The Ethics of Authorship Verification Programs

When reading a research paper or book, the author is clearly displayed near the title, and even if this information isn't included, you could always search the title to find more information about the work. However, this is not always the case for ancient texts, where they either did not record the author or there are multiple versions of the same text but different authors. Traditional methods of authorship verification consist of close reading for stylistic detail (Juola, 2012), but advances in NLP have added to this toolkit. For example, Sari and Stevenson (2016) demonstrated how word embeddings and ngrams could be used for author clustering.

Ethical Matrix of Authorship Verification Programs

	Autonomy	Well-being	Justice
Researchers (NLP)	Pro: Free to develop and integrate new methods Con: Misuse could prevent ability to get funding for future research	-Useful in historical research - Generative adversarial programs could be used to create fake documents	- (Potentially) contributes to development of open-source NLP software - Could be changed to closed-source and restrict access
Educators	- Contributes to automation of education (Free to do other tasks) - Threatens teacher's autonomy and decision making	- Assists cheat detection - Widespread use/reliance could erode trust	- Promotes fairness in cheat detection - As a machine learning technology, has the potential to learn and propagate biases from training data
Wider society	- Prevent fraud / protects your ideas and published work *- Erosion of trust*	- Provides another tool for forensic investigations and testimony verification - False positive/negative could lead to wrongful incarceration *- Generative adversarial*	- Promotes fairness in forensic investigation - Bad actors could utilize or manipulate the program for unequal verification
Future generations	*- Protection of ideas* *- Erosion of trust*	*- Historical research* - Identification of genuine documents - Counterfeiting may improve and become more difficult to distinguish	*- Fairness in forensic investigation* *- Bias propagation*

Narrative based on matrix:

Suppose you are researching Mencius and find the following quote in a digital database:「孟子曰：欲查其人，先鑑其文」*(lit. "Mencius says: those who desire to check a person, should first verify their texts"). Through an authorship verification program, you could identify this quote as actually deriving from someone's final paper for their Chinese class. Therefore, you are able to correctly disregard this quote for your research.* Authorship verification programs have the potential to enrich historical

research and digital humanities, promote fairness in forensic investigations, and even assist educators through the automation of cheat detection. However, similar to generative adversarial networks, other individuals could generate more convincing fake documents or even manipulate the training data to impact fairness in investigations. The availability of such a technology could also cause a mass erosion of trust, since people (generally speaking) regard technology as infallible and may default to using this technology as a shortcut instead of genuinely investigating the origins of a document. Therefore, in education settings, it could be appropriately implemented similar to how *Turnitin* is currently used, where it is a just a tool at the disposal of the educator. And likewise, in the context of a courtroom, it could be appropriately implemented similar to a blood test or fingerprint analysis, where it is just one piece of a larger set of evidence that is logically evaluated. While open-source development would enable greater transparency in the implementation, closed-source development would protect from individuals developing a generative adversarial program to evade detection.