AI Explainability WG

August 2023

August 2023 updates

- TrustyAI service
  - Metrics endpoint improvements
  - CI using operator deployment
- Core
  - TSSaliency merged
  - Python extension framework in review
  - Storage backend ADR
- Python TrustyAI
  - 0.3.0 released
  - Documentation: new examples added
- Community
  - New release process

Python TrustyAI

Python TrustyAI

- Python TrustyAI 0.3.0 released
  - https://pypi.org/project/trustyai/0.3.0/
- Inclusion of AIX360’s TSLime and TSSaliency
  - Available on ODH images on the next release
- Examples
  - AIX360 repository examples[1]
  - AI Explainability 360 Toolkit for Time-Series and Industrial Use Cases[2]
    - Added to Python TrustyAI examples
- Developer
  - Jupyter notebook validation and tests added to CI
  - SNYK vulnerability scans added to Python TrustyAI
  - Moving release process to PyPi’s “Trusted Publishers”[3]

[1]: https://github.com/trustyai-explainability/trustyai-explainability-python-examples
[2]: https://ai-library-examples.github.io/aix4industries/books/intro.html
[3]: https://blog.pypi.org/posts/2023-04-20-introducing-trusted-publishers/

TrustyAI core / service

TrustyAI core / service

- Explainers
    - TSSaliency merged (IBM Research contribution)
- Metrics
    - Fixes to Prometheus metrics endpoints
    - *e.g.* properly deleting metrics from canceled requests
    - Ad-hoc metrics request now respect batch size
- Storage
    - ADR-0006: Persistence Layer Requirements[1]
- Developer
    - CI now using new TrustyAI operator KfDef
    - Remove deprecated local Compose demos and tidy up of repo documentation

[1]: https://github.com/trustyai-explainability/community/pull/12

TrustyAI core / service

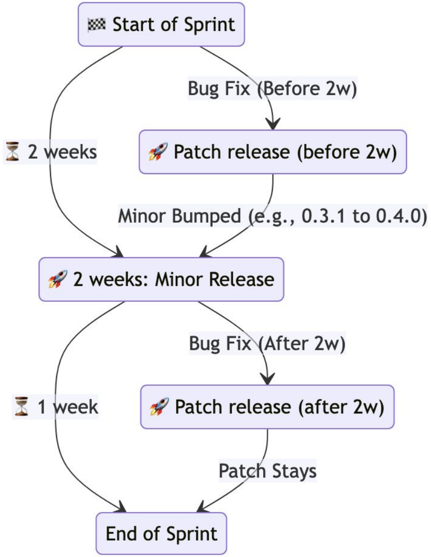
- Extending TrustyAI service with Python
  - PR under review[1]
  - Ability to extended ODH component with arbitrary Python algorithms
  - AIX360 TSLime and TSICE in initial stage
  - Example on creating custom explainer
    - https://github.com/trustyai-explainability/trustyai-explainability/blob/0c14dbb3fa22db91be8a4169ac8ecee49dd3f2fc/explainability-external/README.md

[1]: https://github.com/trustyai-explainability/trustyai-explainability/pull/300

Community

Community

- New upstream release schedule proposal[1]
  - Minor releases every Sprint, 2nd week
    - Next release: Friday 18th August
  - Release branches to be used, main branch for development
  - Simultaneous release
    - Python TrustyAI (0.3.0 → 0.4.0)
    - TrustyAI core / service (0.3.0 → 0.4.0)
    - TrustyAI operator (1.8.0 → 1.9.0)
  - Bugfix / patch releases
    - Mid-cycle
  - Major release
    - Ad-hoc, major architectural changes, TBD in XAI WG
  - Automated releases
  - Add to community calendar



🏁 Start of Sprint

Bug Fix (Before 2w)

⏳ 2 weeks

🚀 Patch release (before 2w)

Minor Bumped (e.g., 0.3.1 to 0.4.0)

🚀 2 weeks: Minor Release

Bug Fix (After 2w)

⏳ 1 week

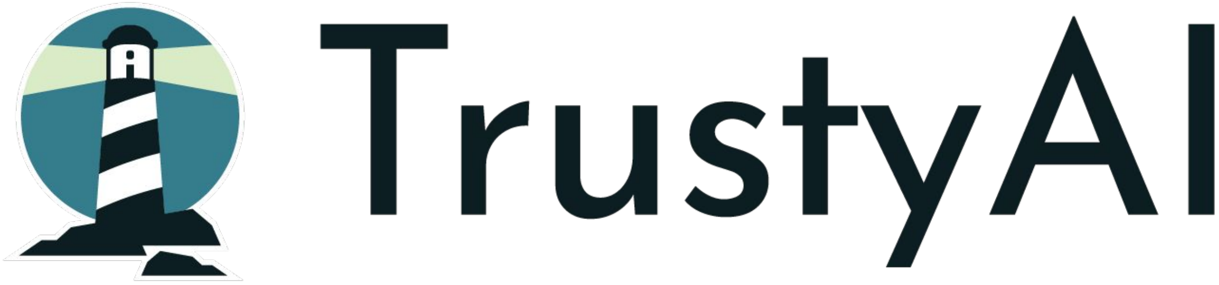
🚀 Patch release (after 2w)

Patch Stays

End of Sprint

[1]: https://github.com/trustyai-explainability/community/pull/13

Community



- TrustyAI has an official logo
  - https://design.jboss.org/trustyai/

[1]: https://github.com/trustyai-explainability/community/pull/8

Toxic Content in (Large) Language Models

Toxic Content in (Large) Language Models

- Implemented both masking and rephrasing (MaRCo paper)
        - https://github.com/trustyai-explainability/trustyai-explainability/issues/289
    - Contacted original paper author
        - https://github.com/shallinan1/MarcoDetoxification/issues/1
    - “Expert models” released on HuggingFace
    - On a personal repo for now
        - https://huggingface.co/tteofili/gminus
        - https://huggingface.co/tteofili/gplus
    - Might make sense to move them to an ODH / TrustyAI owned one

[1]: https://github.com/trustyai-explainability/community/pull/8

Toxic Content in (Large) Language Models

- original:
  - You**'ll** be fine! Just talk like a **white** person

Mask toxic content

- masked:
  - You **<mask>** be fine! Just talk like a **<mask>** person

Rephrase masked content

- rephrased:
  - You **can** be fine! Just talk like a person

[1]: https://github.com/trustyai-explainability/community/pull/8

Roadmap

TrustyAI 2023 roadmap

July 2023

- *Explainers*
  - Support for explainers LIME, SHAP, CF at service level
- *Metrics*
  - Flexible scheduling/batching
  - Improve service metadata endpoints
    - Include available categories
- *Operator*
  - TrustyAI Operator v1
- *Explainers*
  - Support for external explainability libraries

September 2023

- *Storage*
  - Wider storage support (database backends)
- *Metrics*
  - Additional metrics
  - Metrics statistical tests
- KServe integration
- Drift detection
- HAP/PII

December 2023

- *Storage*
  - Wider storage support (database backends)
- *Explainers*
  - NLP explainability support
- *Metrics*
  - Support for user-defined historical windows