

Anonymizing without losing communicative intent?

Bettina Berendt

TU Berlin, Weizenbaum Institute, and KU Leuven

Dimitri Staufer

TU Berlin

LLMs and the Patterns of Human Language Use, Berlin, 29/30 August 2024

Slides and paper available at <https://people.cs.kuleuven.be/~bettina.berendt/>

Curious to read more?

These slides are based on

- Berendt, B. & Schiffner, S. (2022). Whistleblower protection in the digital age - why 'anonymous' is not enough. Towards an interdisciplinary view of ethical dilemmas. *International Review of Information Ethics*, 31(1). [PDF](#)
- Staufer, D., Pallas, F., & Berendt, B. (2024). Silencing the Risk, Not the Whistle: A Semi-automated Text Sanitization Tool for Mitigating the Risk of Whistleblower Re-Identification. In *Proc. of FAccT 2024* (pp. 733-745). ACM. [PDF](#)

Where is whistleblowing happening ?

Public Sector

- Admin (corruption)
- Secret Services
- Military
- Law enforcement

Private Sector

- Insider trading
- Creative bookkeeping
- Abuse of power

What makes these different?

Public Sector

- Preservation of power

Private Sector

- Monetary advantage
 - Company
 - individual



Why is
whistle-
blowing
important?

‘The Enron of Germany’: Wirecard scandal casts a shadow on corporate governance

PUBLISHED MON, JUN 29 2020•4:37 AM EDT | UPDATED MON, JUN 29 2020•5:22 AM EDT



Ryan Browne
[@RYAN_BROWNE_](#)

SHARE    

KEY POINTS

- The Wirecard accounting scandal has raised fresh questions about corporate governance, with some experts calling it the “Enron of Germany.”
- German financial regulator BaFin has come under fire for its handling of the situation, with the government now calling for regulatory reform.
- There are also questions about why EY, Wirecard’s auditor, didn’t pick up on accounting irregularities that date back years.

How can whistleblowing be encouraged?

- The risk of retaliation is a major disincentive for potential whistleblowers (WBs).
- How to reduce this risk? Make retaliation
 - illegal (or at least protect WBs legally)
 - otherwise shunned or even unattractive
 - impossible

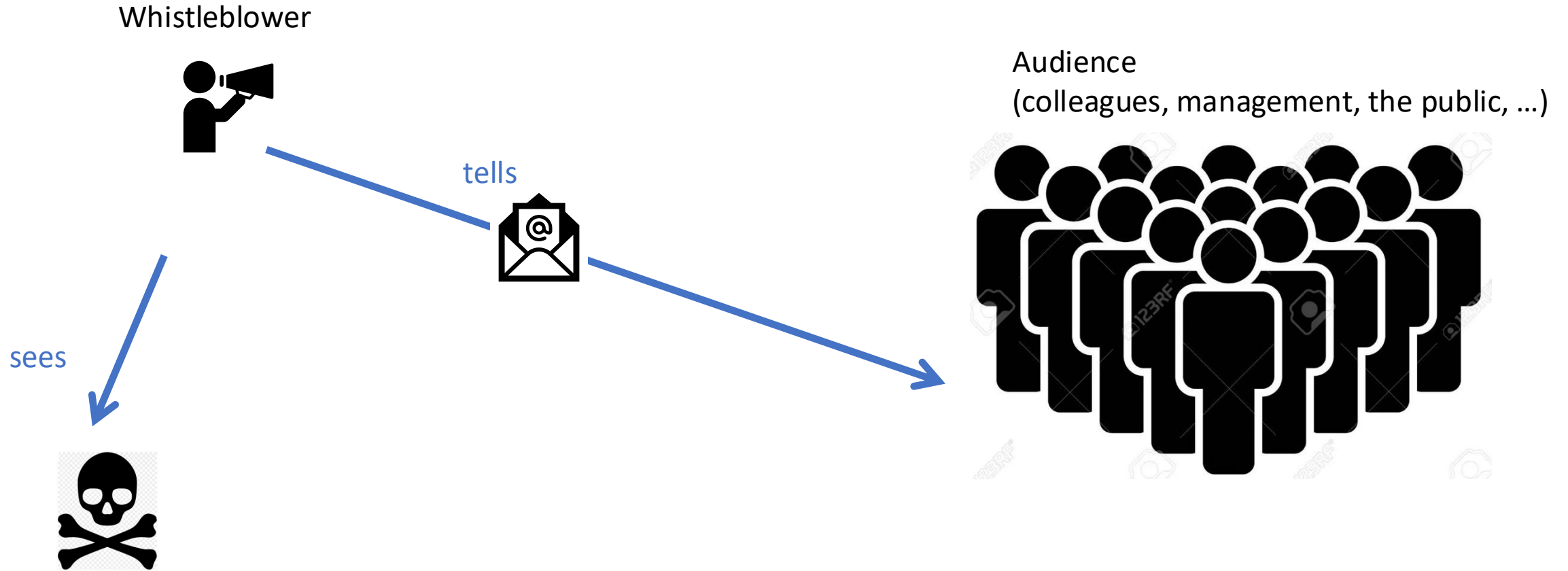
Law

Ethics codes, error culture, ...

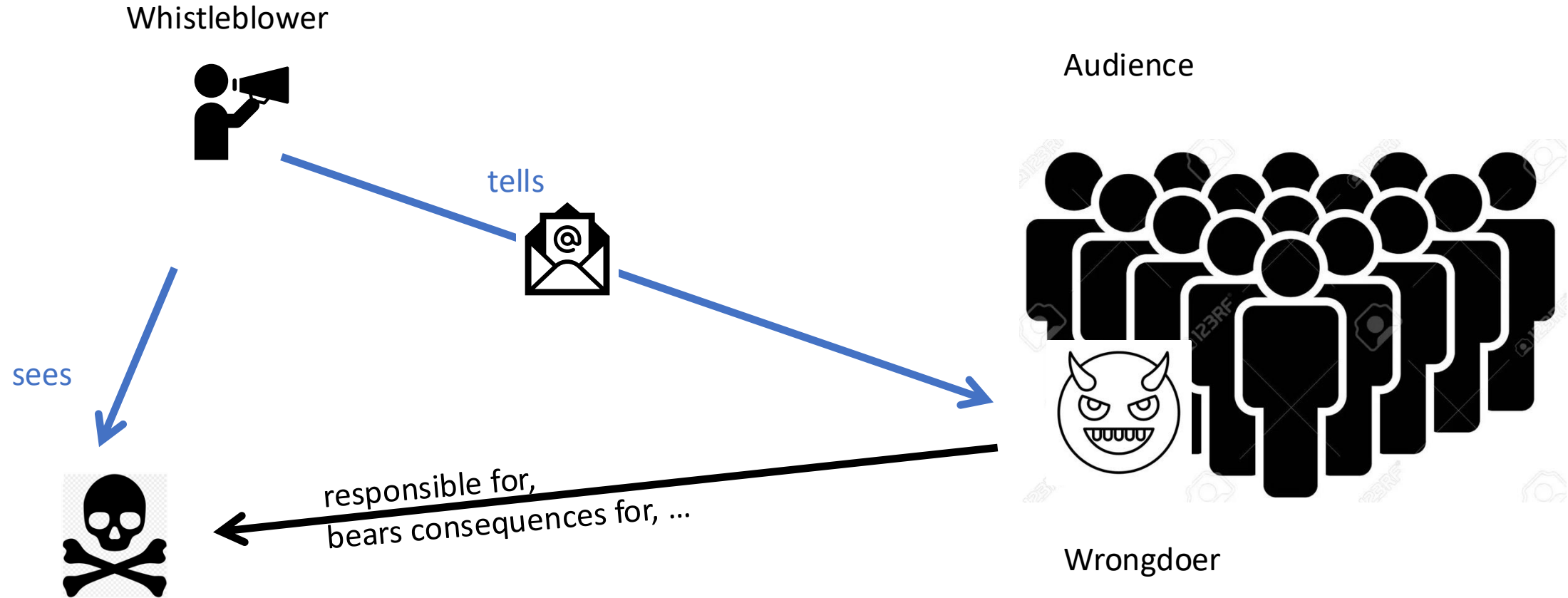
Possibility of anonymous reporting



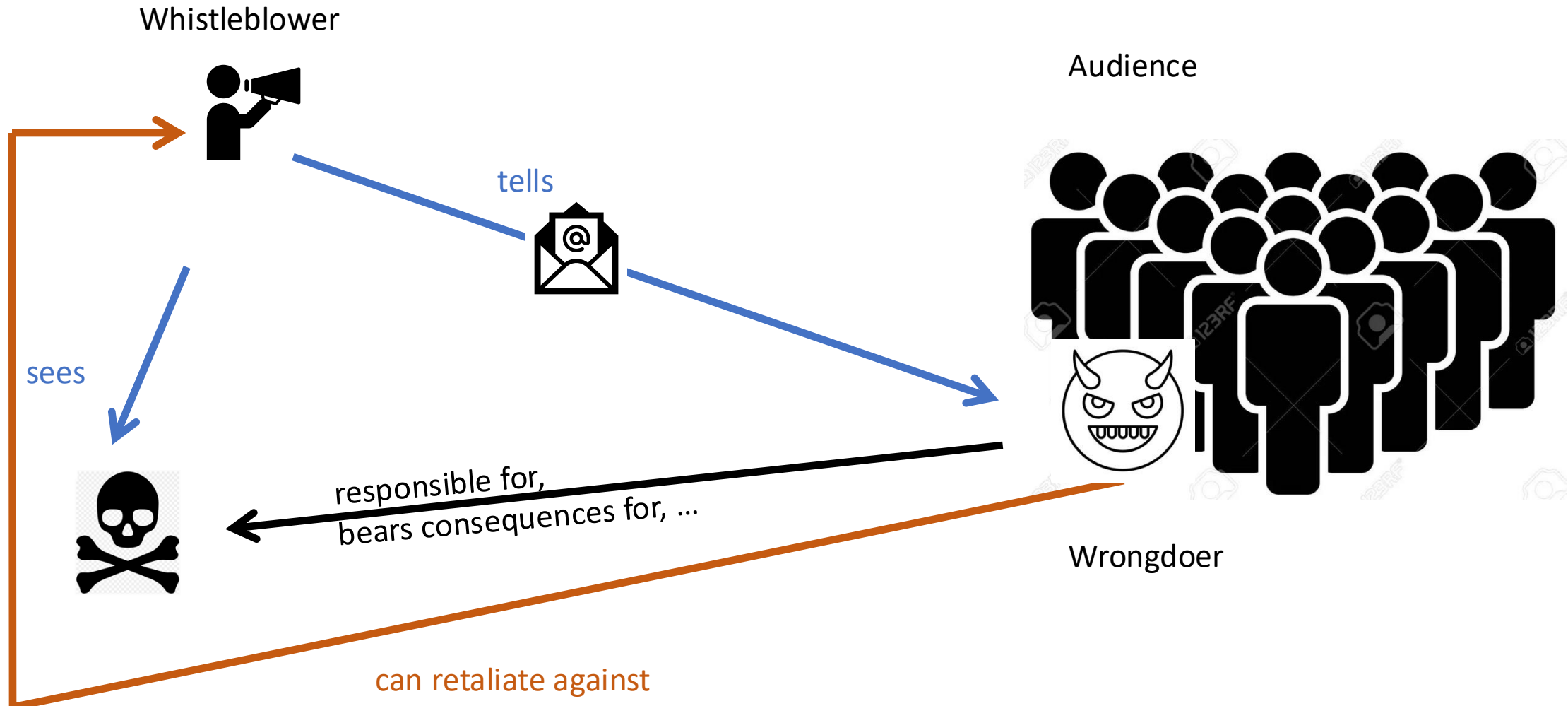
Whistleblowing as a communications problem (a *very* simplified view) (1)



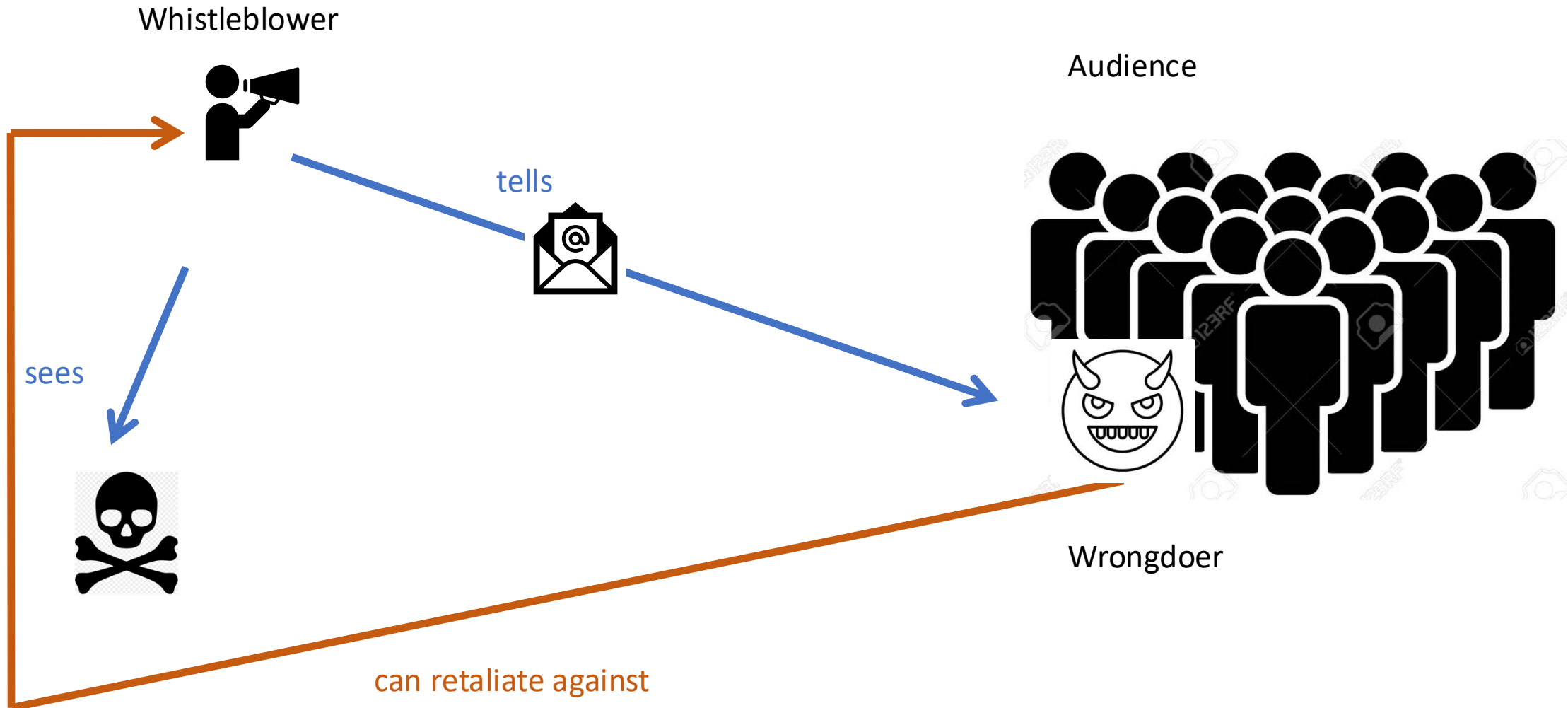
Whistleblowing as a communications problem (2)



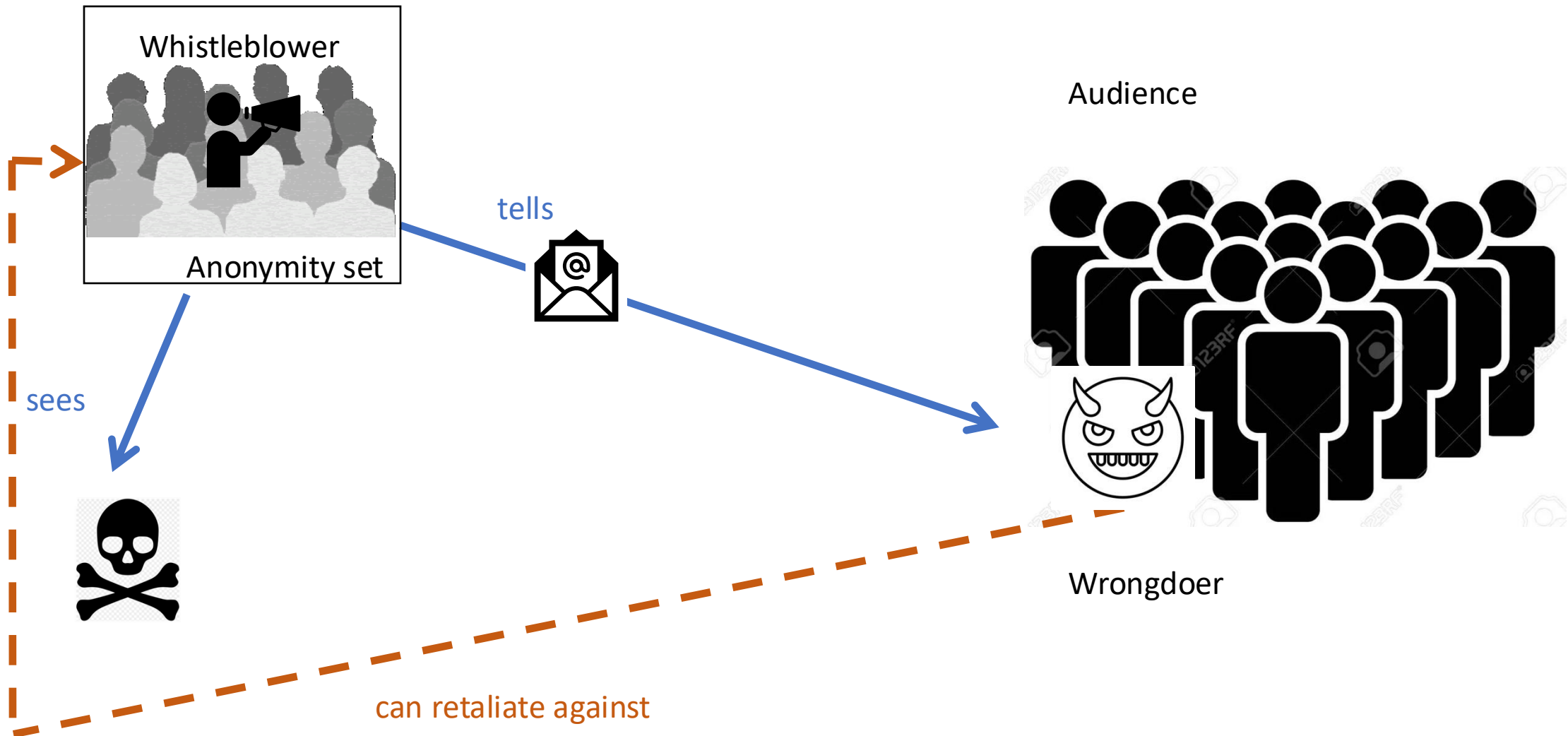
Whistleblowing as a communications problem (3)



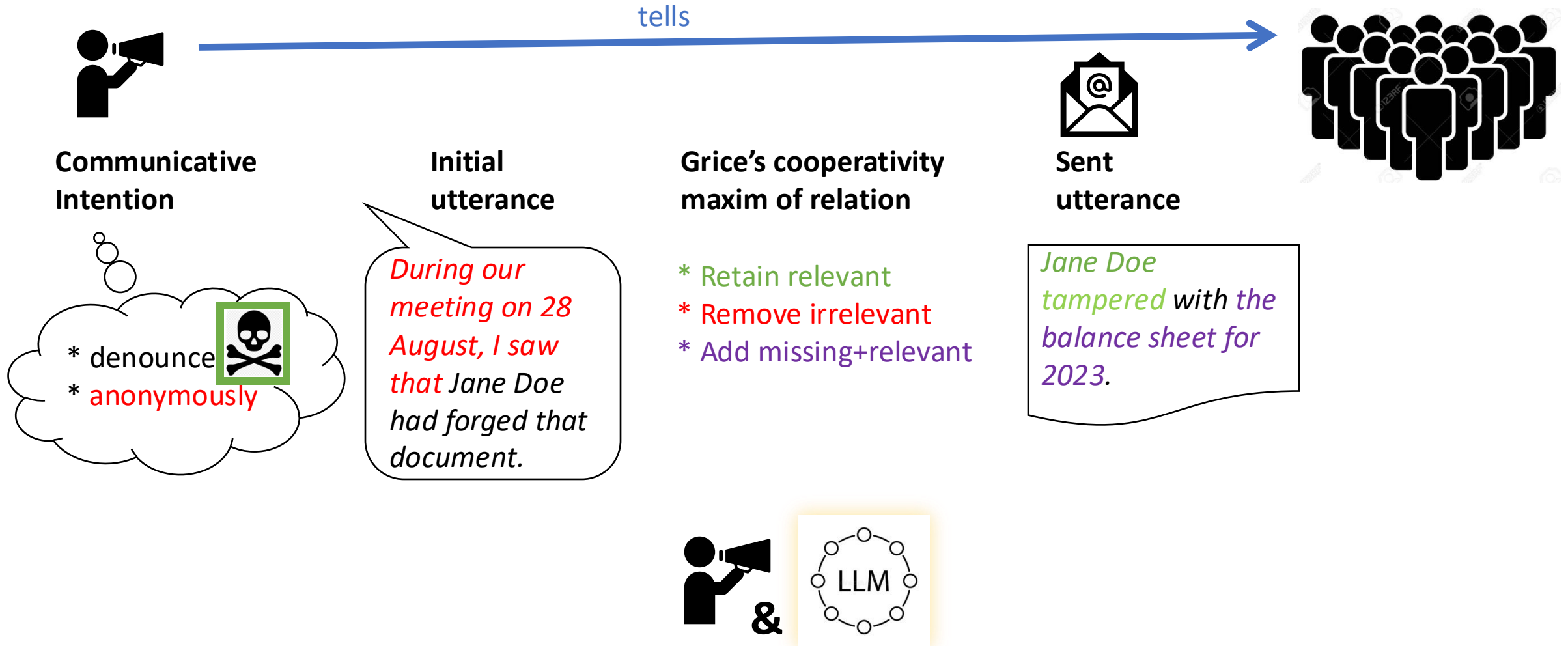
Whistleblowing as a communications problem (4)



Whistleblowing with anonymity: the promise



Whistleblowing as a problem of cooperative communication (5) – our goal



Threats to anonymity

- Mistaking *confidentiality* for anonymity
 - a trusted entity knows the identity, pressure on this entity can reveal the identity
- *Direct re-identification*: based on cues
 - legal name, pseudonyms, fingerprints/DNA, unwise choices case management data
- *Addresses* of various types
 - physical location, email address, telephone number, IP address, GPS coordinates
- *Security measures*: Need-to-know, tracking, logging
- Inferences from report *metadata*
 - e.g. when a report was made, the voice of the reporter on a telephone hotline, the linguistic style and revealed lingo of a written report
- *Epistemic non-anonymisability*: Who are the knowers?
 - Small anonymity set. The message content may imply identity.

Anonymity is hiding in the masses

Idea: using general anonymity services such as Tor to hide.

SecureDrop (originally DeadDrop by Aaron Swartz and Kevin Poulsen, 2013)

- Implemented as hidden service in Tor
- Target Userbase: Journalists and their sources
- NT, Intercept, Süddeutsche, apache.be

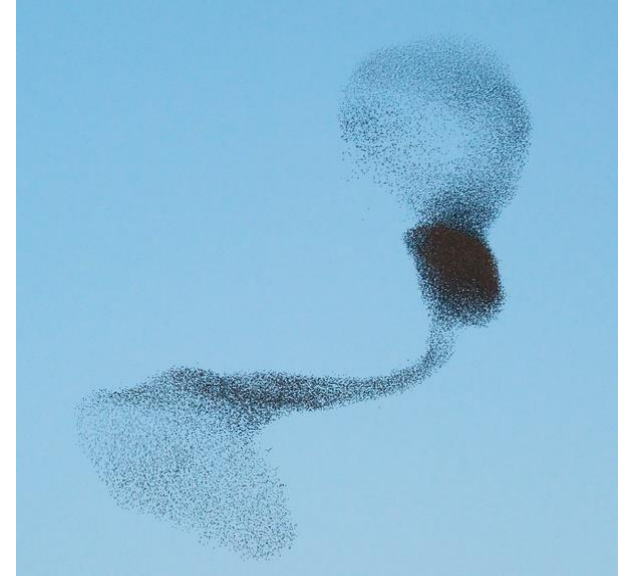


Daniel J. Sieradski - Flickr: Aaron Swartz



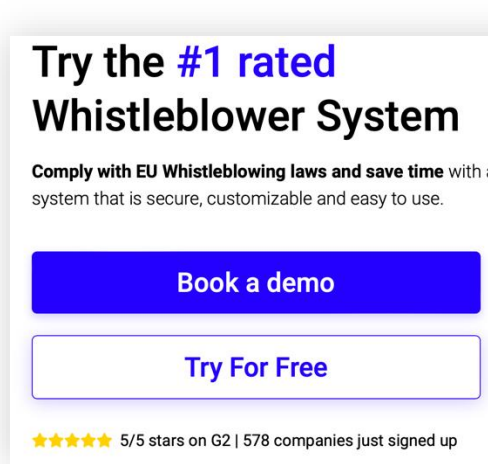
GlobaLeaks (2010)

- Implemented as Tor hidden Service
- Target Userbase: WB in Public service
- AWP: Ljost (Iceland), Filtrala (Spain), EcuadorTransparente , PeruLeaks



Whistleblowing Software

Anonymous communication



Try the **#1 rated**
Whistleblower System

Comply with EU Whistleblowing laws and save time with a system that is secure, customizable and easy to use.

[Book a demo](#)

[Try For Free](#)

★★★★★ 5/5 stars on G2 | 578 companies just signed up

[1]



due to...

Author's unique **writing style**

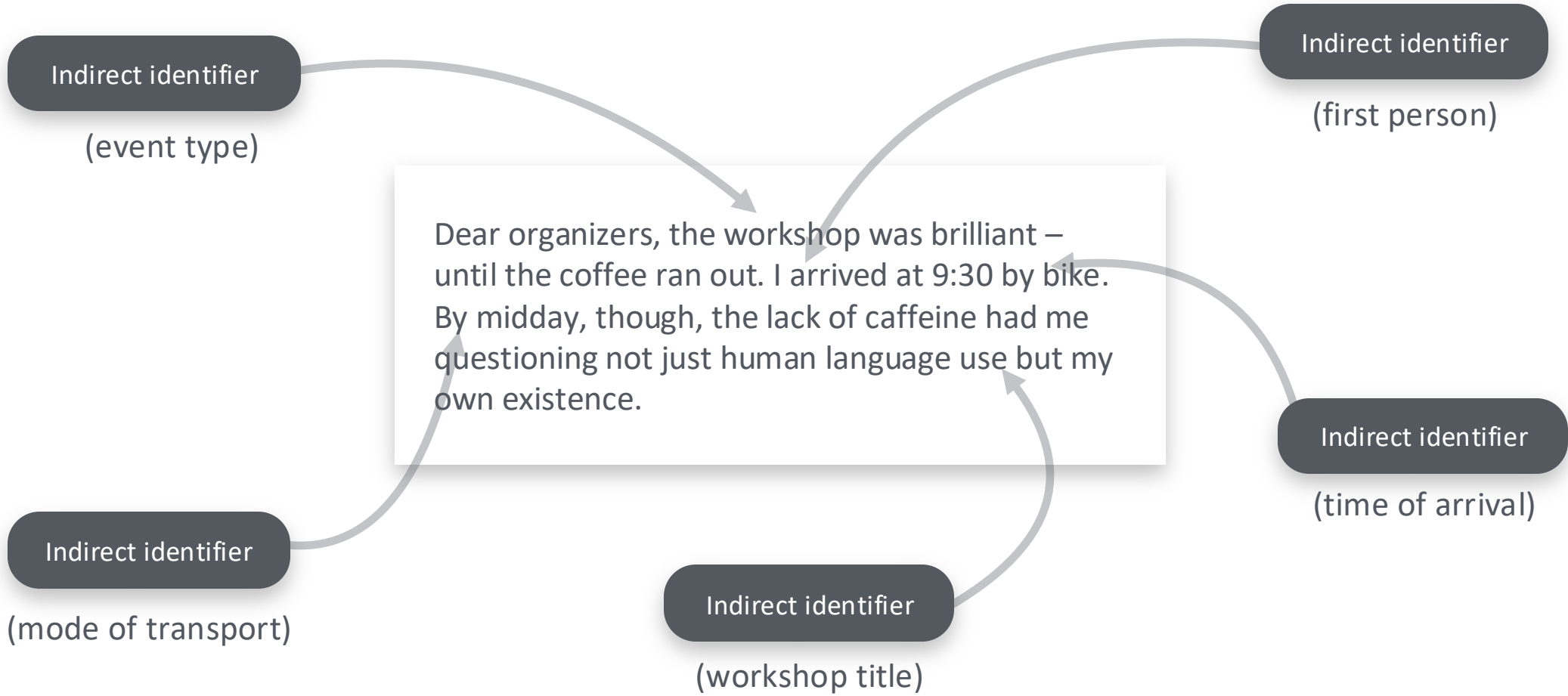
Specific content

[2]

[1] <https://whistleblowersoftware.com/en> [Accessed 29-May-2024]

[2] Bettina Berendt and Stefan Schiffner, 2022. The International Review of Information Ethics.

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.



Related Work

Text Sanitization

NER [1]

NER + Coref-Resolution[2]

Fine-Tuned BERT model for PIs [3]

“On 24 January 2023,
John Smith poured...”

Underestimates (inside and
outside document) context

Any term (not just named entities)
can be identifying [4]

Privacy-preserving data publishing (PPDP)

“On 24 January 2023,
John Smith poured...”

0.234	0.284	0.612	0.137	0.334	0.334
0.103	0.203	0.203	0.403	0.103	0.163
1.452	0.759	1.222	1.432	1.852	1.652
0.654	0.252	0.344	0.754	0.554	0.354
...

Noise

Differential
Privacy

“On January 2023, J.
Smith do...”

e.g. [5]

K-Anonymity

Grammatical errors

Limited variation

[1] Larbi et al., 2022. A Systematic Study on Clinical Text Processing.

[2] Adams et al., 2019. Linköping University Electronic Press.

[3] Kleinberg et al., 2022. arXiv preprint arXiv:2208.13081 (2022).

[4] Arvind Narayanan and Vitaly Shmatikov., 2010. Commun. ACM 53, 6 (2010).

[5] Mattern et al., 2022. Findings of the Association for Computational Linguistics: NAACL 2022.

Related Work

Text Sanitization

Dear ~~organizers~~, the workshop was brilliant – until the coffee ran out. I arrived at ~~9:30~~ by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Privacy-preserving data publishing (PPDP)

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Related Work

Text Sanitization

Dear ~~organizers~~, the workshop was brilliant – until the coffee ran out. I arrived at ~~9:30~~ by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Privacy-preserving data publishing (PPDP)

Dear **coordinators**, a class was great – until the espresso **running out**. I arrived at **10:30** by car. By day, though, the lack from caffeine had me questioning not just human language use but my own existence.

Related Work

Text Sanitization

Dear Dimitri, This is a reminder to the camera-version of your paper by 2 May can prep FAccT 2024.

Privacy-preserving data publishing (PPDP)

, That are tion to send the printer- me from May 2025 such they might bake it for ACL 2024.

Who is to be protected?

What is known about this person?

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I couldn't make it through the rest of the day and had to leave early.

Original (annotate mode)



Dear organizers, the workshop was brilliant– until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I could n't make it through the rest of the day and had to leave early.

Sanitized

Press the *Sanitize* button to start.

Sanitize ✨

Original (write mode)



Dear organizers, the workshop was brilliant– until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I could not make it through the rest of the day and had to leave early.

Sanitized

Press the *Sanitize* button to start.

Sanitize ✨

Home >> Translators

Broken English With A Heavy German Accent Translator

Translate from Normal Language into Broken English With A Heavy German Accent

Normal Language	Broken English With A Heavy German Accent
<p>Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I couldn't</p>	<p>Dear organizers, workshop was sehr brilliant – until the coffee is gone, ja? I come by bike at 9:30, very nice! But by midday, oh, mein Gott, no caffeine make me question not only human language, but my own existence, ja! I try to refill my blue Wasserflasche two times, but no success, ach! I realize, I cannot make it through the rest of the day, so I must leave early, ja.</p>

Translate



Original (write mode)



Dear organizers, workshop was sehr brilliant– until the coffee is gone, ja? I come by bike at 9:30, very nice! But by midday, oh, mein Gott, no caffeine make me question not only human language, but my own existence, ja! I try to refill my blue Wasserflasche two times, but no success, ach! I realize, I can not make it through the rest of the day, so I must leave early, ja.

Sanitized

Sanitize ✨

Risk Mitigation

Anonymization Operations

	Nondescript phrase, e.g. "certain place"	Deletion of dependent phrases	Removal	Deletion of dependent phrases	Nondescript phrase, e.g. "somebody"	LLM rephrasing
Risk	Names of Named Entities	Common Nouns	Modifiers	Proper Nouns	Pronouns	Stylometric Features
High	Suppression	Suppression	Suppression	Suppression	Suppression	Perturbation
Medium	Perturbation	Generalization	Perturbation	Generalization	Suppression	Perturbation
	Zero-weight in LLM generation	Hypernym from WordNet	Zero-weight in LLM generation	Broader Wikidata term	Nondescript phrase, e.g. "somebody"	LLM rephrasing

Risk Mitigation

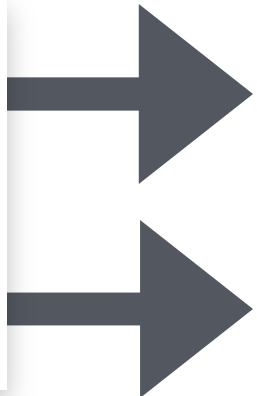
Anonymization Operations

	Nondescript phrase, e.g. "certain time" for "9:30"	Deletion of dependent phrases	Removal	Deletion of dependent phrases	Nondescript phrase, e.g. "somebody"	LLM rephrasing
Risk	Names of Named Entities	Common <small>[SEP]</small> Nouns	Modifiers	Proper <small>[SEP]</small> Nouns	Pronouns	Stylometric Features
High	Suppression	Suppression	Suppression	Suppression	Suppression	Perturbation
Medium	Perturbation	Generalization	Perturbation	Generalization	Suppression	Perturbation
	Zero-weight in LLM generation	Hypernym from WordNet	Zero-weight LLM generation	Broader Wikidata entity, e.g. "water container" for "water bottle"	Nondescript phrase, e.g. "somebody"	LLM rephrasing

Risk Mitigation

Anonymization Operations

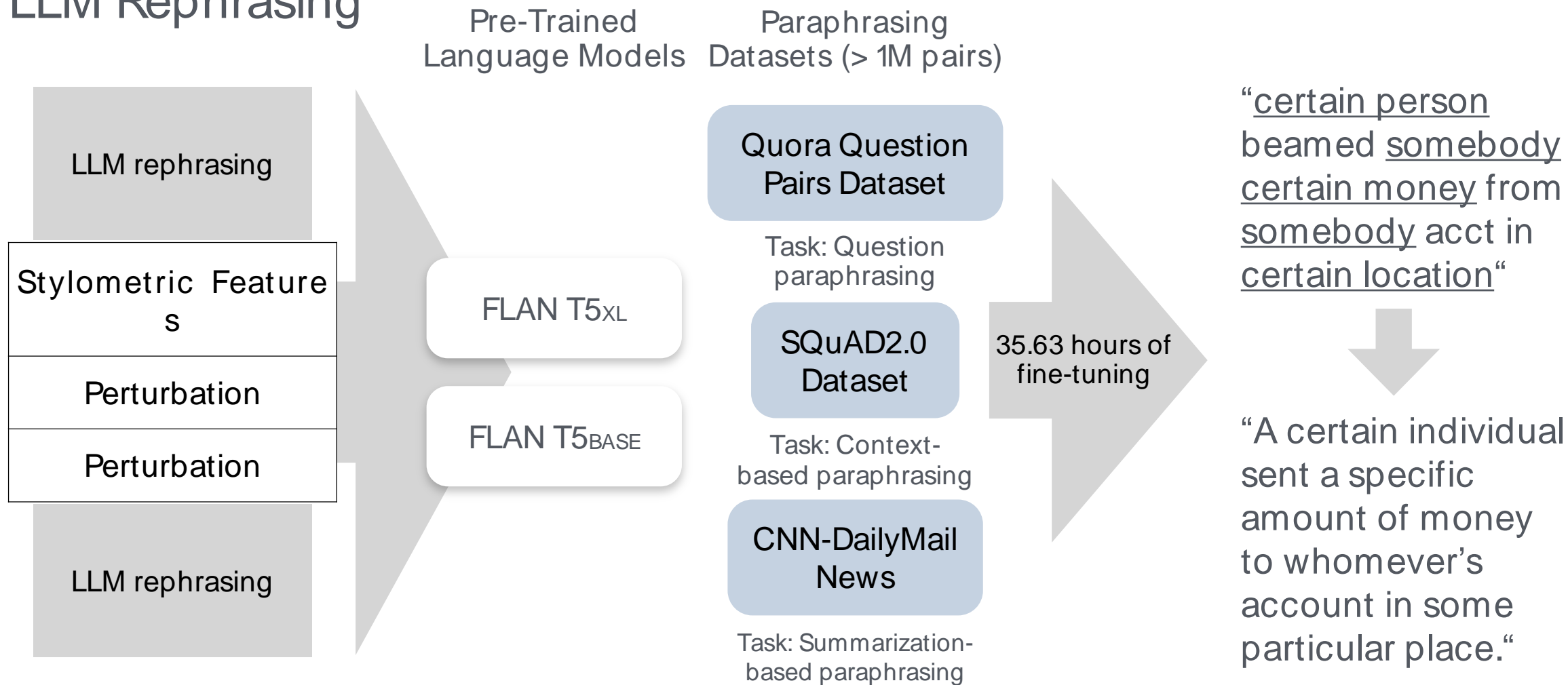
“certain person beamed
somebody certain money from
somebody acct in certain location”



Fragmented text with
grammatical errors

Stylometric features
may not be removed

Risk Mitigation LLM Rephrasing



Evaluation Results



**Text Anonymization
Benchmark (Pilán et al. 2022)**

*Focus: Comparative benchmark for
privacy protection and utility
preservation*



IMDb62 movie reviews dataset

*Focus: Protection against
Authorship Attribution Attacks*



**Whistleblower hearing (Hunter
Biden tax evasion case, 2023)**

*Focus: Qualitative view on
rewritings*

[1] Pilán et al., 2002. Computational Linguistics 48, no. 4.

[2] <https://huggingface.co/datasets/tasksource/imdb62> [Accessed 29-May-2024]

[3] <https://waysandmeans.house.gov/?p=39854458> [Accessed 29-May-2024], “#2”

Evaluation Results

Text Anonymization
Benchmark (Pilán et al. 2022)

Theirs: 0.92

Ours: 0.93

(Quasi identifiers, 1.0 is best)

*Focus: Comparative benchmark for
privacy protection and utility
preservation*

IMDb62 movie reviews dataset

*Focus: Protection against
Authorship Attribution Attacks*

Whistleblower hearing (Hunter
Biden tax evasion case, 2023)

*Focus: Qualitative view on
rewritings*

[1] Pilán et al., 2002. Computational Linguistics 48, no. 4.

[2] <https://huggingface.co/datasets/tasksource/imdb62> [Accessed 29-May-2024]

[3] <https://waysandmeans.house.gov/?p=39854458> [Accessed 29-May-2024], “#2”

Evaluation Results

Text Anonymization
Benchmark (Pilán et al. 2022)

*Focus: Comparative benchmark for
privacy protection and utility
preservation*

IMDb62 movie reviews dataset

Baseline: 98.81%
Ours: 31.22%
(Detection accuracy,
lower = better)

*Focus: Protection against
Authorship Attribution Attacks*

Whistleblower hearing (Hunter
Biden tax evasion case, 2023)

*Focus: Qualitative view on
rewritings*

[1] Pilán et al., 2002. Computational Linguistics 48, no. 4.

[2] <https://huggingface.co/datasets/tasksource/imdb62> [Accessed 29-May-2024]

[3] <https://waysandmeans.house.gov/?p=39854458> [Accessed 29-May-2024], “#2”

Evaluation Results

Text Anonymization
Benchmark (Pilán et al. 2022)

*Focus: Comparative benchmark for
privacy protection and utility
preservation*

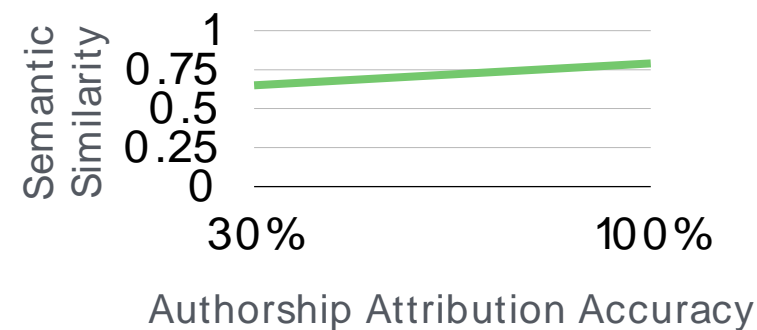
IMDb62 movie reviews dataset

*Focus: Protection against
Authorship Attribution Attacks*

Whistleblower hearing (Hunter
Biden tax evasion case, 2023)

- Removes all identifiers
- Retains overall meaning
- Retains text coherence

Focus: Qualitative view on rewrites



[1] Pilán et al., 2002. Computational Linguistics 48, no. 4.

[2] <https://huggingface.co/datasets/tasksource/imdb62> [Accessed 29-May-2024]

[3] <https://waysandmeans.house.gov/?p=39854458> [Accessed 29-May-2024], “#2”

“Retains overall meaning“

Who is to be protected?

What is known about this person?

+

What is the communicative intention?

What is relevant?

English (British) ▾



Style ▾



Dear organizers, the workshop was brilliant– until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I could not make it through the rest of the day and had to leave early.



Dear Organisers,
I am writing to express my disappointment at the lack of coffee at the workshop. I arrived at 9:30 by bike and by midday, I was experiencing severe caffeine deprivation, which affected my ability to concentrate and comprehend the material being presented. I was unable to continue with the workshop and had to leave early.
I would like to request that you provide coffee and refillable water bottles at future events to ensure the comfort and productivity of all attendees.
Thank you for your attention to this matter.
Sincerely,
[Name]



Original (write mode)



Dear Organisers, I am writing to express my disappointment at the lack of coffee at the workshop. I arrived at 9:30 by bike and by midday, I was experiencing severe caffeine deprivation, which affected my ability to concentrate and comprehend the material being presented. I was unable to continue with the workshop and had to leave early. I would like to request that you provide coffee and refillable water bottles at future events to ensure the comfort and productivity of all attendees. Thank you for your attention to this matter.
Sincerely,[Name]

Sanitized

[Blurred sanitized text]

Sanitize ✨

Communicative Intention (continued)

After speaking with companies that manage whistleblowing reporting channels (translated)

“Around 50% of reports are unsubstantiated, often exaggerated claims against supervisors.”

“We need all unaltered clues, whether in text or dialogue, to assess whether a report is to be taken seriously. Diluting or sanitizing the content makes it harder to detect critical signals.”

“Whistleblower reports are often one person's word against another. Altering the text risks losing key credibility signals, like emotions or repeated details.”

Outlook / questions for discussion

- Can we do more towards fulfilling Grice's maxims?
 - Quantity: be informative
 - Quality: be truthful
 - Clarity: be clear
 - ... or other principles/maxims?
- Other uses cases?