Ética y computación

Sistemas (IA) fiables

PRESENTADO POR:

Dr. Sergio Marcellin Jacques

2020

Maestría en Ciencia e Ingeniería de la Computación

UNAM

INDICE

INDICE

Ética

Es un subcampo de la filosofía que en general, se ocupa de cuestiones como:
¿qué es una buena acción?
¿qué valor tiene la vida humana?
¿qué es la justicia? o
¿qué es una buena vida?

Ética aplicada

En el ´ambito académico, uno de los campos de investigación es el de la ética aplicada, a la que le preocupa que estamos obligados a hacer (o lo que se nos permite hacer) en una situación especifica o en un ‘ambito determinado de posibilidades de acción.

Ética aplicada

La ética de los sistemas de computo, en general, se centra en los problemas normativos que plantea el desarrollo, despliegue, utilización y supervisión de estos sistemas que usan (entre muchas otras) tecnologías la IA.

Los principios éticos en el contexto de sistemas son:

- Respeto a la autonomía humana
- Prevención de daños
- Equidad
- Explicabilidad

Sistema (IA) fiable

Tiene tres componentes: El sistema debe ser:

1. Lícito, es decir, cumplir con todas las leyes y reglamentos aplicables

2. Ético, demostrando el respeto y garantizando el cumplimiento de los principios y valores éticos, y

3. Robusto tanto desde el punto de vista técnico como social, puesto que los sistemas , incluso si las intenciones son buenas, pueden provocar daños accidentales.

La fiabilidad de un sistema no solo concierne al propio sistema, sino también a la de todos los procesos y agentes implicados en el ciclo de vida del sistema.

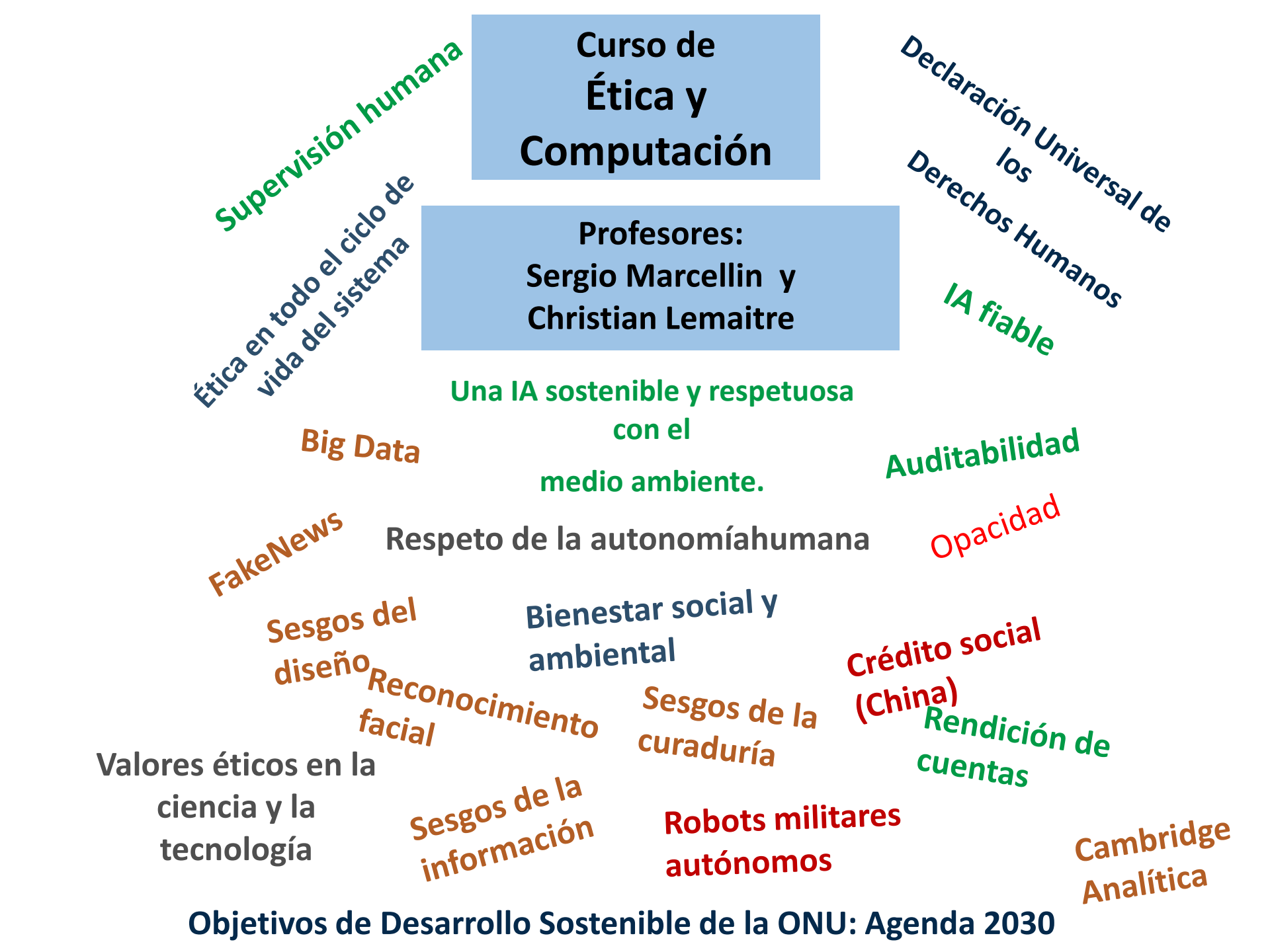
INDICE

Introducción

Ejemplos

Directrices Éticas para una IA Fiable. (Comisión Europea)

Proyecto piloto para la implementación de la lista de evaluación para una IA Fiable


Curso de
Ética y
Computación
Profesores:
Sergio Marcellin y
Christian Lemaitre
Supervisión humana
Ética en todo el ciclo de vida del sistema
Declaración Universal de los Derechos Humanos
IA fiable
Una IA sostenible y respetuosa con el medio ambiente.
Big Data
Auditabilidad
FakeNews
Respeto de la autonomíahumana
Opacidad
Sesgos del diseño
Bienestar social y ambiental
Crédito social (China)
Reconocimiento facial
Sesgos de la curaduría
Rendición de cuentas
Valores éticos en la ciencia y la tecnología
Sesgos de la información
Robots militares autónomos
Cambridge Analítica
Objetivos de Desarrollo Sostenible de la ONU: Agenda 2030

¡2 ejemplos polémicos actuales!

SISTEMA DE CRÉDITO SOCIAL CHINO (社会信用体系)

Asistentes virtuales

Crédito social chino

https://www.youtube.com/watch?v=uReVvICTrCM

2 Tiempo: 5.19 BBC

ttps://www.youtube.com/watch?v=pNf4-d6fDoY

1 Tiempo: 2.49 W Post

Asistentes digitales

https://www.youtube.com/watch?v=DxxAwDHgQhE

Autos Autónomos

Transporte (¿aéreo?)

INDICE

Introducción

Ejemplos

Directrices Éticas para una IA Fiable.

(Comisión Europea)

Proyecto piloto para la implementación de la lista de evaluación para una IA Fiable

Requisitos que incluyen aspectos sistémicos, individuales y sociales, para hacer un sistema fiable:

1. Acción y supervisión humanas
2. Solidez técnica y seguridad. Resistencia a los ataques y seguridad
3. Gestión de la privacidad y de los datos
4. Transparencia
5. Diversidad, no discriminación y equidad
6. Bienestar social y ambiental.
7. Rendición de cuentas

A continuación, de los 7 requisitos anteriores, veremos ejemplos de las 128 preguntas que forman la lista de evaluación para tener sistemas fiables.

Por último, veremos:
Descripción de un proyecto piloto realizado y sus resultados

1. Acción y supervisión humanas

- **¿El sistema mejora o aumenta las capacidades humanas?**

- **En el caso de que el sistema se implante en el proceso de trabajo, ¿ha tenido usted en cuenta la asignación de tareas entre el sistema y los trabajadores humanos para garantizar interacciones adecuadas y una supervisión y control humanas apropiadas?**

- **¿Se han adoptado medidas para evitar que los procesos de trabajo confíen o dependan en exceso del sistema?**

- **¿Qué tipo de mecanismos de detección y respuesta ha establecido para evaluar si algo puede salir mal?**

2. Resistencia a los ataques y seguridad

✓ ¿Ha evaluado las posibles formas de ataque a las que puede ser vulnerable el sistema ?

✓ En particular, ¿ha analizado los diferentes tipos y naturalezas de las vulnerabilidades, como la contaminación de los datos, la infraestructura física o los ciberataques?

✓ ¿Ha evaluado el comportamiento de su sistema en situaciones o entornos imprevistos?

✓ ¿Ha establecido algún mecanismo o comunicación para garantizar a los usuarios (finales) que el sistema es seguro?

3. Gestión de la privacidad y de los datos

✓ **¿Ha introducido mecanismos de aviso y control sobre los datos personales en función del caso de uso (como, por ejemplo, el consentimiento válido y la posibilidad de revocar el uso de dichos datos, cuando proceda)?**

✓ **¿Ha tomado medidas para mejorar la privacidad, por ejemplo a través de procesos como el encriptado, el anonimato,….?**

✓ **En los casos en que exista una persona responsable de la privacidad de los datos, ¿la ha implicado desde una fase inicial del proceso?**

✓ **¿Ha alineado su sistema con las normas potencialmente pertinentes (por ejemplo, ISO, IEEE) o ha adoptado protocolos generales para la gestión y gobernanza cotidianas de sus datos?**

4. Transparencia

✓ ¿Ha adoptado medidas que puedan garantizar la trazabilidad? Esto puede conllevar la documentación de:

➢ los métodos utilizados para diseñar y desarrollar el sistema algorítmico:

➢ los métodos empleados para ensayar y validar el sistema algorítmico:

➢ los resultados del sistema algorítmico: se deberían documentar los resultados del algoritmo o las decisiones adoptadas por este.

✓ ¿Se ha asegurado de que se pueda elaborar una explicación comprensible para todos los usuarios que puedan desearla sobre las razones por las que un sistema adoptó una decisión determinada que diera lugar a un resultado específico?

✓ ¿Ha informado a los usuarios (finales) —mediante cláusulas de exención de responsabilidad u otros medios— de que están interactuando con un sistema y no con otro ser humano?

✓ ¿Ha comunicado con claridad las características, limitaciones y posibles carencias del sistema:

- **en caso de desarrollo: a las personas encargadas de su despliegue en un producto o servicio?**
- **en caso de despliegue: a los usuarios finales o consumidores?**

5. Diversidad, no discriminación y equidad

✓ ¿Se ha asegurado de que exista una estrategia o un conjunto de procedimientos para evitar crear o reforzar un sesgo injusto en el sistema, tanto en relación con el uso de los datos de entrada como en lo referente al diseño del algoritmo?

✓ Dependiendo del caso de uso, ¿se ha asegurado de introducir un mecanismo que permita a otras personas informar sobre posibles problemas relacionados con la existencia de sesgos, discriminación o un rendimiento deficiente del sistema?

✓ ¿Ha evaluado si existe la posibilidad de que las decisiones varíen aunque las condiciones no cambien?

✓ ¿Ha allanado el camino para la introducción del sistema en su organización, informando e implicando previamente a los trabajadores afectados y sus representantes?

6. Bienestar social y ambiental.

✓ **¿Ha establecido mecanismos para medir el impacto ambiental del desarrollo, despliegue y utilización del sistema (por ejemplo, energía consumida por cada centro de datos, tipo de energía utilizada por los centros de datos, etc.)?**

✓ **¿Se ha asegurado de que se entiendan correctamente los efectos sociales del sistema? Por ejemplo, ¿ha evaluado si existe un riesgo de pérdida de puestos de trabajo o de descalificación de la mano de obra? ¿Qué pasos se han dado para contrarrestar esos riesgos?**

✓ **¿Ha evaluado el impacto social global asociado al uso del sistema más allá del que tenga sobre el usuario (final), como, por ejemplo, las partes interesadas que pueden verse indirectamente afectadas por dicho sistema?**

7. Rendición de cuentas

✓ **¿Ha establecido marcos de formación y educación para el desarrollo de prácticas de rendición de cuentas?**

✓ **¿Ha considerado la posibilidad de crear una «junta de revisión ética» u otro mecanismo similar para debatir sobre las prácticas éticas y de rendición de cuentas en general, incluidas las posibles «zonas grises»?**

✓ **¿Existe algún proceso para que los trabajadores o agentes externos (por ejemplo, proveedores, consumidores, distribuidores/vendedores) informen sobre posible vulnerabilidades, riesgos o sesgos en el sistema o su aplicación?**

✓ **¿Se han instaurado mecanismos para proporcionar información a usuarios (finales) y a terceros sobre las oportunidades de obtener compensación?**

INDICE

Resultados del piloto

Aplicando los 7 requisitos que incluyen los aspectos sistémicos, individuales y sociales, para hacer un sistema fiable y que consta de 128 preguntas, con dos equipos de entrevistadores, se obtuvieron, los siguientes resultados:

Resultados del piloto

- Mayor número de respuestas a los puntos:
  fue el **NO** (representando el 42 y 43 %)

- Mayor consideración a los puntos:

1. Acción y supervisión humanas

3. Gestión y privacidad de los datos

- Menor consideración a los puntos:

2. Resistencia a los ataques y seguridad

4. Transparencia

5. Diversidad, no discriminación y equidad

6. Bienestar social y ambiental

7. Rendición de cuentas

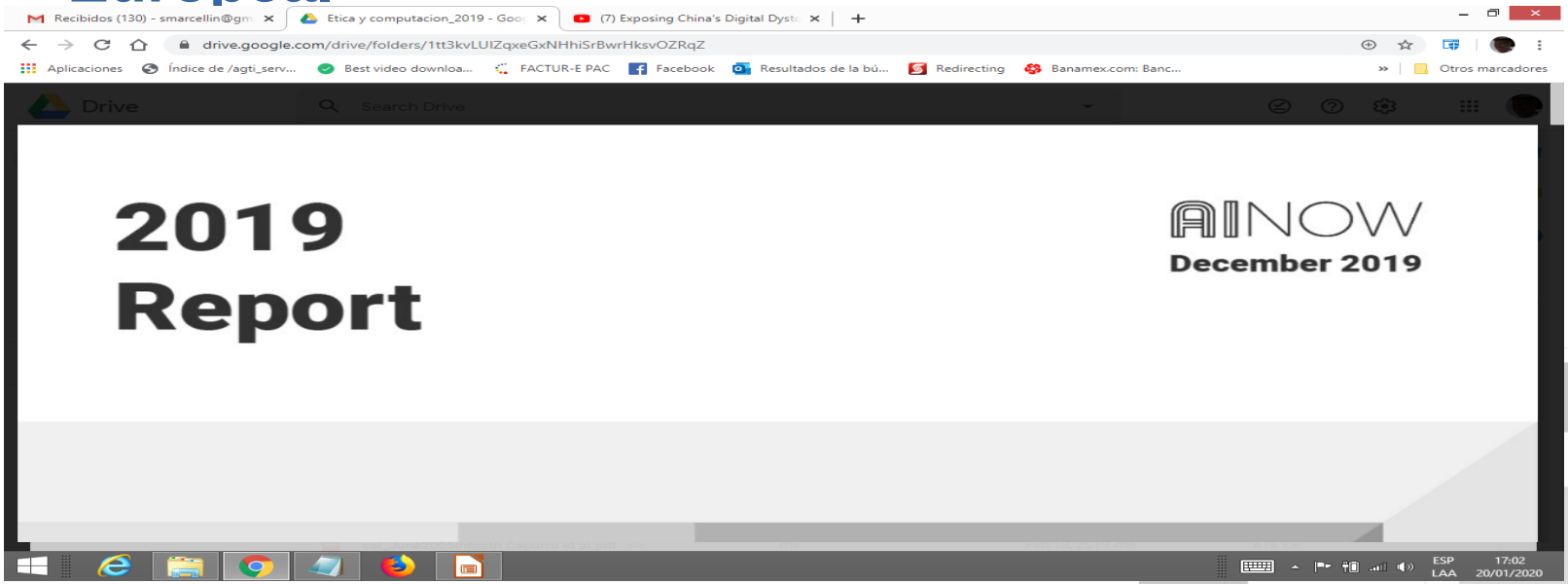
Resumen

Hemos presentado una herramienta (una lista) no exhaustiva de preguntas para apoyar el desarrollo de sistemas fiables a lo largo de todo su ciclo de vida, es decir desde su diseño, hasta su desarrollo, despliegue y utilización.

Tanto entrevistados como entrevistadores opinaron que fue muy útil el ejercicio (un poco largo y tedioso) para poder lograr una mayor fiabilidad en el sistema y admitieron que muchas de las preguntas, ni siquiera se les había ocurrido plantearselas.

Referencias:

Directrices Éticas para una IA Fiable. Grupo Independiente de Expertos de Alto Nivel Sobre Inteligencia Artificial (2018). Comisión Europea.



Con cerca de 500 referencias

Vivir todos simplemente para que todos puedan simplemente vivir

Vivre tous simplement pour que tous puissent simplement vivre.

Living all simply so that all can simply live.

Gandhi

¡GRACIAS POR SU ATENCIÓN!

Dr. Sergio Marcellin Jacques

smarcellin@gmail.com

REFERENCIAS

20- Enero 2020

Google boss Sundar Pichai calls for AI regulation

https://www.bbc.co.uk/news/technology-51178198

Referencias

https://www.youtube.com/watch?v=0cGB8dCDf3c

https://www.youtube.com/watch?v=DxxAwDHgQhE

Crédito social chino

1 Tiempo: 2.48 WPost

https://www.youtube.com/watch?v=uReVvICTrCM

2 Tiempo: 8.06 Puntos NBC

https://www.youtube.com/watch?v=0cGB8dCDf3c

Asistente

https://www.youtube.com/watch?v=DxxAwDHgQhE

3 Tiempo: 5.19 BBC

https://www.youtube.com/watch?v=pNf4-d6fDoY

Tiempo: 2.35 CBS

https://www.youtube.com/watch?v=Onm6Sb3Pb2Y