Academic Integrity in the Age of Artificial Intelligence

Saadia Mahmud
Australian Institute of Business, Australia

A volume in the Advances in Educational Marketing, Administration, and Leadership (AEMAL) Book Series

Published in the United States of America by
IGI Global
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Mahmud, Saadia, 1968- editor.
Title: Academic integrity in the age of artificial intelligence / Edited by Saadia Mahmud.
Description: Hershey, PA : Information Science Reference, [2024] | Includes bibliographical references and index. | Summary: "This book seeks to understand the higher education landscape with the advent of generative AI. More importantly, it seeks to promote a culture of academic integrity in the face of unprecedented pressure to breach academic integrity by members of the academic community"-- Provided by publisher.

Identifiers: LCCN 2023049176 (print) | LCCN 2023049177 (ebook) | ISBN 9798369302408 (hardcover) | ISBN 9798369302415 (paperback) | ISBN 9798369302422 (ebook)
Subjects: LCSH: Artificial intelligence--Educational applications. | Education, Higher--Moral and ethical aspects.
Classification: LCC LB1028.43 .A27 2024 (print) | LCC LB1028.43 (ebook) | DDC 371.9--dc23/eng/20231103
LC record available at https://lccn.loc.gov/2023049176
LC ebook record available at https://lccn.loc.gov/2023049177

This book is published in the IGI Global book series Advances in Educational Marketing, Administration, and Leadership (AEMAL) (ISSN: 2326-9022; eISSN: 2326-9030)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 10
The Inherent Uncertainties of AI-Text Detection and the Implications for Education Institutions: An Overview

Robin Crockett
https://orcid.org/0000-0002-8047-5182
University of Northampton, UK

Robert Howe
https://orcid.org/0000-0002-1344-5530
University of Northampton, UK

ABSTRACT

This chapter focuses on the implications of the improving generative-AI 'chatbot' technologies and the inevitable unreliability of attendant AI-text detection technologies. The goal of generative-AI programmers is to design AIs which produce text indistinguishable from typical human-written text: an eventuality that will render AI-text detectors redundant. The authors outline the underpinning mathematics of AI-generated and human-written text as the basis of AI-text detection, and how this leads to inherent inaccuracies and uncertainties in AI-text detection. The chapter proceeds to overview on how institutions will have to work with both the growth in use of AI and the unreliability of AI-text detection: institutions cannot avoid AI and cannot rely on 'tech' to police it. Students need to be taught how to use AIs ethically with integrity and insight and sanctioned when they do not. At the same time, institutions need to resource people to investigate students suspected of false authorship, whether commissioning a human ghost-writer or using an AI inappropriately.

DOI: 10.4018/979-8-3693-0240-8.ch010

INTRODUCTION

This chapter considers the detection and classification of text produced by generative artificial intelligences (i.e., generative AIs, gAIs), with some extension into other generative AI capabilities (Ibrahim et al, 2023; New Scientist, 2023). It does not consider artificial general intelligence (AGI). A "generative" AI is one that humans can instruct and interrogate. Generative AI supports the creation of text, images and other media using existing sources and rules by which this knowledge may be combined in new ways giving new insights. Due to the wide range of sources which are aggregated, current and emerging AI systems are impacting on many areas of society. There is a range of literature (Lim et al, 2023; Yu and Guo, 2023; Nah et al., 2023), which provides a comprehensive background to gAI but it should be noted that a detailed understanding of these systems is not needed to appreciate the detection processes – and problems therewith – covered in this chapter. Jisc (2023c) has also developed an introduction to gAI to help institutions better understand the current range of tools which are categorised as gAI. The issue being covered in this chapter is the fact that some students are using gAI tools to create some or all of the content for their assignments, sometimes in ways which constitute academic misconduct, and the inherent unreliability of AI-text detectors in the academic integrity context. For clarity to avoid confusion with use of the verb "to generate" and the noun "generation", all references to AIs hereinafter are to generative AIs.

The problem of students submitting unauthorised AI-generated text first manifested in any significant way with OpenAI's launch of GPT-3 in 2020 (Brown et al., 2020; Crompton & Burke, 2023) and the subsequent appearance of numerous essay bots (Crockett, 2023a), many operated by essay mills, that explicitly targeted students seeking to cheat assignments. However, the problem became immeasurably bigger and more significant with OpenAI's release of ChatGPT in November 2022 (OpenAI, 2022), with its vastly improved user interface, e.g., the capability to "chat", as well as its improved and refined algorithms and larger training datasets. That has been followed by, for example, OpenAI's GPT-4 and Google's Bard in March 2023, Anthropic's Claude-2 in July 2023 (OpenAI, 2023b; Google, 2023a&b; Anthropic, 2023).

In response to these technological developments, numerous AI-text detectors/classifiers have been released by a variety of technology and educational technology companies (for example, not an exclusive list: Copyleaks, 2023; Crossplag, 2023; OpenAI, 2023a; Tayeb, 2023 (Draft & Goal); Tian, 2023a (GPTZero); Turnitin, 2023). These all work using essentially the same underpinning mathematics and rely on the intrinsic stylistic differences between typical human-written text and typical AI-generated text. Those differences arise from the intrinsic differences between the ways which human writers actually write and the ways AIs generate text according to their programming.

AIs programmed to generate text, such as ChatGPT, Bard and Claude, sometimes referred to as "chatbots", are based on large language models (LLMs), and the chapter opens with a brief introduction into the way those work. That is followed by a consideration of the differences between AI-generated and human-written text that underpin AI-text detection, including a brief outline of the underpinning mathematics to introduce the terminology commonly encountered on websites, in the academic literature and other reporting. That leads directly to the core of the chapter: i.e., the functionality of AI-text detectors/classifiers and the inherent uncertainties that give rise to the observed and reported unreliabilities (Anderson, 2023; Chakraborty et al., 2023; Ibrahim et al., 2023; Liang et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023). The chapter concludes with an overview of the challenges all this poses

for course and assessment design, and challenges faced by learning technologists in trying to meet the emerging and sometimes conflicting demands.

GENERATIVE AIS AND LARGE LANGUAGE MODELS

The AIs in question are based on large language models (LLMs) (Shanahan, 2023). A language model can be defined as a mathematical/statistical/probability model of natural language, i.e., human-written, texts (DeepLearningAI, 2023). Those texts can be modelled in terms of individual words or phrases or, indeed, parts of words (e.g., stems and affixes) and punctuation, i.e., what are referred to as "tokens". An LLM, such as used by OpenAI for its GPT variants (Generative Pre-trained Transformer), Google for Bard (LaMDA, Language Model for Dialogue Applications; PaLM, Pathways Language Model) and Anthropic for its Claude variants, is a (very) large version of the above, typically using vast numbers of texts scraped from the Internet (Anthropic, 2023; Google, 2023a&b; OpenAI, 2022 & 2023b).

For an individual AI, its artificial-intelligence machine-learning algorithms analyse its LLM dataset to determine patterns of language usage and characterise relationships between words/phrases and the ideas and concepts that it "sees" these as being semantically and/or syntactically linked to (DeepLearningAI, 2023; Hahn, 2023; Havlik, 2023). From there, according to its programming, an artificial neural network (ANN) is constructed: it is that ANN that is at the core of the predictive and generative capabilities.

There is a variety of definitions for AI but a helpful definition in the current context is "the branch of computer science involved with the design of computers or other programmed mechanical devices having the capacity to imitate human intelligence and thought" (Dictionary.com, 2023). However, let us be clear: current AIs (latter half of 2023) are not intelligent in a human-like way and whilst AIs are programmed to generate text that appears human-like, and do so increasingly successfully, the text-generation process does not (currently) resemble a typical human writing process, see the illustrative analogy in this section.

If you ask a human to write an answer to a question, they typically think about what they are writing as they write, whereas if you prompt an AI to generate some text or ask it a question, it does not think about and write an answer. Instead, it parses the prompt/question and searches for the most probable, most commonly-occurring words in its dataset and joins those together to generate grammatically correct but not necessarily subject-fluent text, all according to its programming (Heikkila, 2022; Shanahan, 2023). Every word in the generated text depends on the prompt(s) and the preceding generated words. Every word, every punctuation mark in the generated text occurs somewhere in the LLM dataset, with sentence and paragraph breaks determined algorithmically.

Thus, if you ask an AI a question, what you are in effect doing is giving the LLM software an instruction along the lines (quoting Shanahan, 2023): "Here's a fragment of text. Tell me how this fragment might go on. According to your model of the statistics of human language, what words are likely to come next?" For example, if you enter "Humpty Dumpty" then you are effectively asking the software which words commonly associate with "Humpty Dumpty" in its model, and the expectation is that the LLM will return "sat on a wall", possibly followed by the remainder of the nursery rhyme.

That is very different to how a human typically writes text. In essence, when a human writes, they express their thoughts in the best way for the task in hand, choosing the best vocabulary and phrase/sentence/paragraph structure etc. according to the subject matter and their personal stylistic preferences, often reviewing and editing as they write. The resulting text reflects that individuality of writing style:

indeed, such considerations with regard to individual writing style lie at the core of stylometry (Crockett, 2023a; Crockett & Best, 2020; Fox et al., 2012; Juola, 2013). In contrast, because of the way AIs work in always supplying high-probability previously written text that aligns with the prompts, AI generated text is predictable, with sentence and paragraph breaks determined algorithmically and not necessarily according to immediate changes in subject emphasis, lacking individuality and generally bland, monotone and unvaried in comparison to typical human-written text (Goom, 2023; Reed, 2023). That can sometimes be easy to observe but not always, hence the initial stimulus for the development of AI-text detectors/classifiers. Subsequently, human-observation has generally been becoming more difficult as the AIs improve, also posing increasing challenges for AI-text detectors/classifiers. Furthermore, AIs generally have an element of randomisation coded-in such that if a user clicks to regenerate the text, or another user enters the same prompt, the AI outputs further text that describes the same content in the same underlying way as the previous output, but with different wording, phrasing, sentences, and sometimes different content-precedence. That, coupled with continuing improvements in the AIs, makes it increasingly impossible for tutors, for example, to reproduce AI-text even if they know the specific AI used by a student in question, and the specific prompts used.

In order to help illustrate the AI text generation process, consider the loose analogy described as follows. A person who knows how to write grammatically correct text is tasked with writing about a subject of which they have no knowledge, according to the following method:

i. Search the Internet for the most common opening phrase associated with the subject.
ii. Search the Internet for the words/phrases that most commonly follow that.
iii. Repeat (ii), incorporating grammatically correct punctuation, until the opening sentence is grammatically complete.
iv. Search the Internet for the most common opening phrase for a following sentence.
v. Search the Internet for the words/phrases that most commonly follow that.
vi. Repeat (v), incorporating grammatically correct punctuation, until that sentence is grammatically complete.
vii. Repeat (iv)-(vi), without reviewing previously written sentences, until the body of text is complete in grammatical terms even if not in subject terms, separating groups of sentences into paragraphs at grammatically appropriate points.

A process such as this is very different to normal human writing but it would yield a body of text that is grammatically correct. However, because the writer is inferring correctness and consistency of content as aligning with the most commonly associated words/phrases that appear on the Internet, it would not necessarily be correct or consistent with regard to the specified subject. Also, that reliance on the most commonly associated words/phrases will result in an absence of individuality (Heikkila, 2022).

Further illustration was obtained by providing Claude-2 (Anthropic, 2023) with the user-prompt "Describe ChatGPT's writing style". The output is [verbatim]:

I don't actually have detailed insights into ChatGPT's writing style, since I'm a different AI created by Anthropic. However, from what I've seen, ChatGPT tends to produce very natural and human-like responses by training on a massive dataset of online dialog. It aims to sound conversational and articulate while providing useful information to users. Some key aspects of its style seem to be:

- *Uses common speech patterns and informal language*
- *Varies sentence structure and length to sound more natural*
- *Incorporates facts and data seamlessly into fluent explanations*
- *Adapts tone based on the prompt and conversation context*
- *Avoids overtly sounding like a bot through the use of opinions, quirky remarks etc.*
- *Seeks to continue conversations and engage users through follow-up questions and remarks*

Of course, this is just my impression based on limited interactions and observations. ChatGPT's capabilities and output will likely continue evolving over time as Anthropic refines its model. But overall it tends to exhibit a flexible, human-like narrative voice in service of delivering informative and satisfying responses. (Anthropic, 2023)

Briefly interpreting Claude-2's output. First, there is the "hallucination" that ChatGPT will evolve as Anthropic refines its model, which is not correct: ChatGPT is OpenAI, not Anthropic. An AI hallucination is text (in this context, but also potentially images etc.) that does not reflect reality, arising from the AI addressing a user prompt according to its programming, and joining sections of high-probability text but without any understanding or subject-related error-checking (Bender et al., 2021; Ji et al., 2023). This highlights an important aspect of AIs' programming, at least with the current generation of AIs (latter half of 2023): in essence, AIs conflate correctness with the (high) probability of occurrence in their datasets and so can erroneously link incorrect and/or individually correct but inconsistent sections of text in their outputs. Second, the statements regarding "massive dataset", "common speech patterns", are basic statements reflecting the design and programming as are, to varying extents, the references to fluency and varying sentence structure. Third, it includes the statement "incorporation of facts and data" but does not state that those are necessarily appropriate, correct or consistent and are not hallucinatory: even "opinions" and "quirky remarks" will be generated from the dataset according to the programming, with no guarantee of correctness.

In a different context, there is a good illustration of the pitfalls of uninsightful use of AIs and AI hallucinations reported in July 2023 by Retraction Watch concerning a pre-print posted on Preprints.org (Retraction Watch, 2023). This appears to be research fraud rather than academic misconduct, but it nevertheless illustrates what can happen if a determined user is prepared to prompt and re-prompt an AI in order to obtain highly detailed text, but then not to check in sufficient detail before proceeding, and the Retraction-Watch narrative is informative with regard to several aspects of misuse of AIs.

In summary, an AI outputs the most probable words/phrases according to its programming, and there is no guarantee that facts, data, information, opinions or remarks are relevant or correct in context, and there is no guarantee that the text will not contain hallucinations. AIs are not currently programmed to either understand the text they generate or to error-check for veracity (Heikkila, 2022). In contrast, human writers can typically vary their writing based on their knowledge and understanding of the subject (noting that there are atypical human writers who do not do this (Liang et al., 2023)) and, importantly, can check and verify that facts, data, information, opinions and remarks are relevant and correct in context as their subject knowledge and understanding develop during the writing process.

Thus, hallucinations can be very reliable tell-tales of AI-generated text, and a common and easily-detectable type of hallucination is invented/faked citations and references (where the AI has been tasked to generate referenced text). In the academic integrity context, hallucinations are examples of fabrication/

falsification of data/information and thus represent serious academic misconduct irrespective of whether the fabrication/falsification is attributable to AI text generation or human writing. Furthermore, it must be remembered that it is probable that some hallucinations in AI-generated text will propagate. This is because some hallucinations will appear in documents that are collated for the AI training datasets when those datasets are updated. This will have the unfortunate consequence that, for example, a fabricated citation/reference currently identifiable using text-matching software, which will not flag that citation as similar as it is not present in the software's database, could be incorporated into an updated database. That, in turn, will mean that both staff and students will have to be more careful in their checking of data and information as they undertake research.

Finally, before considering AI-text detection/classification, it is possible to prompt AIs to produce more human-like text, i.e., "smart prompting" (Lu et al., 2023; Sadavisan et al., 2023) and it is also possible to "watermark" AI-generated text by programming the AI with whitelists and blacklists of words to use frequently and infrequently respectively. Whilst watermarking itself is detectable, Sadasivan et al. (2023) and others have shown that putting such text through a paraphrasing or rewriting tool can very effectively reduce or remove the watermarking tell-tales: this is in addition to the documented use of such tools to reduce evidence of "ordinary" plagiarism (Rogerson & McCarthy, 2017). Thus, all that someone trying to render AI-generated less detectable has to do is put the AI-generated text through a readily available paraphrasing/rewriting tool, and then perhaps check with one or more online AI-text detectors/classifiers (noting that AI-text detectors/classifiers are inherently unreliable, see following sections). In that context, Phrasly (2023) explicitly targets students with the lure "Transform AI-generated content into undetectable text with Phrasly. Save time on your assignments, boost grades, and easily bypass AI detectors like TurnItIn and GPTZero". Consequently, AI-text detectors/classifiers that rely on watermarking are particularly vulnerable to paraphrasing/rewriting tools. However, it should be noted that as well as reducing/removing watermarking, putting AI-generated text through such tools can change the style and use-of-language, e.g., grammar and vocabulary, meaning that all AI-text detectors/classifiers have some degree of vulnerability to such tools. Furthermore, this situation will not improve: the next generation(s) of AI-equipped paraphrasing/rewriting tools will produce text that is more human-like and more effectively stripped of watermarking than the current/preceding generations of such tools.

AI-TEXT DETECTION/CLASSIFICATION

Leaving hallucinations and watermarking aside, it is these inherent stylistic differences that underpin the functionality of AI-text detectors/classifiers, but that functionality is clearly not definitive, and never can be. For example, some humans sometimes write in AI like ways (Liang et al., 2023). Conversely, a careful AI user with subject knowledge can smart-prompt an AI to vary its default output style, and putting AI-generated text through a paraphrasing or rewriting tool can reduce its "AI-ness" thereby obfuscating the evidence. However, over-riding all such individual factors is the ongoing improvement of the AIs: each new AI generation produces text that is more human-like than the previous generation, further blurring the distinctions between AI-generated and human-written text, thereby increasing the uncertainty of AI-text detection.

There is subject jargon in the AI-text detection context with some commonly encountered terms it is useful to be aware of. The stylistic factors outlined in the preceding section can be characterised via the text "entropy" (or "information entropy" – broadly, the unpredictability of the words, see next section),

which can be quantified in various ways (Brownlee, 2019; Snow et al., 2016; Tian, 2023b). The basic underpinning quantity is the (information) entropy, but natural language processing (NLP) programmers designing the AI-text generators and detectors/classifiers generally use the related quantity of "perplexity" (Bernstein, 2021; Tian, 2023b). Both of these are related to the "surprisal" of words, i.e., how unpredictable or surprising they are. In outline, words that occur frequently (common, unsurprising words) are low-surprisal and words that occur infrequently (uncommon, surprising words) are high-surprisal, and entropy and perplexity are measures of average surprisal (Crockett, 2023a; Rogers, 2023). There is also "burstiness": there are several definitions of this but, broadly, it is a measure of the variation of entropy over a whole document, often quantified in terms of variations over sentences (or other "chunks" of text), and is also related to surprisal (Tian, 2023b).

Underpinning Mathematics

The details of the mathematics that underpins the functionality of AI-text detectors/classifiers are beyond the scope of this chapter, and it is not necessary to understand the mathematics in order to appreciate the problems inherent to AI-text detection/classification, but it is useful to develop a little of the theory in order to establish the context of the terms and concepts commonly encountered on AI-text detector/classifier and related websites and in the literature, as introduced in the preceding section.

In his landmark analyses, Shannon (1948, 1951) set out the framework for information entropy and defined the entropy of (a body of) text as the expectation of the surprisal of the symbols (i.e., words in the current context) used to represent that text. Surprisal (self-information, Shannon information) is a key quantity which has useful mathematical properties in the context of information theory and natural language processing.

Therefore, customising Shannon's definitions and notation for the current context, the surprisal, *SI*, of a word *x* in a (body of) text, *X*, is defined as

$$SI\left(x\right) = log_2\left(\frac{1}{p\left(x\right)}\right) = -log_2\left(p\left(x\right)\right)$$

where p(*x*) is the probability of *x* occurring in *X*.

Thus, surprisal as the negative log-probability is an inverse measure of probability, i.e., it increases as probability decreases and thus is a measure of unpredictability or surprise.

From there, the entropy, *H*, of the overall (body of) text *X* is defined as

$$H\left(X\right) = \sum\nolimits_{x \in X} p\left(x\right) SI\left(x\right)$$

i.e., the entropy is the expectation of the surprisal.

For surprisal and entropy as above, the perplexity, *PP*, is

$$PP\left(X\right) = 2^{H(X)}$$

There are various definitions of burstiness but the most accessible one is the Fano factor. For a sentence, or other chunk of text, *X*, the Fano factor, *F*, is the ratio of the variance to the mean of the word surprisals, SI_x, i.e.

$$F(X) = \frac{\sigma^2_{SIx}}{\mu_{SIx}}$$

These are core definitions which should help with interpreting material on and about detector websites. However, it should be noted that different programmers use variations of these basic quantities in trying to develop effective and reliable AI detectors/classifiers, and, thus, different but related statements and definitions might be encountered with regard to specific detectors/classifiers depending on their programming.

Relating this to linguistics, there are two basic categories of words, i.e., content words (also referred to as lexical or open-class words, e.g., words such as main nouns and adjectives, main verbs and adverbs) and stop words (also referred to as function or closed-class words, e.g., structural/grammatical words such as articles, conjunctions, prepositions, pronouns, particles, helping verbs). Content words carry the subject content and are generally high-surprisal, whereas stop words carry the grammar between the content words and are generally low-surprisal (Kermes & Teich, 2017; Lancaster University, n.d.; UCL, n.d.). Thus, text containing relatively high and low proportions of stop-words and content-words respectively, as typifies AI-generated text, will be relatively low-entropy, low-perplexity. Conversely, text containing relatively low and high proportions of stop-words and content-words respectively, as typifies human-written text, will be relatively high-entropy, high-perplexity. Extending this reasoning, text consisting of relatively uniformly structured sentences containing a relatively high proportion of stop words plus commonly-used "everyday" content words, as typifies AI-generated text, will be relatively low-burstiness, whereas text consisting of more variably structured sentences containing a relatively low proportion of stop words plus a relatively varied set of content words, as typifies human-written text, will be relatively high-burstiness.

Thus, human-written text, with its innate stylistic individuality, is typically relatively high-surprisal, high-entropy, high-perplexity, high-burstiness. Conversely, AI-generated text, due to an AI being programmed to select the most probable words/phrases from its dataset, i.e., the most commonly used, most everyday words/phrases, and to generate sentences and paragraphs algorithmically rather than according to the immediate subject matter, is typically relatively low-surprisal, low-entropy, low-perplexity, low-burstiness (Anderson, 2023).

Illustrative AI-Text Classification

For clarity, in this section, the term "entropy" is used to include entropy, perplexity, burstiness and any related/derived quantities that are used in AI-text detection/classification.

The detection/classification of a text as "AI" or "human" (sometimes expressed as, in terms, "not-AI") depends (a) on that text being either a typical AI text or a typical human text and the less typical in either sense, the less reliable the classification, and (b) on typical AI texts being distinguishable from typical human texts. However, those dependencies cannot be guaranteed: some AI-generated texts are more human-like, i.e., with higher-than-typical entropy compared to the majority of AI-generated text,

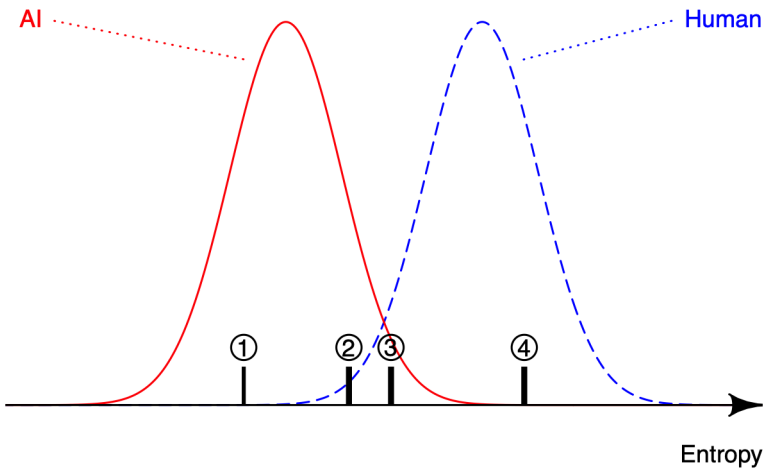
and some human-written texts are more AI-like, i.e., with lower-than-typical entropy compared to the majority of human-written text. This is a real concern: Snow et al. (2016) observe that lower-proficiency students write with more rigidity and less flexibility, which manifests as lower text-entropy, i.e., text that is more AI-like. Thus, in general, there are systematic overlaps between the statistical distributions of entropy statistics from AI-generated and human-written texts. That overlap means that there is always a systematic uncertainty in AI-text detection/classification because an AI-text detector/classifier cannot simply make a determination with reference to an AI-entropy distribution but must also take account of a human-entropy distribution. In essence, it is not a simple case of determining how typically AI-like a piece of text is because any determination also has to account for how typically human-like that piece of text is.

As a consequence of this, AI-text detectors/classifiers are generally configured with detection thresholds, and those detection thresholds are generally set to minimise false-positive probabilities (rates). A false positive is human-written text incorrectly classified as AI-generated; a false negative is AI-generated text incorrectly classified as human-written. It is not possible to set a detection threshold to minimise both the false-positive and false-negative rates simultaneously: indeed, minimising one necessarily maximises the other and, therefore, AI-text detectors/classifiers configured with minimised false-positive rates effectively have maximised false-negative rates (Anderson, 2023; Crockett, 2023b). The small-scale false-negatives investigation reported by Crockett (2023b) demonstrates two important aspects of AI-text detector/classifier performance. First, a detector/classifier only achieves its claimed accuracies for text obtained from an AI-text generator that is explicitly listed by that detector, and even then only if the AI generator was not smart-prompted: accuracies for other categories of AI-generated text are significantly worse. Second, and only for an AI text generator that is explicitly named/listed, a detector/classifier that is "tuned" to minimise its false-positive rate necessarily has a significantly worse false-negative rate with observed false-negative rates of ca. 1-in-4 corresponding to claimed false-positive rates of ca. 1-in-100. That is important: in terms of academic misconduct, any false-negative is potentially a cheated student submission going undetected – "slipping under the radar" – which compromises our abilities to (a) help any such student study with integrity, and (b) protect our institutional integrity and reputation.

Whether explicitly or implicitly, whether as a direct statistical evaluation or as an indirect AI-enabled evaluation where the AI's training – calibration – has taken account of the necessary statistical criteria, AI-text detection/classification involves a complex assessment of entropy. However, a loose but nonetheless informative analogy is to think of AI classification in terms of a Mann-Whitney U-test (McClenaghan, 2022), as follows. A hypothetical AI-text detector has very large reference samples of entropy statistics from both AI-generated and human-written texts and it U-tests the entropy statistics sample from the text in question against its reference samples. That will yield two p-values, one for "AI-ness" and one for "human-ness". Thus, and noting "strong" in the following will depend on the setting of the detection threshold:

- if the AI p-value is strong null-hypothesis and the human p-value is strong alternative hypothesis, then the text in question is classified as "AI-generated";
- if the human p-value is strong null-hypothesis and the AI p-value is strong alternative hypothesis, then the text in question is classified as "human written";
- if one or both of the p-values is insufficiently strong, the classification is determined by the threshold and is not a simple "which p-value is stronger" decision and, for example, if the threshold is set to minimise the false-positive rate, the classification will be "human written".

Figure 1. Illustrative entropy distributions. The AI distribution is shown as a solid/red line, the human distribution is shown as a dashed/blue line.



This is illustrated in Figures 1, 2 and 3. In these figures:

- the shapes of the AI-generated and human-written entropy distributions are arbitrary and are chosen for clarity and should not be interpreted as actual entropy distributions;
- the AI-generated and human-written entropy distributions overlap, reflecting the reality that some AI-generated text is atypically human-like, and some human-written text is atypically AI-like.

The basic situation is illustrated in Figure 1.

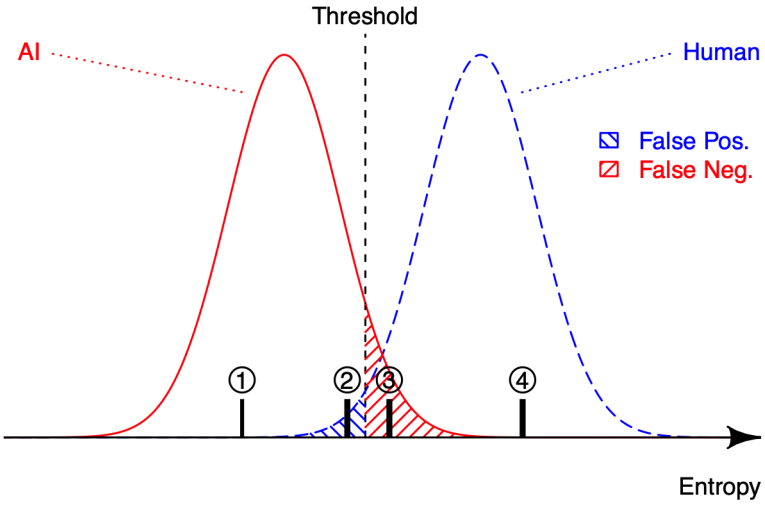
Figure 1 shows four hypothetical entropy statistics (solid vertical bars, labelled 1-4) in relation to the entropy distributions for AI-generated and human-written text. With regard to Figure 1:

- statistic 1 is central within the AI-text entropy distribution (i.e., high probability), and a long way along the tail of the human-text entropy distribution (i.e., low probability), and is therefore classified as "AI-generated" with a high degree of confidence;
- statistic 4 is a long way along the tail of the AI-text entropy distribution (i.e., low probability), and central within the human-text entropy distribution (i.e., high probability), and is therefore classified as "human-written" with a high degree of confidence;
- statistics 2 and 3 are ambiguous, falling in the overlap regions of the two distributions.

In order to classify statistics 2 and 3, a threshold is required as described in the preceding section. This is illustrated in Figure 2.

Figure 2 clearly shows that there is more of the AI-distribution area (probability) on the human side of threshold than there is human-distribution area (probability) on the AI side of the threshold, clearly illustrating that minimising the false-positive probability maximises the false-negative probability, and vice-verse if the threshold had been set to minimise the false-negative probability. With regard to Figure 2:

Figure 2. Illustrative entropy distributions, as Figure 1, with classification threshold set to minimise false-positive rate (probability). The false-positive and false-negative probabilities are shown as shaded areas: false positive shaded leftwards-up/blue, false negative shaded rightwards-up/red.



- statistic 2 is on the AI side (left) of the threshold, so is classified as "AI-generated" but clearly with less confidence than statistic 1;
- statistic 3 is on the human side (right) of the threshold, so is classified as "human-written" but clearly with less confidence than statistic 4.

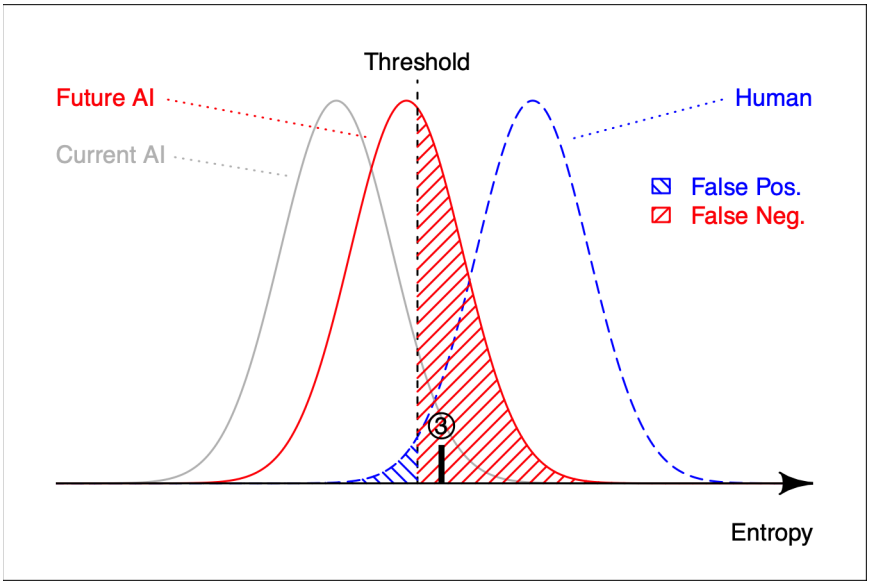
Thus, if statistic 2 arises from AI-generated text, it is classified correctly, however if it arises from human-written text, it is incorrectly classified and is a false positive. Similarly, if statistic 3 arises from human-written text, it is classified correctly, however if it arises from AI-generated text, it is incorrectly classified and is a false negative.

This illustrates the fundamental uncertainty inherent to AI-text detection/classification. Furthermore, if Figures 1 and 2 are regarded as reflecting the current situation with the current generation of AIs, the next generation of AIs will mean that AI-generated text entropies further overlap with human-written text entropies and the AI-text entropy distribution will overlap the human-text entropy distribution to greater extent. This is shown in Figure 3.

Interpreting Figure 3, it is clear that as AI-generated text becomes more human-like, the overlap between the entropy distributions increases, meaning fewer distinguishable differences. It is also clear that keeping the same classification threshold to maintain the same false-positive probability (area shaded leftwards-up/blue) necessitates an increased false-negative probability (area shaded rightwards-up/red). Also consider that if the previous example statistic 3 were to arise in conjunction with this future generation of AIs, it would again be classified as "human-written" despite falling close to the centre of the AI distribution and, therefore, having greater actual probability of being AI-generated than in Figure 2.

Thus, however reliable the current generation of AI-text detectors/classifiers becomes with improved programming and larger datasets (as more people use AI-text generators and thus provide data), these detectors/classifiers will still be less reliable with the next generation of AIs that will produce more

Figure 3. Illustrative entropy distributions: future AI. Future AI distribution shown as a solid/red line, human distribution (same as Figures 1, 2) as a dashed/blue line, current AI entropy distribution (same as Figures 1, 2) shown in grey for comparison



human-like text than the current generation (AI Writing Check, 2023; Anderson, 2023). Also, while custom, focused tools such as that reported by Desaire et al. (2023) for chemistry can perform better than general AI-text detectors in narrowly specified contexts, the probable overfitting and/or overtraining involved significantly constrain the applicability. Furthermore, such focused tools do not address the fundamental issues associated with AIs continuing to improve and produce ever more human-like text.

Even if it transpires that AI-text detectors/classifiers continue to improve (Chakraborty et al., 2023), both within AI-generations and from AI-generation to AI-generation, the goal of (many) AI designers and programmers is to produce AIs that generate text that is indistinguishable from human-written text (Nvidia, 2023). In that eventuality, AI-text detectors will become redundant. Even with the current generation of AIs (latter half of 2023), AI-text detection/classification is unreliable (Crockett, 2023b; Orenstrakh et al., 2023; Sadasivan et al., 2023; Weber-Wulff et al., 2023). Also, OpenAI withdrew their AI-text classifier on 20 July 2023, stating:

As of July 20, 2023, the AI classifier is no longer available due to its low rate of accuracy. We are working to incorporate feedback and are currently researching more effective provenance techniques for text, and have made a commitment to develop and deploy mechanisms that enable users to understand if audio or visual content is AI-generated. (OpenAI, 2023c).

That action by OpenAI explicitly raises the question that if OpenAI – the designers of ChatGPT, with all their (inside) knowledge of AI-functionality – cannot make a reliable AI-text detector, even for their own AI, is it reasonable to expect other AI-text detectors/classifiers to be any more accurate and reliable? Also, there is OpenAI's FAQ which in response to the listed question "Do AI detectors work?", states "In short, no. While some (including OpenAI) have released tools that purport to detect AI-generated

content, none of these have proven to reliably distinguish between AI-generated and human-generated content" before further statements which augment and clarify that statement (OpenAI, 2023d).

IMPLICATIONS FOR ACADEMIC INTEGRITY AND TEACHING

To summarise, AI-text detection/classification is currently inherently unreliable (Al-Sibai, 2023; Anderson, 2023; Crockett, 2023b; Tangermann, 2023; Williams, 2023). Furthermore, this inherent unreliability will become increasingly significant as AIs improve and AI-generated text becomes more human-like with (possibly) fewer hallucinations (Bender et al., 2021). In that context, Atleson (2023) on the USA Federal Trade Commission's business blog has written:

If you're interested in tools to help detect if you're getting the good turtle soup or merely the mock, take claims about those tools with a few megabytes of salt. Overconfidence that you've caught all the fakes and missed none can hurt both you and those who may be unfairly accused, including job applicants and students.

Also, even if watermarking becomes more robust, paraphrasing and rewriting tools will also improve and will more effectively alter the text vocabulary and structure, obfuscating the evidence by reducing and possibly completely removing the watermarking, as well as altering the entropy-related properties. Coupled with that, it is safe to observe that as people discover ways of disguising AI-generated text, "How to beat AI detectors" webpages will proliferate, similar to the proliferation of "How to beat text-matching software" webpages some years ago. Such webpages already exist, and not all necessarily contain reliable advice, just as not all advice with regard to beating text-matching software was reliable. Without making any judgement or endorsement on either of these see, for example, articles by Gluska (2023) and Khan (2023) as indicators of what is already available to students and others who search the Internet for advice. Lastly, there is recent legal opinion from Gaumann and Veale (2023) with regard to the legal position of generative AIs in the context of academic integrity and whether such AIs can be regarded as essay mills.

The consequence of this is that education institutions cannot rely on AI-text detectors/classifiers to (help) police academic integrity (Chan, 2023; Wilhelm, 2023): simply, for the reasons outlined above, AI detectors are not reliable for such purposes and never will be. It should also be noted that expert human "classifiers" cannot reliably detect AI-generated text either (Casal & Kessler, 2023). Thus, it is not meaningful for education institutions to announce bans on the use of AIs for assessments where the students control the environments in which they prepare their submissions.

Therefore, the role of the tutors will be key, but also very difficult at least in the next few years while institutions are establishing their policies and guidance and individual tutors experiment with and assess AI capabilities. However, it is already clear that assessments where use of AIs is not permitted will need to be conducted in institutionally-controlled environments, such as in-class assessments and formal examinations, and a degree of redesign of assessments is probable. Conversely, in this new environment, education institutions need to teach their students how to use AIs ethically, with integrity and insight, but cannot rely on AI-text detectors/classifiers to ensure that students actually use AI text generators for assessments where use of AIs is mandated. Consequently, there is need for assessment strategies to be reviewed, including the balance between subject understanding and evaluation on the one hand and

abilities to access, collate and present information on the other. One possibility with regard to assessment design, in overview, is to actively encourage the use of AI tools but with an assessment focus on how the tools contribute to knowledge and understanding, i.e., learning outcomes that explicitly require students to critically evaluate and analyse subject material rather than more straightforward collation and presentation, because collation and presentation are easily achieved via AI. Thus, as a summary overview statement, institutions should ensure that they are not setting assignments which mean that simple, basic "thought-free" use of an AI tool can reliably meet the learning outcomes (AdvanceHE, 2023; Bowditch, 2023; Jisc, 2023d).

The development of AI tools presents education institutions with opportunities and positive challenges: it is far from being all negative, but it is going to be challenging. Let us consider the reality that current and future students are going to graduate into workplaces and environments where use of such AIs is increasingly the norm (Sabzalieva & Valentini, 2023; Salvagno et al., 2023). Therefore, students who have not been taught how to use AIs ethically, with insight and integrity, might well find themselves at a disadvantage compared to those who have (Jisc, 2023a; Lancaster, 2023; Wang et al., 2023). Therefore, education institutions need to incorporate use of generative AIs into their syllabuses so as not to disadvantage their students. That itself presents a major challenge for many education staff who up to this point have been little if no more familiar – indeed, possibly less familiar – with these AIs than their students: that is a major challenge that it is not going to be possible to avoid.

Lastly, given the potential for AI-misuse and academic misconduct on a par with contract cheating, education institutions are going to have to adequately resource their staff who advise students and enforce academic integrity and misconduct policies. However, whilst a major issue to address, it might not turn out to be as daunting a task as initially anticipated by some (Akbari, 2023; Eaton, 2023; Namik et al., 2023; Susnjak, 2022).

Stylometry

Up to this point, the focus has been the unreliability of AI-text detectors and the implications of that. However, there are alternative approaches to assessing authorship of student submissions. One such approach, although it is time and resource intensive, is stylometry (Crockett, 2023a; Juola, 2017). Stylometry means, reasonably literally, "measurement of style". A more informative definition is statistical analysis of variation of (writing) style with authorship (Eder et al., 2016).

In outline in the academic integrity context, using stylometry to investigate authorship involves the comparison of a suspect submission to other submissions by an individual student in question, i.e., investigation of a portfolio of submissions by an individual student in question for consistency of writing style. The objective in the academic integrity context is to identify inconsistencies in authorship and not necessarily identify specific authors (who are often unidentifiable assignment ghost-writers working for essay-mills), whereas in literary contexts, the objective of stylometry is to identify specific authors. Thus, the objective is to identify two or more subsets of submissions according to writing style that when considered on balance of probabilities, do not align with having been written by a single individual and so align with being written by multiple authors of which the student is, at most, one (Crockett, 2023a).

That objective is superficially similar to using an AI-text detector/classifier as part of an academic integrity investigation: if an AI-text detector/classifier indicates "AI generated" then an investigator has to determine, on balance of probabilities, whether or not that submission was written by the student in question and not simply rely on what is an unreliable classification.

Thus, while the problems associated with AI-text detectors/classifiers as outlined in the preceding sections also apply to stylometry, they do not do so to the same extent. This is because stylometry focuses on the uniqueness of the individual whereas AI-text detectors/classifiers sacrifice individuality in effectively comparing text to a typical aggregate human (who doesn't actually exist). Furthermore, an AI-text detector/classifier might return an "AI-generated" result for a student in question whose natural writing style is AI-like but who has not used an AI (Liang et al., 2023), or even an apparently random "AI-generated" result due to an unfortunate coincidence between something a mainstream student has written and an undocumented feature of the AI-detector's model fitting and training (Al-Sibai, 2023). This would obviously have negative consequences if that determination were to be accepted unquestioningly, or even weighted too highly, by a tutor or institution placing too much faith in what is deeply unsound software (Al-Sibai, 2023).

An AI-text detector takes one document at a time with no individual context, no consideration of an individual's writing style. Conversely stylometry, because it can evaluate an entire portfolio of student submissions, has the demonstrated potential to confirm consistency of an individual student's natural writing style, thereby establishing an impacted student's natural "AI-ness", including where misuse of AI is suspected (Sokol, 2023). Thus, stylometry has at least the potential to be a more reliable and accurate means of assessing whether a student is the writer of (all) the assignments they submit and whether or not they have misused an AI, or contract cheated, but it is time-consuming, requires expertise and needs to be resourced.

It is true that as AIs improve, and AI-generated text becomes more human-like, stylometry will become more difficult, particularly with regard to students who write in more AI-like ways. However, until AIs are programmed with the capability to imitate arbitrary individual-user writing styles "on-the-fly", there will still be differences between AI-generated text and text written by an individual student, with the potential for those differences to be sufficiently significant to provide evidence for an academic integrity disciplinary process (Crockett, 2023a).

IMPLICATIONS FOR LEARNING TECHNOLOGY

As well as posing problems for academic staff, the recent rapid developments in AI technologies pose problems for those who provide the technological support for those staff – and students – and learning technologists are at the forefront of this. Even prior to these developments, learning technology teams had been becoming increasingly central components of educational institutions (Walker & Voce, 2023). In addition to working with staff and students to assess and test the suitability of new technology products, they also have to be alert to any emerging technologies which may impact – negatively as well as positively – learning and teaching.

Whilst AI is not new, the extraordinarily rapid pace of developments and the accompanying intense media attention ("hype") since the release of ChatGPT has made adhering to established procurement processes more complicated. A procurement process typically involves identification of an institutional need, development of a formal specification for a new system, checking that any new systems fulfils the identified needs and does not unnecessarily duplicate existing systems or does not confound existing systems, and some form of bidding/tendering process. Within such processes, there will need to be data-protection and equality-impact assessments, according to institutional and legal requirements, to ensure, for example, that users' confidentiality is not compromised and that identifiable groups of users

are not disadvantaged or excluded. Institutions are currently in the position that their staff and students can access a variety of AIs, including AI-text detectors and paraphrasing and rewriting tools, simply by signing-up with an email address, effectively bypassing formal procurement processes due to the need to keep up-to-date with developments. For example, students are seeing the need to meet employers' expectations and, consequently, signing-up to AI tools to gain the experience and expertise they consider they need ahead of institutional options, advice and guidelines. Similarly, staff out of their own curiosity plus, of course, the need to keep up-to-date with their students and what their students can do with these new tools. Furthermore, those AI tools have much unknown functionality and, despite being procured informally by individuals, those individuals still expect support from learning technologists and others, for example, with regard to ethics and academic integrity. That situation is becoming further complicated by AI tools being increasingly integrated into mainstream products such as Microsoft Office and web-browsers, meaning that that it can be possible for individuals to use AIs without being aware of so doing. This means that many institutions are "playing catchup" with regards to their policies and guidance, as well as procurement, due to having to deal with this unprecedented release of uncontrolled products.

Some institutions initially declared bans on generative AIs, bans which were never credible – because institutions do not control students' out-of-institution study environments and access to online tools/resources. Such bans rapidly proved fruitless because there were too many technical and policy loopholes and, as discussed in the preceding sections, the inherent inaccuracy and unreliability of AI-text detectors/classifiers (Wood, 2023). Subsequently, more realistic attitudes have generally prevailed and institutions have increasingly recognised that such AI tools would be part of a longer-term future (Russell Group, 2023). Institutional student surveys during 2023 picked up that students were often far more aware of the implications of AI on academia than many had originally thought (Jisc, 2023b; Lea, 2023). Rather than being seen as a tool which would allow them to have assignments automatically created, many students recognised the impact of the tool on academic integrity and qualifications. As a result, some students actively avoided the use of the AI tools and discouraged others from using them (Charnovsky, 2023; Hough, 2023; Jisc, 2023a). Lea (2023) noted that the number of students adopting AI tools in their studies was in a surprising minority (during the middle of 2023), and also noted that many of the students who had not adopted tools were very ethically aware – "citing concerns of cheating as the main reason they were not using AI tools". Lea (2023) also noted that some students did not have the skills to use AI or did not feel the need to use them, and sometimes considered their use unfair. As institutions moved to the latter part of 2023, they were increasingly starting to see the benefit of the tools in certain situations and are working to improve staff and student AI skill-sets and confidence (Diplo, 2023; Hough, 2023). From the student perspective, this means that they have clear guidance on when, where and how they should and should not be using AI tools, and who they can approach for guidance – with enquiries often being passed to learning technologists (Hough, 2023; Jisc, 2023b).

In addition to the student perspective, staff are also becoming more aware of the potential of using AI in their own work (Hurix, 2023, Pine Cove, 2023). Those conducting research now have a wide range of tools which allow them to analyse data, conduct literature reviews and assist with the writing of articles. Those staff needing help with creating teaching resources now have access to systems that start to cross over with the role of course and assessment designers (Hurix, 2023). Furthermore, senior and professional services staff now have access to tools that will assist with the production and analysis of, for example, planning documents and institutional reports, and learning technologists need to be able to meet those demands.

The availability of these tools means that responsibility for AI goes far beyond that of a single learning technology team. Wider teams including learning designers, ethics and academic integrity staff, learning developers, academic librarians, academic support teams, as well as academic staff, senior managers and students will all need to be involved in the ethical roll out of systems. Institutions – with learning technologists at the forefront – need to ensure that they maintain full awareness of the implications of AI for both themselves and the wider sector, and need to address the potentially dangerous implications of some AI systems bypassing any form of formal procurement. In addition, relevant policies and procedures need to be regularly reviewed in the light of new technologies.

CONCLUSION

This chapter has outlined the basic operation of generative AIs and presented an overview of the underlying principles of AI-text detection in an effort to de-mystify the many claims and assertions that higher education staff and others have been exposed to – and will continue to be exposed to.

It is not going to be easy: to use a phrase, it is going to be a "bumpy ride" over the next several years as individual institutions and the sector as a whole progress through interim policies and regulations to more considered positions in the "brave new world" we find ourselves in. However, the key is to avoid panicking and avoid being panicked by the hype. We need to move quickly to embrace the new AI technologies, but not so quickly that we circumvent established good practices (Webb, 2023).

To put all this in context, institutions are currently dealing with a previously unencountered situation which will have implications far in excess of those which arose during the covid-19 pandemic. Whilst covid-19 alerted many institutions to gaps and shortcomings in their current teaching and assessment policies and procedures, the recent rapid – and almost certainly ongoing – developments in AI technologies will take education in new, previously unforeseeable directions. It will take effort, perseverance and drive from those inside learning technology teams and beyond to ensure that institutions maintain their position on the "crest of the wave" rather than get left behind.

Lastly, and by way of a little speculation, how long will it take for (some) students to realise that AIs – or at least the current generation thereof – are not "silver bullets"? (Marshall, 2023) As well as its programming, an AI is dependent on its dataset: if that is incomplete, out-of-date, or contains biased or widely-propagated misinformation/disinformation etc., that is reflected in the text it generates. Consequently, a student requires an appropriate degree of subject knowledge to use a generative AI to fulfil an assignment brief to an acceptable standard, noting they will generally have to revise/refine their side of the "chat" a few times in order to obtain acceptable text and not simply enter an assignment title and accept the first version the AI generates. That takes time and effort and then, ideally, the "final" text should be checked and verified (as per ethical use, insight and integrity). Thus, it is possible that after an initial flurry of activity, (some) students will largely revert to creating assignments without AI assistance or, perhaps, using AI solely for "routine" tasks such as initial information-scoping and essay-planning? Time – or, perhaps, the next generation of AIs – will tell...

ETHICS DECLARATION

The authors have not received any payment/reward/support/incentive of any kind from any artificial intelligence or educational technology companies or organisations with regard to this research.

Except for the single referenced section of text obtained from Claude-2, no generative AI was used in the writing of this chapter.

ACKNOWLEDGEMENT

The authors wish to thank their colleagues both within and outwith the University of Northampton for their support and encouragement in the preparation of this chapter.

REFERENCES

AdvanceHE. (n.d.) *Authentic Assessment in the era of AI*. AdvanceHE. https://www.advance-he.ac.uk/membership/all-member-benefit-projects/Authentic-Assessment-in-the-era-of-AI

AI Writing Check. (2023, August) *August 2023 Update – AI Writing Check is no longer available* [Press release]. AI Writing Check. https://aiwritingcheck.org/

Akbari, N. (2023, July 14) *Academic Integrity in the age of AI: Approaching Apocalypse or Achievable Equilibrium?* Linkedin. https://www.linkedin.com/pulse/academic-integrity-age-ai-approaching-apocalypse-achievable

Al-Sibai, N. (2023, June 06). *AI Plagiarism Detection Software Keeps Falsely Accusing Students of Cheating*. Futurism. https://futurism.com/ai-plagiarism-software-false-accusing-students

Anderson, C. (2023, June 01) *The False Promise of AI Writing Detectors*. Linkedin. https://www.linkedin.com/pulse/false-promise-ai-writing-detectors-carol-anderson

Anthropic. (2023, July 14) *Claude 2* [Press release]. Anthropic PBC. https://www.anthropic.com/index/claude-2

Atleson, M. (2023, July 06). *Watching the detectives: Suspicious marketing claims for tools that spot AI-generated content.* USA Federal Trade Commission. https://www.ftc.gov/business-guidance/blog/2023/07/watching-detectives-suspicious-marketing-claims-tools-spot-ai-generated-content

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (pp. 610–623). Association for Computing Machinery. 10.1145/3442188.3445922

Bernstein, M. N. (2021, October 08) *Perplexity: a more intuitive measure of uncertainty than entropy.* Github. https://mbernste.github.io/posts/perplexity/

Bowditch, E. (2023, September 12). *Assessment Menu: Designing assessment in an AI enabled world.* Jisc. https://nationalcentreforai.jiscinvolve.org/wp/2023/09/12/designing-assessment-in-an-ai-enabled-world/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., & Amodei, D. (2020) Language Models are Few-Shot Learners. arXiv. https://doi.org//arXiv.2005.14165 doi:10.48550

Brownlee, J. (2019, October 14) *A Gentle Introduction to Information Entropy*. Machine Learning Mastery. https://machinelearningmastery.com/what-is-information-entropy/

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, *2*(3), 1–12. doi:10.1016/j.rmal.2023.100068

Chakraborty, S., Bedi, A. S., Zhu, S., An, B., Manocha, D., & Huang, F. (2023). *On the Possibilities of AI-Generated Text Detection*. arXiv. https://doi.org//arXiv.2304.04736 doi:10.48550

Chan, C. K. Y. (2023) *Is AI Changing the Rules of Academic Misconduct? An In-depth Look at Students' Perceptions of 'AI-giarism'*. arXiv. https://doi.org//arXiv.2306.03358 doi:10.48550

Charnosky, C. (2023, October 02). *Pre-Law Students Avoid Using AI for Admissions*. ALM Global LLC. https://www.law.com/2023/10/02/pre-law-students-nix-using-ai-for-admissions-kaplan-survey-reports/

Copyleaks. (2023, January 12) *Copyleaks Officially Launches First-Of-Its-Kind Multi-Language AI Content Detection Solution with 99 Percent Accuracy* [Press release]. Copyleaks Ltd. https://copyleaks.com/about-us/media/copyleaks-launches-ai-content-detector

Crockett, R. (2023a). Forensic Assignment Stylometry. In S. E. Eaton (Ed.), *Handbook of Academic Integrity* (pp. 1–19). Springer. doi:10.1007/978-981-287-079-7_154-1

Crockett, R. (2023b, November 20) Testing the AI detectors Part 2. University of Northampton. NorthHampton. https://blogs.northampton.ac.uk/learntech/2023/11/20/testing-the-ai-detectors-part-2/

Crockett, R., & Best, K. (2020). Stylometric Comparison Of Professionally Ghost-Written And Student-Written Assignments. In Z. R. Khan, C. Hill, & T. Foltynek (Eds.), *Integrity in Education for Future Happiness* (pp. 35–49). Mendel University Press. doi:10.11118/978-80-7509-772-9-0035

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 1–22. doi:10.1186/s41239-023-00392-8

Crossplag. (2023, July 31) *AI Content Detector* [Press release]. Inspera AS. https://crossplag.com/ai-content-detector/

DeepLearningAI. (2023, January 11). *A Complete Guide to Natural Language Processing*. DeepLearning.AI LLC. https://www.deeplearning.ai/resources/natural-language-processing/

Desaire, H., Chua, A. E., Kim, M.-G., & Hua, D. (2023). Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Reports. Physical Science*, *4*(11), 1–13. doi:10.1016/j.xcrp.2023.101672 PMID:38078148

Diplo. (2023, September 24) *Universities stop using AI detection tool such as Turnitin.*. Diplo Foundation. https://www.diplomacy.edu/updates/universities-stop-using-ai-detection-tool-such-as-turnitin/

Eaton, S. E. (2023, March 04) Artificial intelligence and academic integrity, post-plagiarism. *University World News*. https://www.universityworldnews.com/post.php?story=20230228133041549

Fox, N. P., & Ehmoda, O. (2012). *Statistical Stylometrics and the Marlowe-Shakespeare Authorship Debate.* Brown University. https://cs.brown.edu/research/pubs/theses/masters/2012/ehmoda.pdf

GaumannN.VealeM. (2023) AI Providers as Criminal Essay Mills? Large Language Models meet Contract Cheating Law. socarXiv. https://osf.io/preprints/socarxiv/cpbfd doi:10.31235/osf.io/cpbfd

Gluska, J. (2023, August 09). *How to Bypass ChatGPT Writing Detection Tools With Other Tools*. Gold Penguin. https://goldpenguin.org/blog/avoiding-ai-detection-for-chatgpt-writing/

Google. (2023a, February 06) *An important next step on our AI journey* [Press release]. Google LLC. https://blog.google/technology/ai/bard-google-ai-search-updates

Google. (2023b, March 21). *Try Bard and share your feedback* [Press release]. Google LLC. https://blog.google/technology/ai/try-bard

Goom, H. (2023, July 12). *AI-Generated vs. Human-Written Text: Technical Analysis*. Artmap Inc. https://hackernoon.com/ai-generated-vs-human-written-text-technical-analysis

Hahn, W. W. (2023, September 15). *ChatGPT and Large Language Models: Syntax and Semantics.* CFA Institute. https://blogs.cfainstitute.org/investor/2023/09/25/chatgpt-and-large-language-models-syntax-and-semantics/

Havlik, V. (2023) Meaning and understanding in large language models. arXiv. https://doi.org//arXiv.2310.17407 doi:10.48550

Heikkila, M. (2022, December 19). How to spot AI-generated text. *MIT Technology Review*. https://www.technologyreview.com/2022/12/19/1065596/how-to-spot-ai-generated-text/

Hough, D. (2023, October 10). *Student guidance for the responsible use of AI*. Association for Learning Technology. https://altc.alt.ac.uk/blog/2023/10/student-guidance-for-the-responsible-use-of-ai/

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., Adnan, W., Alhanai, T., AlShebli, B., Baghdadi, R., Bélanger, J., Beretta, E., Celik, K., Chaqfeh, M., Daqaq, M., El Bernoussi, Z., Fougnie, D., Garcia de Soto, B., Gandolfi, A., & Zaki, Y. (2023) Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports, 13*, 1–13. Nature Publishing. doi:10.1038/s41598-023-38964-3

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1–38. doi:10.1145/3571730

Jisc. (2023a). *Artificial intelligence (AI) in tertiary education*. 3rd edition. Joint Information Systems Committee (JISC). https://beta.jisc.ac.uk/reports/artificial-intelligence-in-tertiary-education

Jisc. (2023b). *Student perceptions of generative AI*. Joint Information Systems Committee (JISC). https://beta.jisc.ac.uk/reports/student-perceptions-of-generative-ai

Jisc. (2023c). Generative AI – a primer. Version 1.1. Joint Information Systems Committee (JISC). https://repository.jisc.ac.uk/9182/1/generative-ai-a-primer.pdf

Jisc. (2023d, July 31). How UCL is redesigning assessment for the AI age. Joint Information Systems Committee (JISC). Retrieved, January 05, 2024, from https://www.jisc.ac.uk/member-stories/how-ucl-is-redesigning-assessment-for-the-ai-age

Juola, P. (2013, August 20). How a Computer Program Helped Show J. K. Rowling wrote A Cuckoo's Calling. *Scientific American*. https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/

Juola, P. (2017). Detecting Contract Cheating via Stylometric Methods. *Proceedings Plagiarism across Europe and Beyond*, 187–198. Mendel University Press. https://academicintegrity.eu/conference/proceedings/2017/Juola_Detecting.pdf

Kermes, H., & Teich, E. (2017). Average surprisal of parts-of-speech. *Proceedings Corpus Linguistics 2017*, 1–6. University of Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2017/general/paper207.pdf

Khan, U. (2023, May 16). How To Bypass AI Content Detectors: Remove AI Detection. Linkedin. https://www.linkedin.com/pulse/how-bypass-ai-content-detectors-uzair-khan

Lancaster, T. (2022, December 04). Artificial Intelligence, Generated Text and Academic Integrity: Navigating the Ethics of AI in Academia. Thomas Lancaster's Blog. Retrieved, January 05, 2024, from https://thomaslancaster.co.uk/blog/artificial-intelligence-generated-text-and-academic-integrity-navigating-the-ethics-of-ai-in-academia/

Lancaster University. (n.d.). Ling 131 – Language & Style. Department of Linguistics and English Language (LAEL), Lancaster University. Retrieved, January 05, 2024, from https://www.lancaster.ac.uk/fass/projects/stylistics/index.htm

Lea, K. (2023, November 24). Students are still confused about AI. WonkHE. Retrieved, January 05, 2024, from https://wonkhe.com/blogs-sus/students-are-still-confused-about-ai/

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns (New York, N.Y.)*, *4*(7), 1–4. doi:10.1016/j.patter.2023.100779 PMID:37521038

Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarok or reformation? A paradoxical perspective from management educators. *International Journal of Management Education*, *21*(2), 1–13. doi:10.1016/j.ijme.2023.100790

Lu, N., Liu, S., He, R., Wang, Q., & Tang, K. (2023). Large Language Models can be Guided to Evade AI-Generated Text Detection. arXiv. https://doi.org//arXiv.2305.10847 doi:10.48550

Marshall, A. J. (2023, July 17). AI: There Is No Such Thing As A Silver Bullet. Linkedin. https://www.linkedin.com/pulse/ai-thing-silver-bullet-alexander-james-marshall

McClenaghan, E. (2022, July 06). *Mann-Whitney U Test: Assumptions and Example*. Labx Media Group Inc. https://www.technologynetworks.com/informatics/articles/mann-whitney-u-test-assumptions-and-example-363425

Namik, H., Sims, A., & Withy, A. (2023, May 22). Can academic integrity prevail when AI is so good? *Ingenio*. University of Auckland. https://www.auckland.ac.nz/en/news/2023/05/22/ingenio-taking-issue-opinion-chatgpt.html

New Scientist. (2023, July 29). How To Think About AI... And How To Live With It. *New Scientist, 259*, 32–40. https://www.sciencedirect.com/science/article/pii/S0262407923014276. doi:10.1016/S0262-4079(23)01427-6

Nvidia. (2023, March 20). *What is Generative AI?* Nvidia Corporation. https://www.nvidia.com/en-us/glossary/data-science/generative-ai/

OpenAI. (2022, November 30) *Introducing ChatGPT* [Press release]. OpenAI LLC. https://openai.com/blog/chatgpt

OpenAI. (2023a, January 31). *New AI classifier for indicating AI-written text.* [Press release]. OpenAI LLC. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

OpenAI. (2023b, March 14). *GPT-4*. [Press release]. OpenAI LLC. https://openai.com/research/gpt-4

OpenAI. (2023c, July 20). *As of July 20, 2023, the AI classifier is no longer available…* [Press release] OpenAI LLC. https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

OpenAI. (2023d, September 22). *How can educators respond to students presenting AI-generated content as their own?* OpenAI LLC. https://help.openai.com/en/articles/8313351-how-can-educators-respond-to-students-presenting-ai-generated-content-as-their-own

Orenstrakh, M. S., Karnalim, O., Suarez, C. A., & Liut, M. (2023). Detecting LLM-Generated Text in Computing Education: A Comparative Study for ChatGPT Cases. arXiv. https://doi.org//arXiv.2307.07411 doi:10.48550

Phrasly. (2023, September 21). *AI Powered Writing for...* Phrasly LLC. https://phrasly.ai/?gclid=EAIaIQobChMIk_-J6puxgQMVi_ftCh0rugu0EAMYASAAEgIgQ_D_BwE

Pine Cove Consulting. (2023, April 07). *How Can an Educator Can Prevent Students from Using AI Writers? Pine Cove Consulting*. Pine Cove Consulting. https://marketing.pinecc.com/blog/how-can-an-educator-can-prevent-students-from-using-ai-writers

Reed, T. (2023, March 20). *Decoding Humanity: How to Differentiate Between AI and Human Writing*. Linkedin. https://www.linkedin.com/pulse/decoding-humanity-how-differentiate-between-ai-human-writing-reed

Retraction Watch. (2023, July 07). *Publisher blacklists authors after preprint cites made-up studies.* Center for Scientific Integrity. https://retractionwatch.com/2023/07/07/publisher-blacklists-authors-after-preprint-cites-made-up-studies/

Rogers, R. (2023, February 08). *How to Detect AI-Generated Text, According to Researchers*. Advance Magazine Publishers Inc. https://www.wired.com/story/how-to-spot-generative-ai-text-chatgpt/

Rogerson, A. M., & McCarthy, G. (2017). Using Internet based paraphrasing tools: Original work, patch-writing or facilitated plagiarism? *International Journal for Educational Integrity*, *13*(1), 1833–2595. doi:10.1007/s40979-016-0013-y

Russell Group. (2023, July 04). *New principles on use of AI in education*. The Russell Group. https://russellgroup.ac.uk/news/new-principles-on-use-of-ai-in-education/ -:~:text=

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide*. UNESCO. https://www.iesalc.unesco.org/wp-content/uploads/2023/04/ChatGPT-and-Artificial-Intelligence-in-higher-education-Quick-Start-guide_EN_FINAL.pdf

Sadasivan, V. S., Kumar, A., Balasubramanian, S., Wang, W., & Feizi, S. (2023) Can AI-Generated Text be Reliably Detected? arXiv. https://doi.org//arXiv.2303.11156 doi:10.48550

Salvagno, M., Taccone, F. S., & Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, *27*(1), 75–79. doi:10.1186/s13054-023-04380-2 PMID:36841840

Shanahan, M. (2023) Talking About Large Language Models. arXiv. https://doi.org//arXiv.2212.03551 doi:10.48550

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, *27*(3), 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shannon, C. E. (1951). Prediction and entropy of printed English. *The Bell System Technical Journal*, *30*(1), 50–64. doi:10.1002/j.1538-7305.1951.tb01366.x

Snow, E. L., Allen, L. K., Jacovina, M. E., Crossley, S. A., Perret, C. A., & McNamara, D. S. (2016). Keys to Detecting Writing Flexibility Over Time: Entropy and Natural Language Processing. *Journal of Learning Analytics*, *2*(3), 40–54. doi:10.18608/jla.2015.23.4

Sokol, D. (2023, July 10) It is too easy to falsely accuse a student of using AI: a cautionary tale. *Times Higher Education (THE)*. https://www.timeshighereducation.com/blog/it-too-easy-falsely-accuse-student-using-ai-cautionary-tale

Susnjak, T. (2022). ChatGPT: The End of Online Exam Integrity? arXiv. https://doi.org//arXiv.2212.09292 doi:10.48550

Tangermann, V. (2023, January 09). There's a Problem With That App That Detects GPT-Written Text: It's Not Very Accurate. Futurism. https://futurism.com/gptzero-accuracy

Tayeb, N. (2023, January 23). *Today we are launching our #chatgpt Detector* [Press Release]. Linkedin. https://www.linkedin.com/posts/nabil-tayeb_chatgpt-schools-universities-activity-7021140020723933184-YZcB

Tian, E. (2023a, January 03). *I spent New Years building GPTZero…* [Press release]. X/Twitter https://twitter.com/edward_the6/status/1610067688449007618?s=20&t=KgkIlG9q3Zkw_AeyXQMRVA

Tian, E. (2023b, June 01). *A Statistical Defence for AI Detection*. GPTZero LLC. https://gptzero.me/blogs/statistical-approach

Turnitin. (2023, April 04). *The launch of Turnitin's AI writing detector and the road ahead* [Press release]. Turnitin LLC. https://www.turnitin.com/blog/the-launch-of-turnitins-ai-writing-detector-and-the-road-ahead

UCL. (n.d.). *The Internet Grammar of English. The Survey of English Usage, University College London 1996-1998 & Jisc.* University College London. https://www.ucl.ac.uk/internet-grammar/home.htm

Walker, R., & Voce, J. (2023). Post-Pandemic Learning Technology Developments in UK Higher Education: What Does the UCISA Evidence Tell Us? *Sustainability (Basel)*, *15*(17), 1–16. doi:10.3390/su151712831

Wang, W., Wang, G., Marivate, V., & Hufton, A. L. (2023). On the transparency of large AI models. *Patterns (New York, N.Y.)*, *4*(7), 1–2. doi:10.1016/j.patter.2023.100797 PMID:37521049

Webb, M. (2023, September 18). *AI Detection – Latest Recommendations. Jisc National Centre for AI in Tertiary Education*. National Center for Education. https://nationalcentreforai.jiscinvolve.org/wp/2023/09/18/ai-detection-latest-recommendations/

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltynek, T., Guerrero-Dib, J., Popoola, O., Sigut, P., & Waddington, L. (2023) Testing of Detection Tools for AI-Generated Text. arXiv. https://doi.org//arXiv.2306.15666 doi:10.48550

Wilhelm, I. (2023, June 12). Nobody Wins in an Academic-Integrity Arms Race: How artificial intelligence is changing the way colleges think about cheating. *The Chronicle of Higher Education*. Retrieved, January 05, 2024, from https://www.chronicle.com/article/nobody-wins-in-an-academic-integrity-arms-race

Williams, R. (2023, July 07). AI-text detection tools are really easy to fool. *MIT Technology Review*. https://www.technologyreview.com/2023/07/07/1075982/ai-text-detection-tools-are-really-easy-to-fool/

Wood, P. (2023, February 28). Oxford and Cambridge ban ChatGPT over plagiarism fears but other universities choose to embrace AI bot. *iNews*. https://inews.co.uk/news/oxford-cambridge-ban-chatgpt-plagiarism-universities-2178391

Yu, H., & Guo, Y. (2023). Generative artificial intelligence empowers educational reform: Current status, issues, and prospects. *Frontiers in Education*, *8*, 1–10. doi:10.3389/feduc.2023.1183162