AI Can Triage Evidence. It Still Cannot Verify the Story Alone

Maria Cattini | 24/07/2026 | OSINT

What BBC Eye's Haystack system teaches OSINT teams about agents, social media evidence and human review.

The problem is not that investigators have too little data.

The problem is that they often have too much of it.

A social movement, a conflict, a disinformation campaign or an extremist network may leave thousands of posts, images, captions, comments, videos and reposts across public platforms. Somewhere inside that material there may be useful leads. But the volume itself becomes a barrier.

The OSINT question is no longer only:

```
Can we find the evidence?
```

It is also:

```
Can we sort the evidence without letting automation decide the story?
```

BBC Eye's recent Haystack project is a useful case study because it does not present AI as a replacement for investigation. It presents AI as a way to expand triage under human supervision.

According to a June 2026 Reuters Institute account by Christopher Giles, Serdar Tumgoren, Chris Zubak-Skees and Marc Perkins, BBC Eye built a multi-agent AI system named Haystack while investigating Russia's rising nationalist vigilante movement. The system helped a team of open-source investigators, reporters and computational journalists collect and analyse Russian social media material at a scale that would have been difficult to handle manually.

The reported numbers are important.

Haystack gathered around 10,000 social media posts from more than ten Russian nationalist groups and produced about 55,000 assessments of that content. Those assessments looked for signals such as nationalist ideology, references to migrants, anti-migrant raids and expressions of violence against minority groups.

But the most important part of the case is not the number.

It is the workflow.

The mistake to avoid

The weak version of this story would be:

```
AI investigated 10,000 posts.
```

That is not the useful lesson.

The stronger version is:

```
AI helped investigators triage 10,000 posts, surface leads and query the dataset, whi
le reporters retained control over direction, review and interpretation.
```

That distinction matters for OSINT.

An AI system can label, cluster, translate, count, search and route material. It can make a large corpus easier to inspect. It can surface posts that a human team may not have had time to find. It can reduce the cost of asking a first-pass question across thousands of items.

But it cannot verify the story alone.

It does not know whether a lead is legally safe to publish. It does not know whether a pattern is politically meaningful. It does not know whether a term is being used literally, ironically, euphemistically or as an in-group signal. It does not know whether a single post is representative of a movement or only loud enough to be noticed.

The risk is not simply hallucination.

The risk is misplaced authority.

If the output of an AI triage system is treated as a conclusion, the investigation becomes fragile. If the output is treated as a queue of leads, the system becomes more useful.

What Haystack did

The Reuters Institute article describes Haystack as a multi-agent system built to mirror several tasks inside an investigations team.

Those tasks included:

• collecting posts from Russian social media sites;
• assessing image and text-based posts for relevant signals;
• allowing journalists to ask natural-language questions about the dataset;
• helping non-technical reporters run data analysis that would usually require database or programming skills.

The system used multiple AI agents rather than a single all-purpose prompt.

From a newsroom perspective, this matters. A single prompt asking a model to collect, classify, analyse and conclude creates too much room for hidden decision-making. A multi-step process can make the work more inspectable if each stage is bounded.

The Reuters Institute account says BBC Eye explored a more automated system, but found that it was safer to keep reporters involved at each stage. Journalists provided instructions and

clarifications as agents moved through the process.

That design choice is the core OSINT lesson.

The system did not remove the investigator.

It changed where the investigator spent attention.

The human-in-the-loop part is not decoration

"Human in the loop" is often used as a vague safety phrase.

In this case, it has practical meaning.

The team did not only ask the system for answers. They refined the system's behaviour, reviewed assessments, checked outputs and manually verified leads. The Reuters Institute article notes that once the team was confident enough to run Haystack across more posts, the leads it surfaced were still verified by team members who manually reviewed the evidence.

That is the line OSINT teams should keep.

AI can support triage.

It can support lead generation.

It can support counting and querying.

It can support translation and cross-language discovery.

But the finding still needs a human evidence chain.

A useful evidence chain should answer:

- What source material was collected?
- What query or prompt was used?
- What did the model classify?
- What confidence or category was assigned?
- What reasoning did the model provide?
- What did a human reviewer check?
- What source independently supports the lead?
- What remains uncertain?

Without that chain, AI-assisted OSINT becomes a black box with a newsroom interface.

Ambiguity is an investigative variable

One of the most useful details in the Reuters Institute account is small but important.

The BBC Eye team improved Haystack by reducing ambiguity in the labels used for classification. Early prompts asked whether a post contained references to raids using categories such as "definitely", "definitely not", "probably" and "probably not". The team found more precise leads when they narrowed the options to:

```
yes
no
not sure
```

That is a practical lesson for anyone using AI in OSINT.

Complex labels can feel more nuanced. They can also create false precision.

"Probably" may look more informative than "not sure", but it can hide a weak inference. A simple category system forces the reviewer to make the next step explicit.

If the model says yes, inspect the evidence.

If the model says no, sample-check the misses.

If the model says not sure, route the item to human review or a narrower query.

The goal is not to make the model sound sophisticated.

The goal is to create a useful review queue.

AI can find language humans may miss

Haystack also surfaced terms and euphemisms used to refer to migrant workers, according to the Reuters Institute account. That matters because extremist, vigilante and political communities rarely use one stable vocabulary.

They use slang.

They use code.

They use jokes.

They use euphemisms.

They use platform-specific shorthand.

They use words that are obvious to insiders and invisible to outsiders.

For OSINT teams, this is where AI can help. A model can detect recurring language patterns across a large corpus and bring them to the surface. It can help non-native or non-specialist reviewers notice terms that deserve expert review.

But this is also where the risk increases.

Language is contextual.

A term can be derogatory in one setting, neutral in another and ironic in a third. A phrase can be a signal of ideology, a reposted quote, a counter-speech example or a false positive.

The right workflow is not:

```
The model found a term, so the term proves the pattern.
```

The right workflow is:

```
The model found a term. Now we test how it is used, who uses it, where it appears, an
d whether experts or native speakers confirm the interpretation.
```

That is slower.

It is also safer.

The OSINT workflow

For ProjectOSINT, the Haystack case can be turned into a reusable workflow.

1. Build the seed list carefully

The first question is not what the model can do.

The first question is what material enters the system.

A weak seed list creates a weak investigation. If the initial groups, accounts, channels or sources are poorly chosen, the system may scale the wrong sample.

Record:

- source names;
- platform;
- why each source was included;
- date range;
- access method;
- collection limits;
- known bias in the dataset.

The dataset is never neutral.

It is a constructed object.

2. Separate collection from assessment

Do not let one opaque step collect, classify and summarise everything.

Separate the phases.

Collection answers:

```
What did we gather?
```

Assessment answers:

```
What signals may appear in the material?
```

Analysis answers:

```
What pattern can be supported after review?
```

Keeping those stages separate makes errors easier to find.

3. Use narrow labels

The label system should be simple enough to audit.

For first-pass classification, use categories such as:

```
yes
no
not sure
```

Then preserve the reason for the label.

A classification without a reason is difficult to review. A reason without the source material is weak. A source without a claim component is not yet evidence.

4. Review the model's misses, not only its hits

It is tempting to review only the items the system flags as relevant.

That can hide systematic failure.

Sample the "no" pile.

Sample the "not sure" pile.

Check whether the model misses slang, visual cues, coded language, sarcasm, screenshots, memes, reposts or text embedded in images.

An OSINT triage system should be tested for recall as well as precision.

5. Keep the finding human-authored

The model can surface the lead.

The investigator must build the finding.

A publishable finding should not be:

```
The AI found 427 examples.
```

It should be:

```
We collected a defined corpus, used AI-assisted triage to surface candidate posts, ma
nually reviewed the relevant evidence, checked examples against source material, and
found a pattern supported by the following records.
```

The second sentence is longer because the evidence chain is visible.

That visibility is the point.

What this case does not prove

Haystack is not proof that AI agents can investigate independently.

It is not proof that every newsroom should build a multi-agent system.

It is not proof that a model can classify political speech without risk.

It is not proof that scale automatically produces truth.

What it does show is narrower and more useful:

```
AI-assisted systems can help OSINT teams triage large public datasets when the workfl
ow keeps humans in control of direction, review, source interpretation and final clai
ms.
```

That is enough.

In OSINT, enough is often better than too much.

The practical rule

Use AI where volume breaks human attention.

Do not use AI where judgment defines the finding.

The investigation still needs:

- source selection;
- evidence preservation;
- prompt and query logging;
- classification review;
- expert language checks;
- manual verification;
- uncertainty statements;
- human accountability.

The useful future is not an autonomous investigator.

It is a better evidence desk.

One that helps humans find the right material faster, without pretending that sorting is the same as verifying.

That distinction is where the work begins.

What BBC Eye's Haystack system teaches OSINT teams about agents, social media evidence and human review.

The problem is not that investigators have too little data.

The problem is that they often have too much of it.

A social movement, a conflict, a disinformation campaign or an extremist network may leave thousands of posts, images, captions, comments, videos and reposts across public platforms. Somewhere inside that material there may be useful leads. But the volume itself becomes a barrier.

The OSINT question is no longer only:

```
Can we find the evidence?
```

It is also:

```
Can we sort the evidence without letting automation decide the story?
```

BBC Eye's recent Haystack project is a useful case study because it does not present AI as a replacement for investigation. It presents AI as a way to expand triage under human supervision.

According to a June 2026 Reuters Institute account by Christopher Giles, Serdar Tumgoren, Chris Zubak-Skees and Marc Perkins, BBC Eye built a multi-agent AI system named Haystack while investigating Russia's rising nationalist vigilante movement. The system helped a team of open-source investigators, reporters and computational journalists collect and analyse Russian social media material at a scale that would have been difficult to handle manually.

The reported numbers are important.

Haystack gathered around 10,000 social media posts from more than ten Russian nationalist groups and produced about 55,000 assessments of that content. Those assessments looked for signals such as nationalist ideology, references to migrants, anti-migrant raids and expressions of violence against minority groups.

But the most important part of the case is not the number.

It is the workflow.

The mistake to avoid

The weak version of this story would be:

```
AI investigated 10,000 posts.
```

That is not the useful lesson.

The stronger version is:

```
AI helped investigators triage 10,000 posts, surface leads and query the dataset, whi
le reporters retained control over direction, review and interpretation.
```

That distinction matters for OSINT.

An AI system can label, cluster, translate, count, search and route material. It can make a large corpus easier to inspect. It can surface posts that a human team may not have had time to find. It can reduce the cost of asking a first-pass question across thousands of items.

But it cannot verify the story alone.

It does not know whether a lead is legally safe to publish. It does not know whether a pattern is politically meaningful. It does not know whether a term is being used literally, ironically, euphemistically or as an in-group signal. It does not know whether a single post is representative of a movement or only loud enough to be noticed.

The risk is not simply hallucination.

The risk is misplaced authority.

If the output of an AI triage system is treated as a conclusion, the investigation becomes fragile. If the output is treated as a queue of leads, the system becomes more useful.

What Haystack did

The Reuters Institute article describes Haystack as a multi-agent system built to mirror several tasks inside an investigations team.

Those tasks included:

• collecting posts from Russian social media sites;
• assessing image and text-based posts for relevant signals;
• allowing journalists to ask natural-language questions about the dataset;
• helping non-technical reporters run data analysis that would usually require database or programming skills.

The system used multiple AI agents rather than a single all-purpose prompt.

From a newsroom perspective, this matters. A single prompt asking a model to collect, classify, analyse and conclude creates too much room for hidden decision-making. A multi-step process can make the work more inspectable if each stage is bounded.

The Reuters Institute account says BBC Eye explored a more automated system, but found that it was safer to keep reporters involved at each stage. Journalists provided instructions and clarifications as agents moved through the process.

That design choice is the core OSINT lesson.

The system did not remove the investigator.

It changed where the investigator spent attention.

The human-in-the-loop part is not decoration

"Human in the loop" is often used as a vague safety phrase.

In this case, it has practical meaning.

The team did not only ask the system for answers. They refined the system's behaviour, reviewed assessments, checked outputs and manually verified leads. The Reuters Institute article notes that once the team was confident enough to run Haystack across more posts, the leads it surfaced were still verified by team members who manually reviewed the evidence.

That is the line OSINT teams should keep.

AI can support triage.

It can support lead generation.

It can support counting and querying.

It can support translation and cross-language discovery.

But the finding still needs a human evidence chain.

A useful evidence chain should answer:

• What source material was collected?
• What query or prompt was used?
• What did the model classify?
• What confidence or category was assigned?
• What reasoning did the model provide?
• What did a human reviewer check?
• What source independently supports the lead?
• What remains uncertain?

Without that chain, AI-assisted OSINT becomes a black box with a newsroom interface.

Ambiguity is an investigative variable

One of the most useful details in the Reuters Institute account is small but important.

The BBC Eye team improved Haystack by reducing ambiguity in the labels used for classification. Early prompts asked whether a post contained references to raids using categories such as "definitely", "definitely not", "probably" and "probably not". The team found more precise leads when they narrowed the options to:

```
yes
no
not sure
```

That is a practical lesson for anyone using AI in OSINT.

Complex labels can feel more nuanced. They can also create false precision.

"Probably" may look more informative than "not sure", but it can hide a weak inference. A simple category system forces the reviewer to make the next step explicit.

If the model says yes, inspect the evidence.

If the model says no, sample-check the misses.

If the model says not sure, route the item to human review or a narrower query.

The goal is not to make the model sound sophisticated.

The goal is to create a useful review queue.

AI can find language humans may miss

Haystack also surfaced terms and euphemisms used to refer to migrant workers, according to the Reuters Institute account. That matters because extremist, vigilante and political communities rarely use one stable vocabulary.

They use slang.

They use code.

They use jokes.

They use euphemisms.

They use platform-specific shorthand.

They use words that are obvious to insiders and invisible to outsiders.

For OSINT teams, this is where AI can help. A model can detect recurring language patterns across a large corpus and bring them to the surface. It can help non-native or non-specialist reviewers notice terms that deserve expert review.

But this is also where the risk increases.

Language is contextual.

A term can be derogatory in one setting, neutral in another and ironic in a third. A phrase can be a signal of ideology, a reposted quote, a counter-speech example or a false positive.

The right workflow is not:

```
The model found a term, so the term proves the pattern.
```

The right workflow is:

```
The model found a term. Now we test how it is used, who uses it, where it appears, an
d whether experts or native speakers confirm the interpretation.
```

That is slower.

It is also safer.

The OSINT workflow

For ProjectOSINT, the Haystack case can be turned into a reusable workflow.

1. Build the seed list carefully

The first question is not what the model can do.

The first question is what material enters the system.

A weak seed list creates a weak investigation. If the initial groups, accounts, channels or sources are poorly chosen, the system may scale the wrong sample.

Record:

- source names;
- platform;
- why each source was included;
- date range;
- access method;
- collection limits;
- known bias in the dataset.

The dataset is never neutral.

It is a constructed object.

2. Separate collection from assessment

Do not let one opaque step collect, classify and summarise everything.

Separate the phases.

Collection answers:

```
What did we gather?
```

Assessment answers:

```
What signals may appear in the material?
```

Analysis answers:

```
What pattern can be supported after review?
```

Keeping those stages separate makes errors easier to find.

3. Use narrow labels

The label system should be simple enough to audit.

For first-pass classification, use categories such as:

```
yes
no
not sure
```

Then preserve the reason for the label.

A classification without a reason is difficult to review. A reason without the source material is weak. A source without a claim component is not yet evidence.

4. Review the model's misses, not only its hits

It is tempting to review only the items the system flags as relevant.

That can hide systematic failure.

Sample the "no" pile.

Sample the "not sure" pile.

Check whether the model misses slang, visual cues, coded language, sarcasm, screenshots, memes, reposts or text embedded in images.

An OSINT triage system should be tested for recall as well as precision.

5. Keep the finding human-authored

The model can surface the lead.

The investigator must build the finding.

A publishable finding should not be:

```
The AI found 427 examples.
```

It should be:

```
We collected a defined corpus, used AI-assisted triage to surface candidate posts, ma
nually reviewed the relevant evidence, checked examples against source material, and
found a pattern supported by the following records.
```

The second sentence is longer because the evidence chain is visible.

That visibility is the point.

What this case does not prove

Haystack is not proof that AI agents can investigate independently.

It is not proof that every newsroom should build a multi-agent system.

It is not proof that a model can classify political speech without risk.

It is not proof that scale automatically produces truth.

What it does show is narrower and more useful:

```
AI-assisted systems can help OSINT teams triage large public datasets when the workfl
ow keeps humans in control of direction, review, source interpretation and final clai
ms.
```

That is enough.

In OSINT, enough is often better than too much.

The practical rule

Use AI where volume breaks human attention.

Do not use AI where judgment defines the finding.

The investigation still needs:

- source selection;
- evidence preservation;
- prompt and query logging;
- classification review;

• expert language checks;
• manual verification;
• uncertainty statements;
• human accountability.

The useful future is not an autonomous investigator.

It is a better evidence desk.

One that helps humans find the right material faster, without pretending that sorting is the same as verifying.

That distinction is where the work begins.