

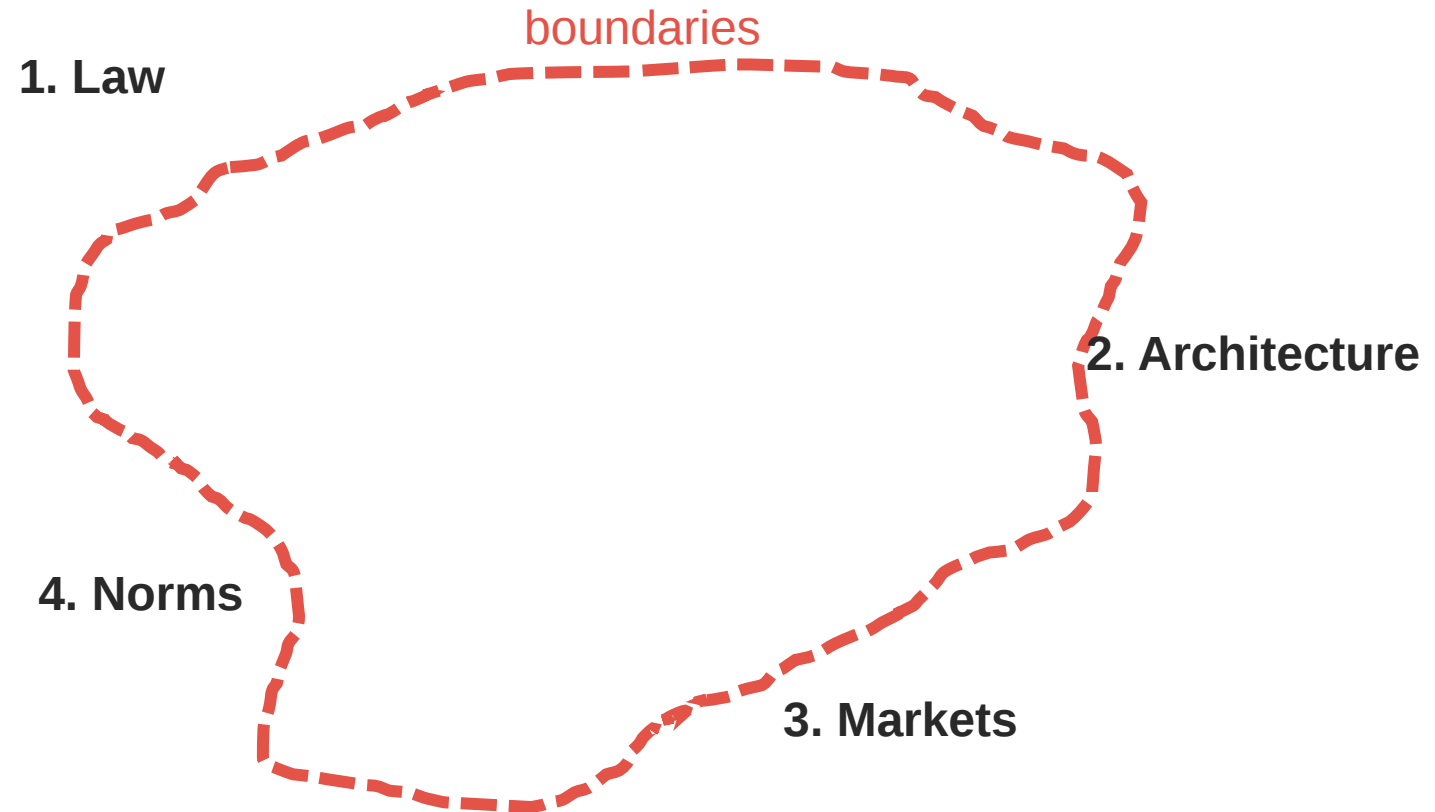
Open Ethics Transparency Protocol

How self-disclosure can help
to bring safer digital diets

Nikita Lukianets

Founder @OpenEthicsAI, PocketConfidant
Fellow, the Alliance of Democracies Foundation

Human decisions are bounded with **4 factors**



Disclosure help us to make informed choices

Ingredients

Allergen labeling

Barcode

Sell-by date information

Caloric and nutritional value information and GDA

Manufacturer information

Identification mark

Mobile tagging





Allowing informed choice

What if we bring labels to
digital products?



Choose a License

https://creativecommons.org/choose/

creative commons

Share your work

Use & remix

What We do

Blog

Your choices on this panel will update the other panels on this page.

Allow adaptations of your work to be shared?

Yes

No

Yes, as long as others share alike

?

Allow commercial uses of your work?

Yes

No

?

Selected License

Attribution-NoDerivatives 4.0 International

CC

=

This is not a Free Culture License.

*

Help others attribute you!

This part is optional, but filling it out will add machine-readable metadata to the suggested HTML!

Have a web page?

CC

BY

ND

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0



Motivation

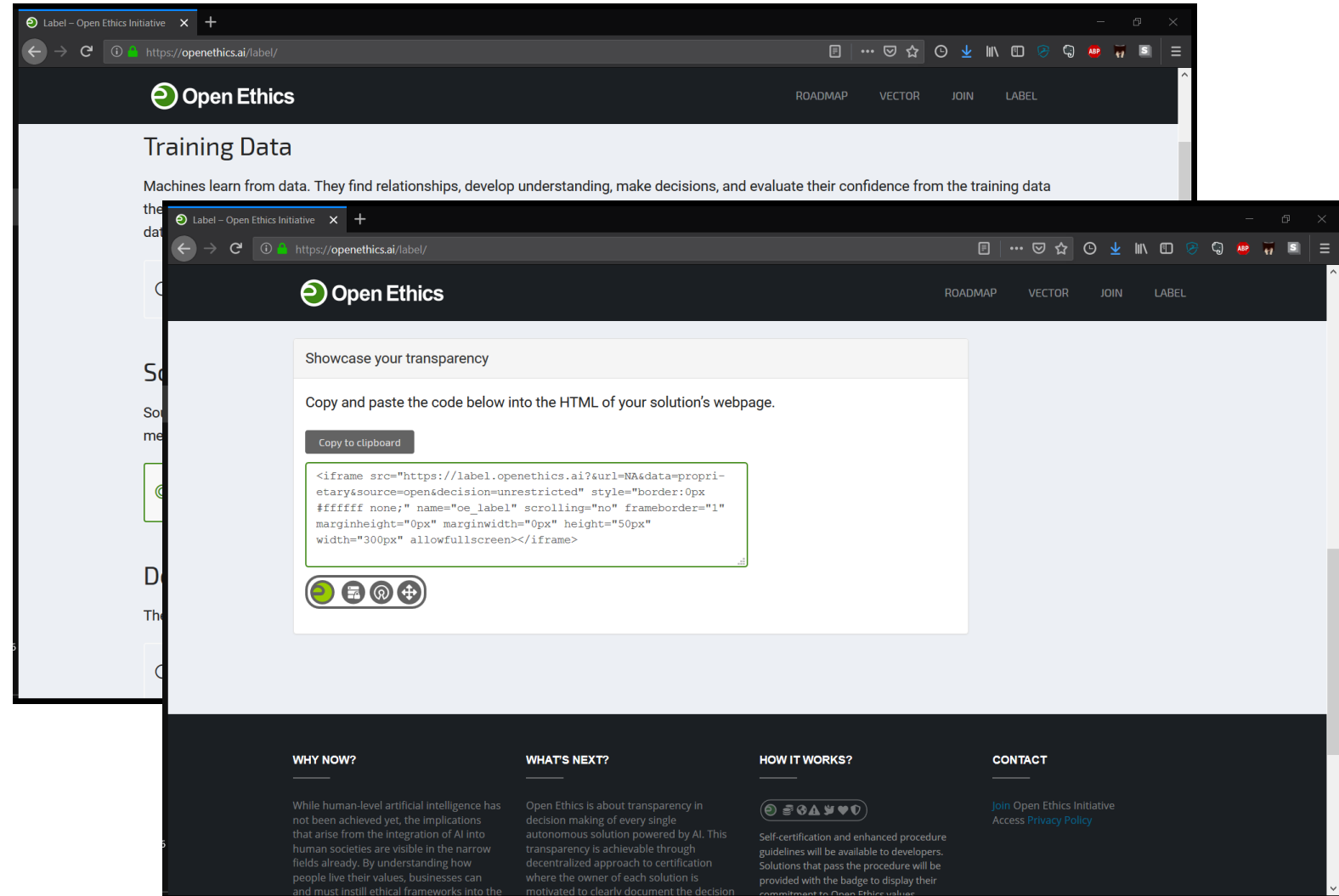
- **Informed consumer choices** : End-users able to make informed choices based on their own ethical preferences and ADM disclosure.
- **Industrial scale monitoring** : Discovery of best and worst practices within market verticals, technology stacks, and product value offerings.
- **Legally-agnostic guidelines** : Suggestions for developers and product-owners, formulated in factual language, which are legally-agnostic and could be easily transformed into product requirements and safeguards.
- **Iterative improvement** : Digital products, specifically, the ones powered by artificial intelligence could receive a nearly real-time feedback on how their performance and ethical posture could be improved to cover security, privacy, diversity, fairness, power-balance, non-discrimination, and other requirements.
- **Labeling and certification** : Mapping to existing and future regulatory initiatives and standards.



Extending The Open Ethics Label



A trust-building platform.
Self-disclosure
for AI systems.



<https://openethics.ai/label>



Open Ethics Transparency Protocol

ai ethics
PASSPORT



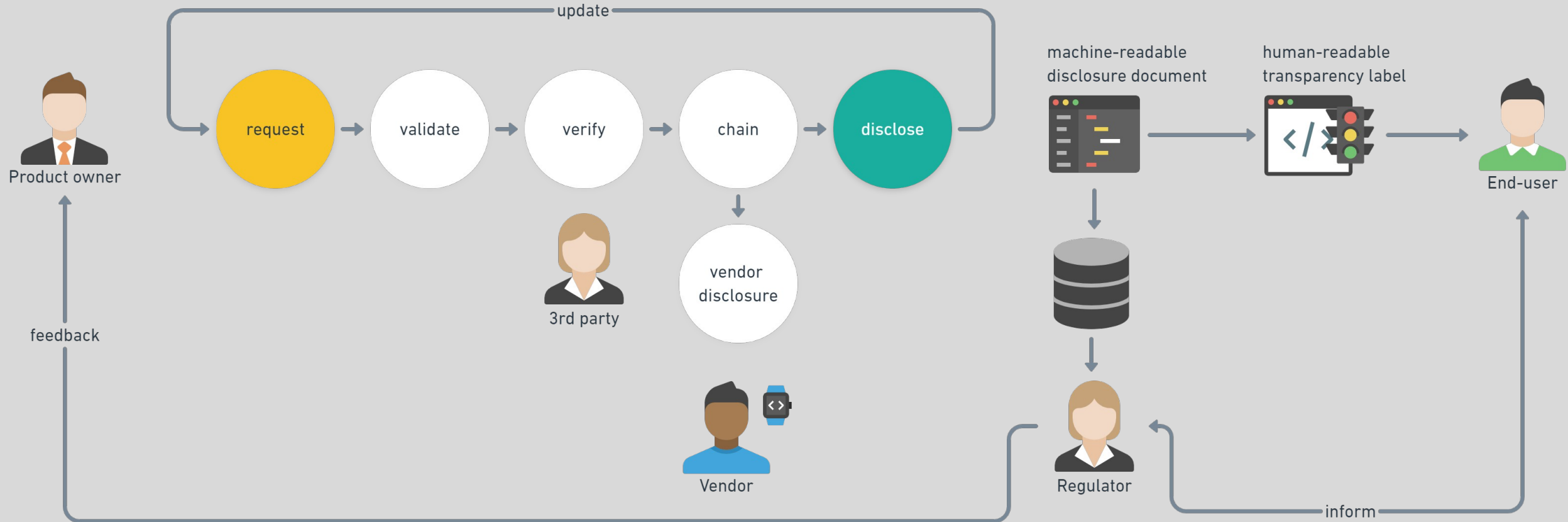
```
1 {
2   "schema": {
3     "name": "Open Ethics Transparency Protocol",
4     "version": "0.9.2 RFC"
5   },
6   "data": {
7     "availability": "public",
8     "batch": 14032507,
9   >   "sources": [ ...
11     ]
12
13   },
14 >   "algorithms": [ ...
53   ],
54 >   "decision-space": { ...
59   },
60   "validation": {
61     "critical": true,
62     "methods": [
63       {
64         ..... "name": "HITL",
65         ..... "objective": "To confirm the diagnosis by
66                   human doctors before presenting to the
67                   patient",
66         ..... "type": "council"
67       }
68     ]
69   }
70 }
```

Display “ethical” posture
in both human and
machine-readable ways.



Open Ethics Transparency Protocol

transparency chaining lifecycle

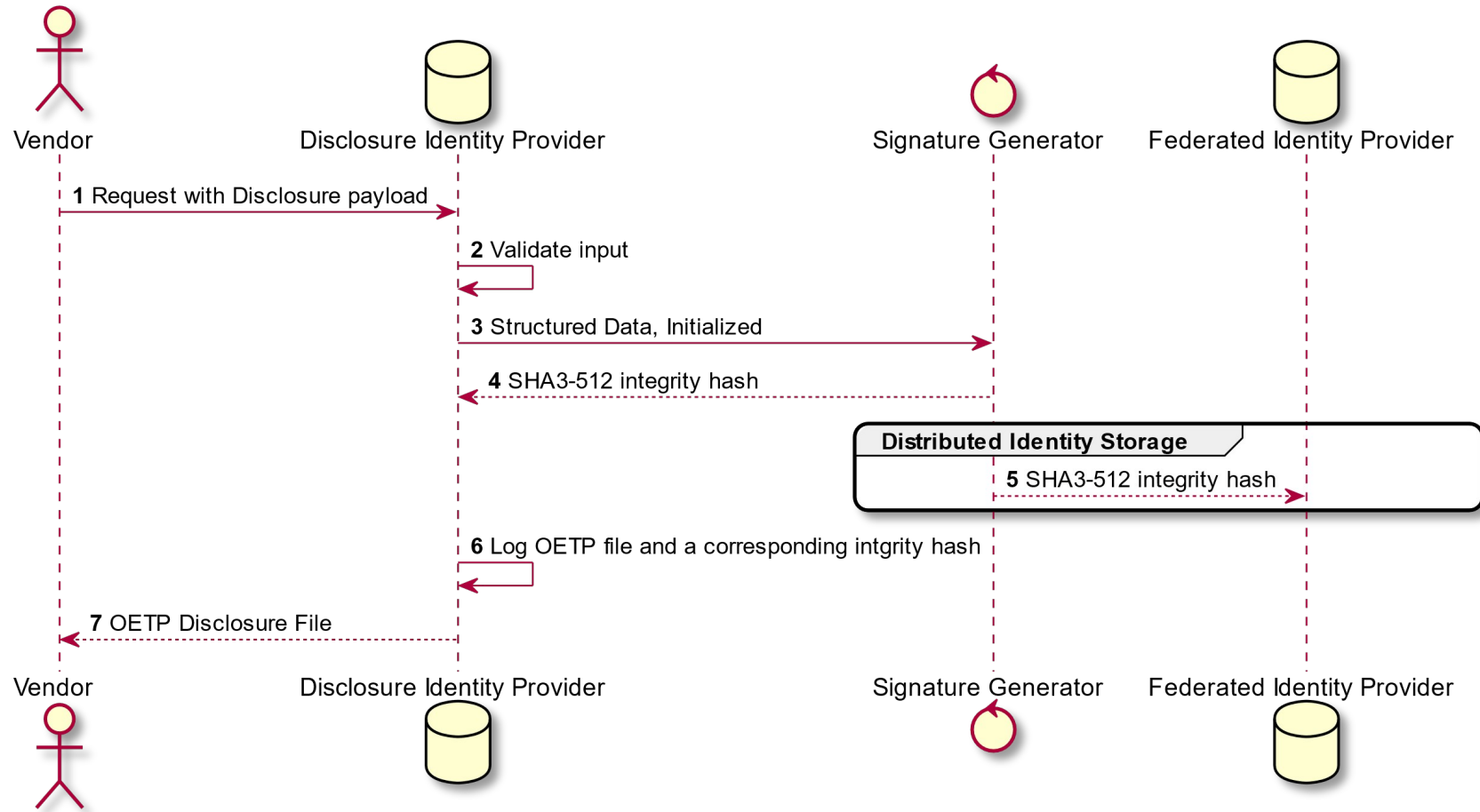




Implementation



Basic Disclosure Submission



Generate Open Ethics Label

+

←→↺🛡🔒https://openethics.ai/label/generate/

🇺🇸🌟📄🕒🔌📶🗑9🔄☰

Open Ethics

MANIFESTO

CANVAS

EVENTS

PROJECTS

ABOUT

JOIN

Human-in-the-Loop component

AI swarm + Human

Humans confirm

HITL design objective

Personal Data Record

Account ID

Source of the data record

permanent

Purpose of the data record processing

☐ Consent obtained directly

☐ Identifying record

Source

Privacy Policy

https://example.com/privacy

Add comments

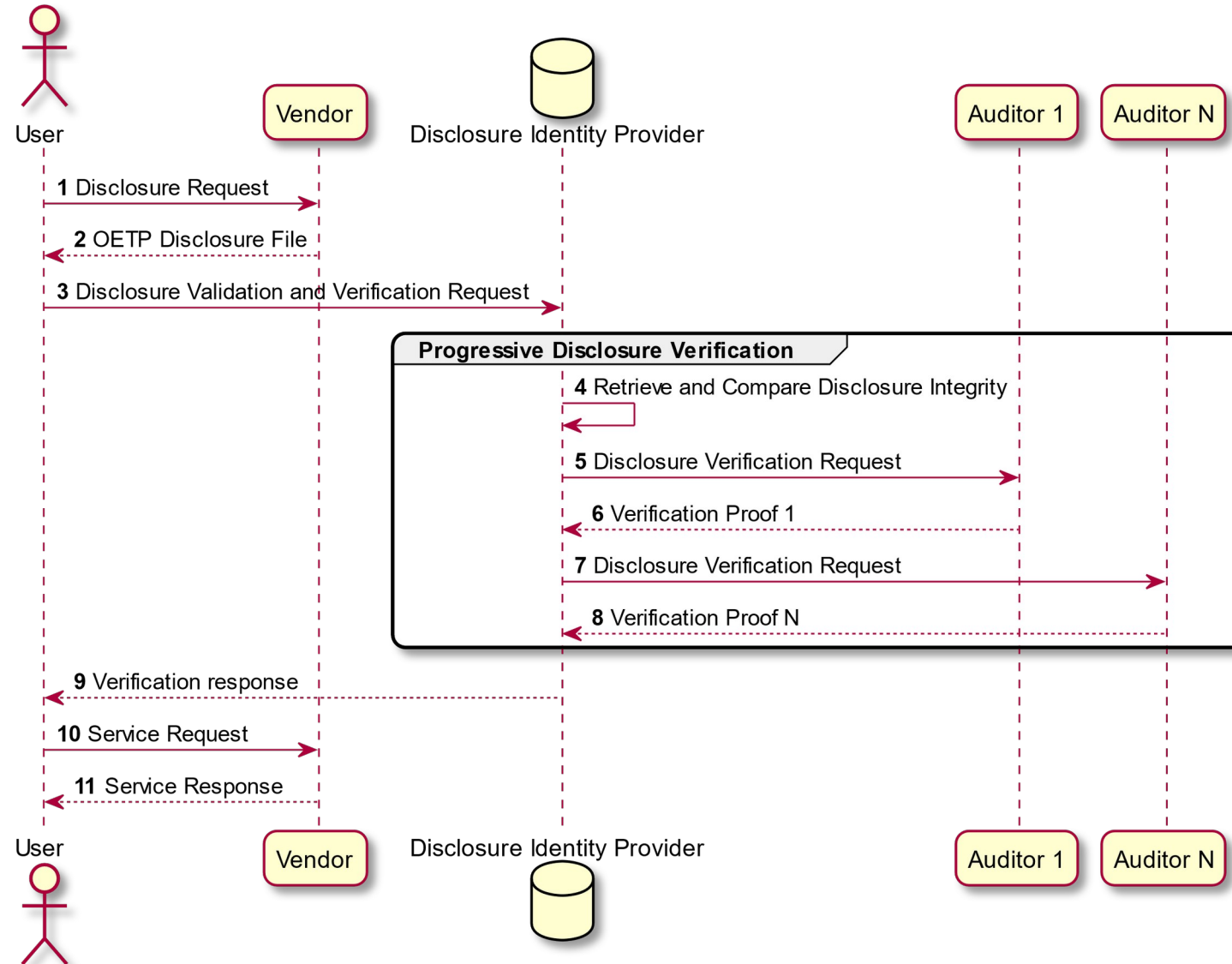
+ Personal Data Record

Display transparency

This is the preview. After submission of the form, you will receive the code to paste Open Ethics Label on your website.



Progressive Verification with multiple Auditors




Generate Open Ethics Label – ×

+

← → ↺ 🛡️ 🔒 https://openethics.ai/label/generate/

📄 🇺🇸 ☆ 📄 ⌚ ⬇️ 🏠 ABP S 🗑️ M+ 🗑️ 9 🔄 ⋮

MANIFESTO CANVAS EVENTS PROJECTS ▾ ABOUT ▾ JOIN

Integrity

Open Ethics has generated the following SHA3-512 hash for the snapshot of submitted data.

093b41c91413496d71c839105f0c2c882d3d34ded7456103cdd84e0a4c4aa6ac6ca35141865cd00c99bb2adaa1240ff306325dd80555e2fe9c7f9d7f0b2752cc


Open Ethics Transparency Protocol

Save the code below as a text file named *oetp.json* in the root of your product's website. It will allow integrity validation, as well as machine readable scenarios for AI transparency management.

Download oetp.json

```
1 {
2   "schema": {
3     "name": "Open Ethics Transparency Protocol",
4     "version": "0.9.3 RFC",
5     "integrity": "093b41c91413496d71c839105f0c2c882d3d34ded7456103cdd84e0a4c4aa6ac6ca35141865cd00c99bb2adaa1240ff306325dd80555e2fe9c7f9d7f0b2752cc"
6   },
7   "snapshot": {
8     "product": {
9       "url": "ietftest.com",
10      "description": ""
11    }
12  }
13 }
```

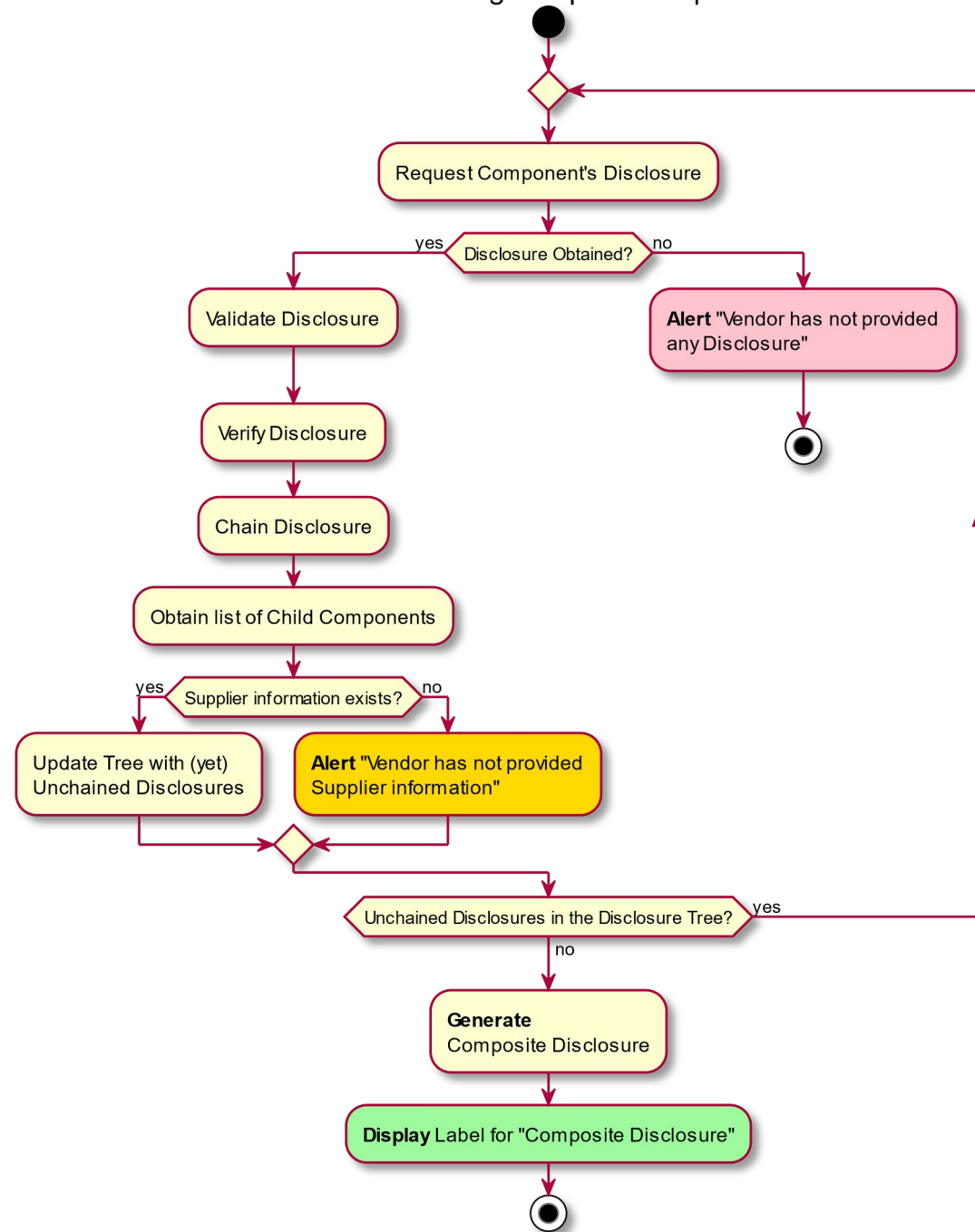
Display transparency



Save and paste the Open Ethics Label HTML code into the HTML of your product's webpage.

Download oel.html

Disclosure Chaining: Request-Response





Implications

AI alignment



Safety



Failure Transparency



Judicial Transparency



Responsibility



Value Alignment



Personal Privacy



Liberty and Privacy



Shared Benefit



Human Control



Non-subversion



Compassion



Shared Prosperity



Emotional Value



Cross-national Security



Non-influence



Traceability of
decisions



Safety

Failure Transparency

Judicial Transparency



Value Alignment



Human Control



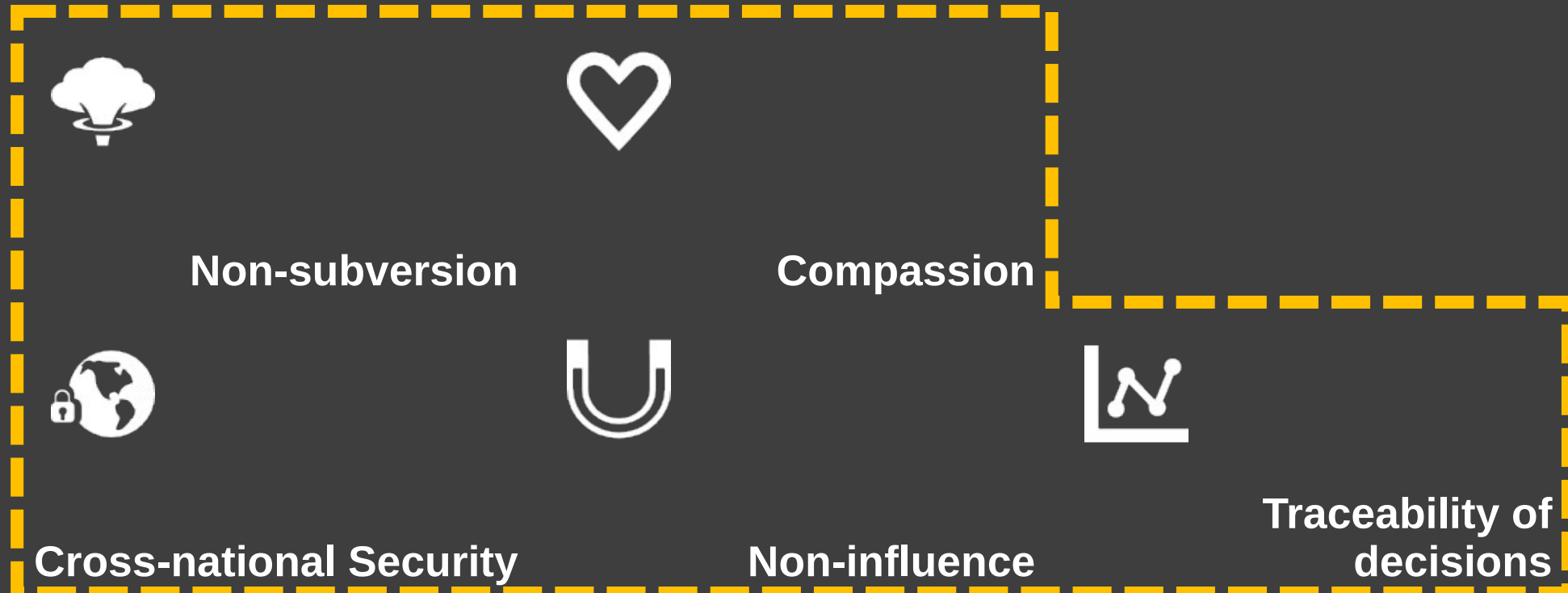
Emotional Value



POV 1



POV 2





Alternative 1



Safety

Failure Transparency

Judicial Transparency

Responsibility



Value Alignment



Personal Privacy



Liberty and Privacy



Shared Benefit



Human Control



Non-subversion



Compassion



Shared Prosperity



Emotional Value



Cross-national Security



Non-influence



Traceability of
decisions



Alternative 2



Safety

Failure Transparency

Judicial Transparency

Responsibility



Value Alignment

Personal Privacy

Liberty and Privacy

Shared Benefit



Human Control

Non-subversion

Compassion

Shared Prosperity



Emotional Value

Cross-national Security

Non-influence

Traceability of
decisions



Alternative 3



Safety

Failure Transparency

Judicial Transparency

Responsibility



Value Alignment

Personal Privacy

Liberty and Privacy

Shared Benefit



Human Control

Non-subversion

Compassion

Shared Prosperity



Emotional Value

Cross-national Security

Non-influence

Traceability of
decisions

Alternatives



Aligned choice



Alternative 1			
Safety	Failure Transparency	Judicial Transparency	Responsibility
Value Alignment	Personal Privacy	Liberty and Privacy	Shared Benefit
Human Control	Non-subversion	Compassion	Shared Prosperity
Emotional Value	Cross-national Security	Non-influence	Traceability of decisions

Alternative 2			
Safety	Failure Transparency	Judicial Transparency	Responsibility
Value Alignment	Personal Privacy	Liberty and Privacy	Shared Benefit
Human Control	Non-subversion	Compassion	Shared Prosperity
Emotional Value	Cross-national Security	Non-influence	Traceability of decisions

Alternative 3			
Safety	Failure Transparency	Judicial Transparency	Responsibility
Value Alignment	Personal Privacy	Liberty and Privacy	Shared Benefit
Human Control	Non-subversion	Compassion	Shared Prosperity
Emotional Value	Cross-national Security	Non-influence	Traceability of decisions



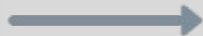
Use-case

example

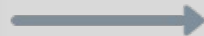
Public Surveillance Transparency Project (PST)

Station Entrance →

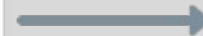
Citizen in the
public space



Scan QR code
on the sticker



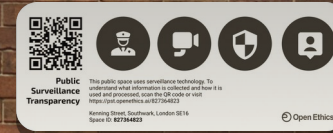
Read the
disclosure



Provide
feedback



1. Visual transparency labels for citizens
2. Machine-readable disclosure
3. Disclosure as Open Data [governance/market]



Market creation: Location-based services

Open Data

McKinsey reference: **5.2 bn EUR**

ODI estimates: **1.2 bn EUR**

LBS

in Europe alone estimates (2025): **10 bn EUR**

<https://theodi.org/article/the-value-of-open-data-for-the-private-sector/>

<https://www.cio.com/article/3626390/6-ways-to-harness-the-value-of-open-data.html>



Sharing is caring

Open Ethics Canvas



Talk #ethics, include your #tech team

Download and use it for free

<https://openethics.ai/canvas/>

Designed For		Designed By	Date	Version
Scope <ul style="list-style-type: none">What is this product designed for?In which context it operates?	Training Data <ul style="list-style-type: none">How was the training data collected?How do you ensure its representativeness?Does your training dataset contain personal data?Who annotates the data and how quality is controlled?What is the data labeling process that you employ?	Algorithms & Source Code <ul style="list-style-type: none">Do you use open or proprietary sources? Why? Which?Who in the team is setting the heuristics (rules) which influence the output?How do you ensure the quality of used third-party codebases?What is your process of making the key architectural choices?	Decision Space <ul style="list-style-type: none">What exactly does the product do?Can you provide the list of all possible outputs?How incorrectly supplied inputs are spotted?Is there anomaly detection in place?	
Users <ul style="list-style-type: none">What type of users does this product have? (customers/admins/ etc)What are their roles?				
Key Stakeholders <ul style="list-style-type: none">Who are the key stakeholders?What influence do they have over the product?How do stakeholders interact with each other?How is the power distributed?	Personal Data Processing <ul style="list-style-type: none">Which personal data is collected by the product?What is the purpose of collecting personal data?How is this data processed? Used? Stored? Deleted?	Components & Subprocessing <ul style="list-style-type: none">Which third parties are engaged by the product?How do you evaluate the potential impacts of API on the quality of your product's output?How do you check the reliability of your data processing contractors?	Failure Modes <ul style="list-style-type: none">How failures are detected and monitored?What are the possible failures of a product?What actions are performed if a product fails?	
Values & Interests <ul style="list-style-type: none">What values do stakeholders/users have?Where these values can clash or create tensions?What is known at the moment and how assumptions are tested?How can you align your technology to the values you want to support/people desire?	Explainability <ul style="list-style-type: none">How is interpretability defined for the system?What interpretability methods are used?What metrics are used in result interpretation?How interpretations of the output are communicated?	Human in the Loop (HITL) <ul style="list-style-type: none">What is the role of a human agent in the validation/verification of the outputs?What is the role of a human agent in refining the model performance?What is the decision-making power assigned to human agents responsible for the quality of output?	Model Performance Metrics <ul style="list-style-type: none">Which metrics are used to evaluate the product performance?Which measures are used to re-evaluate Accuracy, Recall, Precision, and F1- Score?	
			Decision Feedback & Objection <ul style="list-style-type: none">How does the product allow for structured feedback?How can the user challenge the application output?Which are the third parties involved in claims/objectation resolution?	
Impact Assessment <ul style="list-style-type: none">What potential harms can your product cause? (loss of opportunity, discrimination, economic loss, social stigma, detriment, emotional distress, etc)?What are the risks of the product's failure?What impact product can cause if deployed at scale?How is the product influencing the existing markets?		Regulatory Landscape <ul style="list-style-type: none">What is the regulatory context in which the product operates?Is the model portable to other market verticals?What are the involved regulatory risks?	Mitigation <ul style="list-style-type: none">How do you test for bias and fairness? What fairness definitions do you employ and why?Does your team reflect a diversity of opinions, backgrounds, and thoughts?Do you have a process for redress if people are harmed by the outputs?How fast can you shut down your product in production if it behaves badly?Who and how should be informed?	
Changes in Behavior <ul style="list-style-type: none">Do the automated decisions have significant legal or similar effects on the users/stakeholders?How the users may change their behavior after use?What are the potentials for power imbalance?	Group Interactions <ul style="list-style-type: none">What are potential changes in group behavior?How is the product addressing group interests?What new groups could be born due to the product deployment at scale?	Comments		

The Open Ethics Canvas v1.0 © 2021 by Open Ethics contributors
Designed by Nikita Lukianets, Alice Pavliou, Vlad Nekutenko
Licensed under Attribution-ShareAlike 4.0 International
<https://openethics.ai/canvas>

Open Ethics



Nikita Lukianets
n.lukianets@openethics.ai