New provisions and methods for detecting AI-generated content

Ali Knott, Victoria University of Wellington

In today's talk

In today's talk

- I'll start by summarising an argument for AI-content detection mechanisms, from our group at GPAI. (See e.g. here, here)

In today's talk

- I'll start by summarising an argument for AI-content detection mechanisms, from our group at GPAI. (See e.g. here, here)
- I'll discuss what the EU's AI Act says about content detection (and more widely, transparency about AI content generators).

In today's talk

- I'll start by summarising an argument for AI-content detection mechanisms, from our group at GPAI. (See e.g. here, here)
- I'll discuss what the EU's AI Act says about content detection (and more widely, transparency about AI content generators).
- I'll summarise what Biden's Executive Order on AI says about content detection.

In today's talk

- I'll start by summarising an argument for AI-content detection mechanisms, from our group at GPAI. (See e.g. here, here)
- I'll discuss what the EU's AI Act says about content detection (and more widely, transparency about AI content generators).
- I'll summarise what Biden's Executive Order on AI says about content detection.
- I'll broaden the discussion, to include proposals about provenance-authentication schemes.

The impending deluge of AI-generated content

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework
- In ‘traditional’ media—newspapers, broadcasters.

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework
- In ‘traditional’ media—newspapers, broadcasters.

Some recent examples:

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework
- In ‘traditional’ media—newspapers, broadcasters.

Some recent examples:

- Faked audio of Joe Biden used in robocall to NH voters. (Nomorobo estimates this disseminated to 5–25K people.)

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework
- In 'traditional' media—newspapers, broadcasters.

Some recent examples:

- Faked audio of Joe Biden used in robocall to NH voters. (Nomorobo estimates this disseminated to 5–25K people.)
- Fake images of Trump surrounded by Black voters (see e.g. here)

The impending deluge of AI-generated content

The world is about to be flooded with AI-generated content.

- On social media, discussion boards, news sites, reviews sites
- In company reports, in student homework
- In 'traditional' media—newspapers, broadcasters.

Some recent examples:

- Faked audio of Joe Biden used in robocall to NH voters. (Nomorobo estimates this disseminated to 5–25K people.)
- Fake images of Trump surrounded by Black voters (see e.g. here)
- Videos of nonexistent news presenters (e.g. videos from Wolf News applauding China's policies, criticising US policies)

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

- *Not* because AI content is always worse than human content!

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

- *Not* because AI content is always worse than human content!
- *Not* because fake news only comes from AI!

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

- *Not* because AI content is always worse than human content!
- *Not* because fake news only comes from AI!

Two better reasons:

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

- *Not* because AI content is always worse than human content!
- *Not* because fake news only comes from AI!

Two better reasons:

1. AI-generated content undermines the accountability of human organisations (companies, universities): we need new institutions to preserve accountability and reputation.

The new problem for content consumers

Consumers of content have a completely new *attribution problem*:
what content comes from people, and what comes from machines?

Why do we need to know where content comes from?

- *Not* because AI content is always worse than human content!
- *Not* because fake news only comes from AI!

Two better reasons:

1. AI-generated content undermines the accountability of human organisations (companies, universities): we need new institutions to preserve accountability and reputation.
2. AI content generators threaten to destabilise information ecosystems, because individuals can generate *much more*.

Unpacking those arguments

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

- For journalism: standards & codes of conduct, libel laws

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

- For journalism: standards & codes of conduct, libel laws
- For commerce: advertising, contract law

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

- For journalism: standards & codes of conduct, libel laws
- For commerce: advertising, contract law
- In the free market, *reputation* is of great importance.

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

- For journalism: standards & codes of conduct, libel laws
- For commerce: advertising, contract law
- In the free market, *reputation* is of great importance.
- Educational institutions certify people as reliable content providers. (Institutions have reputation of their own.)

Unpacking those arguments

Premise: communication between humans through the creation of enduring content is fundamental to the ordering of our societies.

- In the dissemination of news and opinion
- In commercial communication (contracts, marketing)
- In education (content for students and from students)

Societies have developed institutions that let citizens *trust* content, and certify individuals and organisations as good providers.

- For journalism: standards & codes of conduct, libel laws
- For commerce: advertising, contract law
- In the free market, *reputation* is of great importance.
- Educational institutions certify people as reliable content providers. (Institutions have reputation of their own.)

AI-generated content can *escape* our current institutions.

The accountability argument

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

- But now, an individual can produce content they don't know much about, or maybe didn't even look at.

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

- But now, an individual can produce content they don't know much about, or maybe didn't even look at.
- E.g. students, teachers, company employees.

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

- But now, an individual can produce content they don't know much about, or maybe didn't even look at.
- E.g. students, teachers, company employees.

Of course, if AI content generation is *well used*, it can be a fantastic tool.

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

- But now, an individual can produce content they don't know much about, or maybe didn't even look at.
- E.g. students, teachers, company employees.

Of course, if AI content generation is *well used*, it can be a fantastic tool.

- But the way we *assess* a piece of content will be *different* if we know it was AI-generated.

The accountability argument

Individuals, and organisations, are *accountable* for the content they produce.

- But now, an individual can produce content they don't know much about, or maybe didn't even look at.
- E.g. students, teachers, company employees.

Of course, if AI content generation is *well used*, it can be a fantastic tool.

- But the way we *assess* a piece of content will be *different* if we know it was AI-generated.
- In particular, we will have different questions for the *human provider* of this content. (Are they 'in the loop'?)

The scale argument

The scale argument

It's easier to *destabilise* information ecosystems, now that individuals can produce more content.

The scale argument

It's easier to *destabilise* information ecosystems, now that individuals can produce more content.

- Using AI, individuals can produce *more* content, and *better* content.

The scale argument

It's easier to *destabilise* information ecosystems, now that individuals can produce more content.

- Using AI, individuals can produce *more* content, and *better* content.
- This means they can impersonate *many* individuals, who can act collectively in service of a single goal.

The scale argument

It's easier to *destabilise* information ecosystems, now that individuals can produce more content.

- Using AI, individuals can produce *more* content, and *better* content.
- This means they can impersonate *many* individuals, who can act collectively in service of a single goal.
- Organisations have the same ability to scale the content they produce. (E.g. companies, pressure groups, political parties, states.)

The scale argument

It's easier to *destabilise* information ecosystems, now that individuals can produce more content.

- Using AI, individuals can produce *more* content, and *better* content.
- This means they can impersonate *many* individuals, who can act collectively in service of a single goal.
- Organisations have the same ability to scale the content they produce. (E.g. companies, pressure groups, political parties, states.)
- There are threats to news reporting, democratic processes, the free market.

Summarising the problem

Summarising the problem

Yuval Noah Harari has an interesting (dramatic) way of putting it:

Summarising the problem

Yuval Noah Harari has an interesting (dramatic) way of putting it:

- Texts, images and sounds are ‘the stuff human culture is made of’.

Summarising the problem

Yuval Noah Harari has an interesting (dramatic) way of putting it:

- Texts, images and sounds are ‘the stuff human culture is made of’.
- AI systems that can generate such content ‘have hacked the operating system of our civilisation.’

Summarising the problem

Yuval Noah Harari has an interesting (dramatic) way of putting it:

- Texts, images and sounds are ‘the stuff human culture is made of’.
- AI systems that can generate such content ‘have hacked the operating system of our civilisation.’

I don’t know about active ‘hacking’, but AI systems certainly interfere!

What should we do?

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

- These tools would guide people in assessing content (e.g. in education, business).

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

- These tools would guide people in assessing content (e.g. in education, business).
- They would also guide organisations that *disseminate* content (e.g. social media platforms, news organisations).

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

- These tools would guide people in assessing content (e.g. in education, business).
- They would also guide organisations that *disseminate* content (e.g. social media platforms, news organisations).

But automated AI-content detectors are hard to build.

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

- These tools would guide people in assessing content (e.g. in education, business).
- They would also guide organisations that *disseminate* content (e.g. social media platforms, news organisations).

But automated AI-content detectors are hard to build.

- As AI content generators improve, it's getting ever harder.

What should we do?

To solve these problems, we need *tools* for *reliably identifying* AI-generated content.

- These tools would guide people in assessing content (e.g. in education, business).
- They would also guide organisations that *disseminate* content (e.g. social media platforms, news organisations).

But automated AI-content detectors are hard to build.

- As AI content generators improve, it's getting ever harder.
- *If you're just analysing content*, distinguishing AI-generated and 'natural' content will likely become *impossible* as generators improve.

Our proposal at the Global Partnership on AI

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.
  - They can include (invisible) ‘watermarks’ in generated content.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.
  - They can include (invisible) ‘watermarks’ in generated content.
  - They can *log* the content they generate, and implement a detector as a plagiarism detector on that log.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.
  - They can include (invisible) ‘watermarks’ in generated content.
  - They can *log* the content they generate, and implement a detector as a plagiarism detector on that log.
- The detector should distinguish *different levels of human involvement* in generated content.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.
  - They can include (invisible) ‘watermarks’ in generated content.
  - They can *log* the content they generate, and implement a detector as a plagiarism detector on that log.
- The detector should distinguish *different levels of human involvement* in generated content.
  - Generating a document ‘from scratch’ is very different from tinkering with an existing document.

Our proposal at the Global Partnership on AI

Our GPAI group argues the only way to deliver reliable detectors is by placing responsibility for detection *with the providers of AI generators*.

- That means, in practice, with the big AI companies.

Two reasons why responsibility should be with providers.

- Providers can *set up their generators* to *support detection*.
  - They can include (invisible) ‘watermarks’ in generated content.
  - They can *log* the content they generate, and implement a detector as a plagiarism detector on that log.
- The detector should distinguish *different levels of human involvement* in generated content.
  - Generating a document ‘from scratch’ is very different from tinkering with an existing document.
  - *Companies are the only ones who can distinguish*, because they’re the only ones *with access to prompts*.

Our proposal at the Global Partnership on AI

Our specific proposal was as follows:

> A company developing an AI generation system must demonstrate a reliable detection tool for the content this system generates, as a condition of its public release.

Our proposal at the Global Partnership on AI

Our specific proposal was as follows:

> A company developing an AI generation system must demonstrate a reliable detection tool for the content this system generates, as a condition of its public release.

This proposal had lots of traction with policymakers.

Our proposal at the Global Partnership on AI

Our specific proposal was as follows:

> A company developing an AI generation system must demonstrate a reliable detection tool for the content this system generates, as a condition of its public release.

This proposal had lots of traction with policymakers.

- It was incorporated into the EU's AI Act (officially approved last week).

Our proposal at the Global Partnership on AI

Our specific proposal was as follows:

> A company developing an AI generation system must demonstrate a reliable detection tool for the content this system generates, as a condition of its public release.

This proposal had lots of traction with policymakers.

- It was incorporated into the EU's AI Act (officially approved last week).
- It was also likely influential in shaping Biden's Executive Order on AI.

Our proposal at the Global Partnership on AI

Our specific proposal was as follows:

> A company developing an AI generation system must demonstrate a reliable detection tool for the content this system generates, as a condition of its public release.

This proposal had lots of traction with policymakers.

- It was incorporated into the EU's AI Act (officially approved last week).
- It was also likely influential in shaping Biden's Executive Order on AI.
  - Two of our coauthors, Yoshua Bengio and Stuart Russell, gave evidence to the Senate Judiciary Committee hearing on AI, whose findings fed into the Order.

Here's what the EU's AI Act says. . .

Here's what the EU's AI Act says. . .

Article 52 (1a)
Providers of AI systems, including GPAI systems, generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate and prosecute criminal offences.

Here's what the EU's AI Act says...

Article 52 (1a)
Providers of AI systems, including GPAI systems, generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate and prosecute criminal offences.

Here's what the EU's AI Act says...

Article 52 (1a)
Providers of AI systems, including GPAI systems, generating synthetic audio, image, video or text content, shall ensure the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated. Providers shall ensure their technical solutions are effective, interoperable, robust and reliable as far as this is technically feasible, taking into account specificities and limitations of different types of content, costs of implementation and the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards. This obligation shall not apply to the extent the AI systems perform an assistive function for standard editing or do not substantially alter the input data provided by the deployer or the semantics thereof, or where authorised by law to detect, prevent, investigate and prosecute criminal offences.

Here's what the EU's AI Act says...

Recital 70a:
A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem (...) In the light of those impacts, (...) it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods...

Here's what the EU's AI Act says...

Recital 70a:
A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem (...) In the light of those impacts, (...) it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods...

Here's what the EU's AI Act says...

Recital 70a:
A variety of AI systems can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content. The wide availability and increasing capabilities of those systems have a significant impact on the integrity and trust in the information ecosystem (...) In the light of those impacts, (...) it is appropriate to require providers of those systems to embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human. Such techniques and methods should be sufficiently reliable, interoperable, effective and robust as far as this is technically feasible, taking into account available techniques or a combination of such techniques, such as watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods...

Some useful conditions in the AI Act's obligations

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?
   - If there are ways of evading detection that can't be countered, companies aren't required to do the impossible.

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?
   - If there are ways of evading detection that can't be countered, companies aren't required to do the impossible.
   - If there are *known counters* to current evasion methods, companies can't ignore these.

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?
   - If there are ways of evading detection that can't be countered, companies aren't required to do the impossible.
   - If there are *known counters* to current evasion methods, companies can't ignore these.
   - Technical standards do useful work here.
     (I think this is common in tech legislation.)

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?
   - If there are ways of evading detection that can't be countered, companies aren't required to do the impossible.
   - If there are *known counters* to current evasion methods, companies can't ignore these.
   - Technical standards do useful work here.
     (I think this is common in tech legislation.)

2. 'Cost': support methods can't be unduly costly.

Some useful conditions in the AI Act's obligations

1. Technical feasibility: what's the current state of the art?
   - If there are ways of evading detection that can't be countered, companies aren't required to do the impossible.
   - If there are *known counters* to current evasion methods, companies can't ignore these.
   - Technical standards do useful work here.
     (I think this is common in tech legislation.)

2. 'Cost': support methods can't be unduly costly.
   - Should EU governments perhaps *subsidise* the costs of detection tools?

AI-content detection and provenance schemes

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.
  - Provenance information needs to be added when content is captured/created, transmitted, modified.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.
  - Provenance information needs to be added when content is captured/created, transmitted, modified.
- Some commentators argue we should focus on positively authenticating content from reputable providers, with unauthenticated content being less trusted.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.
  - Provenance information needs to be added when content is captured/created, transmitted, modified.
- Some commentators argue we should focus on positively authenticating content from reputable providers, with unauthenticated content being less trusted.

Legislation beyond the AI Act would be needed for provenance schemes.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.
  - Provenance information needs to be added when content is captured/created, transmitted, modified.
- Some commentators argue we should focus on positively authenticating content from reputable providers, with unauthenticated content being less trusted.

Legislation beyond the AI Act would be needed for provenance schemes.

- We think both provenance schemes and AI-content-detection schemes have their place.

AI-content detection and provenance schemes

Recital 70a also refers to provenance-authentication methods.

- These require action from providers right across the information ecosystem, not just from AI providers.
  - Provenance information needs to be added when content is captured/created, transmitted, modified.
- Some commentators argue we should focus on positively authenticating content from reputable providers, with unauthenticated content being less trusted.

Legislation beyond the AI Act would be needed for provenance schemes.

- We think both provenance schemes and AI-content-detection schemes have their place.
- Perhaps detection schemes are more useful in the shorter term.

The AI Act says a few other things on AI disclosure. . .

The AI Act says a few other things on AI disclosure...

1. Article 52(1): 'Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious (...)'

The AI Act says a few other things on AI disclosure. . .

1. Article 52(1): ‘Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious (. . .)’
   - This relates to *direct interaction* with AI systems, and transparency to ‘users’.

The AI Act says a few other things on AI disclosure...

1. Article 52(1): 'Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious (...)'
   - This relates to *direct interaction* with AI systems, and transparency to 'users'.

2. Recital 70b envisages an obligation on publishers of 'AI-generated text'.

The AI Act says a few other things on AI disclosure...

1. Article 52(1): 'Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious (...)'

- This relates to *direct interaction* with AI systems, and transparency to 'users'.

2. Recital 70b envisages an obligation on publishers of 'AI-generated text'.

- But there's an exception if the AI content has 'undergone a process of human review or editorial control and a natural or legal person holds editorial responsibility for publication'.

The AI Act says a few other things on AI disclosure...

1. Article 52(1): 'Providers shall ensure that AI systems intended to directly interact with natural persons are designed and developed in such a way that the concerned natural persons are informed that they are interacting with an AI system, unless this is obvious (...)'

- This relates to *direct interaction* with AI systems, and transparency to 'users'.

2. Recital 70b envisages an obligation on publishers of 'AI-generated text'.

- But there's an exception if the AI content has 'undergone a process of human review or editorial control and a natural or legal person holds editorial responsibility for publication'.
- Our group doesn't agree with that exception.

US Policy on AI-generated content identification

US Policy on AI-generated content identification

There was a Senate Judiciary Committee Hearing on AI Safety in July 2023, where two co-authors of our GPAI papers gave evidence.

US Policy on AI-generated content identification

There was a Senate Judiciary Committee Hearing on AI Safety in July 2023, where two co-authors of our GPAI papers gave evidence.

- There was a lot of discussion of AI-generated content detection...

US Policy on AI-generated content identification

There was a Senate Judiciary Committee Hearing on AI Safety in July 2023, where two co-authors of our GPAI papers gave evidence.

- There was a lot of discussion of AI-generated content detection...
- Senators were definitely thinking about how AI generators may disrupt this year's US election.

US Policy on AI-generated content identification

There was a Senate Judiciary Committee Hearing on AI Safety in July 2023, where two co-authors of our GPAI papers gave evidence.

- There was a lot of discussion of AI-generated content detection...
- Senators were definitely thinking about how AI generators may disrupt this year's US election.

Biden's Executive Order on AI built on these discussions.

US Policy on AI-generated content identification

There was a Senate Judiciary Committee Hearing on AI Safety in July 2023, where two co-authors of our GPAI papers gave evidence.

- There was a lot of discussion of AI-generated content detection...
- Senators were definitely thinking about how AI generators may disrupt this year's US election.

Biden's Executive Order on AI built on these discussions.

- Its aim: to strengthen public trust in the authenticity of government communications, and to tackle AI-generated disinformation.

Biden's Executive Order

Biden's Executive Order

Section 4.5(a) asks for a review of work on AI content detection.

- The Secretary of Commerce (. . .) shall submit a report (. . .) identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for (. . .) (ii) labeling synthetic content, such as using watermarking; (iii) detecting synthetic content.

Biden's Executive Order

Section 4.5(a) asks for a review of work on AI content detection.

- The Secretary of Commerce (. . .) shall submit a report (. . .) identifying the existing standards, tools, methods, and practices, as well as the potential development of further science-backed standards and techniques, for (. . .) (ii) labeling synthetic content, such as using watermarking; (iii) detecting synthetic content.

Section 4.5(b) asks for guidance on detection and provenance-authentication.

- The Secretary of Commerce, in coordination with the Director of OMB [the Office of Management and Budget], shall develop guidance regarding the existing tools and practices for digital content authentication and synthetic content detection measures.

Biden's Executive Order

Section 10.1.(b) (viii)(c) tasks the Director of OMB with making

- recommendations to [executive departments and] agencies regarding (. . .) reasonable steps to watermark or otherwise label output from generative AI.

Biden's Executive Order

Section 10.1.(b) (viii)(c) tasks the Director of OMB with making

- recommendations to [executive departments and] agencies regarding (. . .) reasonable steps to watermark or otherwise label output from generative AI.

A couple of comments:

Biden's Executive Order

Section 10.1.(b) (viii)(c) tasks the Director of OMB with making

- recommendations to [executive departments and] agencies regarding (. . .) reasonable steps to watermark or otherwise label output from generative AI.

A couple of comments:

- These orders don't impose legal obligations on companies. But they impact government procurement processes, and create expectations that may have impacts in civil lawsuits.

Biden's Executive Order

Section 10.1.(b) (viii)(c) tasks the Director of OMB with making

- recommendations to [executive departments and] agencies regarding (. . .) reasonable steps to watermark or otherwise label output from generative AI.

A couple of comments:

- These orders don't impose legal obligations on companies. But they impact government procurement processes, and create expectations that may have impacts in civil lawsuits.
- Biden's order is like the AI Act, in equivocating between AI-content-detection mechanisms and provenance mechanisms.

What are companies doing?

What are companies doing?

In February, 20 companies signed the Tech Accord to Combat Deceptive Use of AI in 2024 Elections.

What are companies doing?

In February, 20 companies signed the Tech Accord to Combat Deceptive Use of AI in 2024 Elections.

Generator providers commited to:

- Supporting the development of technological innovations to [identify] realistic AI-generated images and/or [certify] the authenticity of content and its origin, with the understanding that all such solutions have limitations. This work could include but is not limited to developing classifiers or robust provenance methods like watermarking

What are companies doing?

In February, 20 companies signed the Tech Accord to Combat Deceptive Use of AI in 2024 Elections.

Generator providers commited to:

- Supporting the development of technological innovations to [identify] realistic AI-generated images and/or [certify] the authenticity of content and its origin, with the understanding that all such solutions have limitations. This work could include but is not limited to developing classifiers or robust provenance methods like watermarking

Content disseminators commited to:

- Seek to detect the distribution of Deceptive AI election content hosted on our platforms, where such content is intended for public distribution and could be mistaken as real...

What are companies doing?

In February, 20 companies signed the Tech Accord to Combat Deceptive Use of AI in 2024 Elections.

Generator providers commited to:

- Supporting the development of technological innovations to [identify] realistic AI-generated images and/or [certify] the authenticity of content and its origin, with the understanding that all such solutions have limitations. This work could include but is not limited to developing classifiers or robust provenance methods like watermarking

Content disseminators commited to:

- Seek to detect the distribution of Deceptive AI election content hosted on our platforms, where such content is intended for public distribution and could be mistaken as real...
- ...and 'react appropriately', consistent with free expression.

A few questions to finish

A few questions to finish

- Did our group influence the wording of the AI Act?

A few questions to finish

- Did our group influence the wording of the AI Act?
- Did we have any influence on the Executive Order?

A few questions to finish

- Did our group influence the wording of the AI Act?
- Did we have any influence on the Executive Order?
- Did the Act or the Order have any influence on companies' new commitments?

A few questions to finish

- Did our group influence the wording of the AI Act?
- Did we have any influence on the Executive Order?
- Did the Act or the Order have any influence on companies' new commitments?
- Will their new commitments make any difference?

A few questions to finish

- Did our group influence the wording of the AI Act?
- Did we have any influence on the Executive Order?
- Did the Act or the Order have any influence on companies' new commitments?
- Will their new commitments make any difference?