear responsibility for war crimes and crimes against humanity, including the war crime of starvation as a method of warfare. It also found reasonable grounds to believe that they bear criminal responsibility as civilian superiors for the war crime of intentionally directing an attack against the civilian population.

An arrest warrant is not a conviction. But it means the questions this film raises already sit inside an international judicial process.

The Palestinian sees the outcome... the system sees the data

Palestinians, journalists, doctors, rescue workers and victims' families, have spoken from the very first day about the scale of civilian loss. But when someone from inside the military explains how a target is produced, the information receives a different kind of international attention. That in itself raises a question about whose testimony is given more weight.

Still, these testimonies matter, because they reveal what the victim cannot see: the other side of the screen.

```
The Palestinian sees:     The system sees:
```

A home	Entity
A family	Device
A child	Location
A neighbor	Feature
A name	Score
A story	Target
	Collateral Damage Estimate

The Palestinian narrative says: **there was a family in this house.**

NAZA pushes us to ask: what appeared on the screen of the person who decided to strike it? Did the number of civilians appear? Did a confidence score appear? Was the targeted person even home? And how many seconds did the review take?

Conclusion: behind the score is a human being

These words are familiar to anyone who works in software:

`Data → Feature → Model → Score → Threshold → Target → Trigger → Approve`

They all look technical and neutral. But put them inside a military system and their meanings change:

The data may be a Palestinian's entire life, gathered through surveillance.
The feature may be a phone call, or a new phone after the old one was lost under the rubble.
The score may determine how suspicious he is.
The threshold may decide whether he moves onto the target list.
The tracking may know when he came home.
The NAZA may be the number of his family members expected to die with him.
And then a person at the end of the chain, with 20 seconds, presses Approve.

All of this happens while the system's owners know that at least one in every ten people on the list is not who they think he is.

In software, when we find a bug, we release a patch. We retrain the model. We change the threshold. We roll back to the previous version.

But when the output is a bomb that fell on a home, there is no rollback.

And a false positive does not stay a number in a performance report.

It becomes a name on a gravestone.