VDW VEREINIGUNG DEUTSCHER WISSENSCHAFTLER

Informe sobre los principios Asilomar en Inteligencia Artificial

Grupo de Estudio

Evaluación de la Tecnología de la Digitalización

Autores (miembros del grupo de estudio VDW)

Prof. Dr. Ulrich Bartosch

Prof. Dr. Stefan Bauberger SJ

Tile von Damm

Dr. Rainer Engels

Prof. Dr. Malte Rehbein

Frank Schmiedchen (líder)

Prof. Dr. Heinz Stapf-Finé

Angelika Sülzen

Contactos:

Editorial: Grupo de Estudio Evaluación de la tegnología de la digitalización de la Federación de Científicos Alemanes (VDW)
Marienstraße 19/20, 10117 Berlin

Impreso en Alemania Licencia: CC BY-NC-ND

Diciembre 2018

Indice de contenido

1. Introducción

La Federación de Científicos Alemanes e.V. (VDW), a través de su Grupo de Estudio "Evaluación de la Tecnología de la Digitalización" presenta su primera obra sobre los aspectos éticos, políticos y legales de la investigación, el desarrollo y la aplicación de la inteligencia artificial, concediendo especial importancia a la inteligencia artificial general (o fuerte). Como punto de partida, se refiere a los principios de Asilomar sobre inteligencia artificial adoptados el 6 de enero de 2017, que reflejan bien el estado actual de la discusión sobre los temas antes mencionados iniciada por la comunidad de desarrolladores de IA[1] Un grupo muy grande de actores relevantes a nivel mundial involucrados en la investigación y el desarrollo de la inteligencia artificial se identifican con estos principios. Entre ellos se encuentran muchos de los actores claves en países occidentales industrializados, algunos de los cuales ocupan puestos de liderazgo.

La investigación, el desarrollo y la aplicación de la inteligencia artificial pueden significar una amenaza fundamental para las personas y su coexistencia pacífica. Es evidente que es de interés público reconocer estos avances en una fase temprana y permitir que se tomen medidas razonables para evitar el peligro. Pertenece al principio de prevención de riesgos actuar de forma preventiva antes que los acontecimientos que implican una amenaza hayan llegado a una dinámica irreversible en la que ya no se permiten contramedidas eficaces. El progreso en el campo de IA está creando una amenaza fundamental. Esto ha sido reconocido por los principales especialistas en esta área de investigación y entre otras cosas llevó a la formulación de los principios Asilomar. Esto es altamente reconocido y apreciado.

Nuestra revisión de los principios Asilomar ha demostrado que éstos no cumplen con la lógica de una seguridad exitosa. Si la contención de la IA sólo siguiera la experiencia de los principios de Asilomar, desde nuestro punto de vista se aceptaría un riesgo a escala

[1] Future of Life Institute (2017).

existencial, que debe considerarse igual a otras amenazas a la humanidad y para la cual ya se acepta la necesidad y la lógica de la seguridad preventiva incondicional.

La estrategia militar nuclear, por ejemplo, sigue esta lógica en todo el mundo. Con pocas excepciones, todavía hoy se reconoce que es necesario prevenir el uso de armas nucleares (incluyendo las armas tácticas), porque una escalada incontrolable amenaza con llevar a la autodestrucción de la humanidad, a una guerra de extinción sin un vencedor. Desde la confianza mutua no podía y no puede ser asumido como una garantía contra acciones hostiles, es importante vincular de forma fiable la capacidad de control mutuo con el potencial simultáneo de aniquilación mutua y aplicar un concepto de "seguridad común".[2]

La política climática global también sigue esta lógica. Desde 1992, la comunidad internacional ha tenido que limitar un aumento adicional de la temperatura del planeta. La peligrosa situación ha sido claramente reconocida y se necesita una acción concertada para evitar que la humanidad se queme a sí misma.[3] Dado que los intereses en el campo de la política climática son mucho más diversos y ramificados que en el caso de la amenaza militar que representan las armas nucleares, un consenso para la acción política global compartida es mucho más difícil. Los hechos del cambio climático antropogénico, diagnosticados sin lugar a dudas por los científicos de la naturaleza, a menudo se ven desbordados por intereses económicos, políticos y también culturales en conflicto. Sin embargo, la agenda de transformación de la Convención de París sobre el Cambio Climático[4] y los Objetivos de Desarrollo Sostenible (SDG) [5] representan la lógica de la prevención incondicional de las amenazas para la humanidad.

[2] VDW-miembro Egon Bahr. Bahr / Lutz (1992); Comisión Independiente de Desarme y de Seguridad (1982).
[3] VDW-miembro Hans Joachim Schellnhuber. Schellnhuber (2015).
[4] Naciones Unidas (2015a).
[5] Naciones Unidas (2015b).

También en el ámbito de la biotecnología, las intervenciones irreversibles en el material genético natural de los seres humanos, los animales y las plantas suponen una amenaza cada vez mayor para la humanidad en su conjunto. La transformación progresiva de la naturaleza humana y la naturaleza como el medio ambiente humano es aceptada comunmente aunque esto produce efectos todavía más incontrolables. Aquí, también, las causas son complejas y confusas. Por encima de todo, la curiosidad científica, el optimismo sobre el progreso, los intereses comerciales y las reivindicaciones de poder a menudo se superponen. Los últimos avances hacen posible que los “bio-hackers” en "laboratorios de garaje" puedan producir organismos peligrosos de acuerdo con "recetas de cocina" y retiren la aplicación tecnológica de cualquier control estatal. Gene Drive rompió las "leyes de la evolución" y puso en marcha una dinámica de proporciones sin precedentes. Los principales científicos en este campo demandan: "proteger a la sociedad de nuestros inventos"[6]. "Es una amenaza real para la humanidad. Y los periódicos no lo reportan[7].

Lo que estas amenazas fundamentales tienen en común es que tienen una dinámica que conduce a una pérdida de control sobre sus desarrollos. La guerra nuclear, la catástrofe climática, el cambio evolutivo genético tienen el potencial de exterminar la vida humana en la Tierra. Al mismo tiempo, este peligro potencial se esconde detrás de la supuesta soberanía de control del creador de este desarrollo: el ser humano. Por lo tanto, es necesario contrarrestar los peligros que son poco o nada visibles (al menos para la mayoría de los observadores). En el área de la amenaza nuclear, la atención del público está más sensibilizada. Un cambio climático peligroso hasta catastrófico provocado por el hombre también es motivo de preocupación en círculos más amplios. La incomodidad con las intervenciones en los procesos evolutivos es bastante baja. Una nueva amenaza de proporciones similares a las de la época, que sólo se ha vuelto realista en los últimos tres

[6] Oye, et.al. (2014).
[7] VDW-miembro Ernst-Ulrich von Weizsäcker. Weizsäcker/Wijkman (2017).

años, también es en gran medida desconocida para el público. Este es el objeto de la presente declaración de la VDW.

Con inteligencia artificial, los humanos abren otra caja de Pandora. Tiene el potencial de socavar la lógica del control, incluidos los peligros antes mencionados. En la era digital, las herramientas para el pensamiento humano pueden desplegar una posición autónoma que se dirige poderosamente contra el ser humano impotente. Ya es hora de que se refleje una respuesta de emergencia activa. El siguiente texto tiene por objeto recordárselo.

La Asociación de Científicos Alemanes está comprometida con este recordatorio. Su fundación se remonta a la advertencia del "Göttinger Achtzehn"[8] de 1957 contra los peligros del armamento nuclear. [9] La responsabilidad de la ciencia hoy en día no es otra que la de hacer que los peligros no reconocidos sean visibles para el público y utilizar los conocimientos científicos para evitarlos. Teniendo esto en cuenta, el "Grupo de Estudio Evaluación de la Tecnología de la Digitalización" de la VDW ha comenzado a trabajar y ahora presenta sus primeros resultados.

[8] Göttinger 18 (1957).
[9] https://vdw-ev.de/ueber-uns/geschichte-und-ziele/

2. Objetivo, definiciones y tesis inicial

El objetivo del dictamen es contribuir al debate europeo e internacional sobre las posibles consecuencias de las IA y las medidas necesarias para abordarlas. Como grupo de estudio de la VDW argumentamos desde el punto de vista de la protección de los derechos humanos individuales, económicos, sociales y culturales[10] y, además, consideramos las posibles consecuencias de una IA general o fuerte para la humanidad en su conjunto. El principio de precaución de la UE es la brújula legal a considerar.[11]

El concepto de inteligencia es controvertido. Por lo tanto, utilizamos una comprensión básica generalmente aceptada que describe la inteligencia como un fenómeno no directamente observable que describe las capacidades cognitivas de una persona y los conocimientos que están a su disposición en un momento dado.[12]

La inteligencia artificial es una rama (empírica) de la informática y se ocupa de métodos que permiten a un ordenador resolver tareas que, cuando resuelto por los seres humanos, requieren de inteligencia. El término "inteligencia artificial" fue utilizado por primera vez en 1956 por el científico de la computación estadounidense, John McCarthy.[13] Ya en 1950 Alan Turing describe la posibilidad de la inteligencia simulada por ordenador.[14] Las características que definen IA son la capacidad incorporada de aprender desde el principio y de hacer frente a la incertidumbre, la inexactitud, la borrosidad y las probabilidades.[15]

[10] Naciones Unidas (1948); Naciones Unidas (1966).
[11] Comisión de las Comunidades Europeas (2000).
[12] Cognición se entiende aquí como “cognición situada”, lo que incluye no solamente los procesos de cálculo interno de adición en el cerebro, sino también y sobre todo la interacción recíproca en tiempo real de un sistema físicamente compuesto de cierto modo con su enterno (Walter, 2014).
[13] BITKOM (2017).
[14] EFI (2018), S. 69.
[15] BITKOM (2018) desccribe una taxonomía de la automatización de la toma de decisiones y desarrolla un modelo de 5 pasos a este proposito.

IA débil son aquellas formas de IA en las que las máquinas sólo simulan un comportamiento inteligente utilizando las matemáticas y la informática en un área específica de aplicación y poseen la capacidad de aprender. Por el contrario, IA general o fuerte significa una capacidad de aprendizaje en general, incluyendo la capacidad de desarrollarse de manera autónoma. La superinteligencia se define como un IA fuerte que es superior a la del cerebro humano por lo menos en muchas áreas. Surge hipotéticamente cuando un IA fuerte se mejora y se expande a través de recursión[16]. Las posibles direcciones de desarrollo para la realización de una superinteligencia son los algoritmos (también en máquinas, robots, etc.), el transhumanismo (por ejemplo, mejoras genéticas de los humanos o la fusión de ambos en cyborgs), así como el uso de redes neuronales artificiales.

La inteligencia artificial tiene una historia variada. Mientras tanto, ha alcanzado una etapa de desarrollo en la que puede cambiar drásticamente todas las áreas de la vida a través de su aplicación. El punto de partida de nuestro trabajo es la preocupación de que los probables peligros de una inteligencia artificial general o fuerte no se discutan adecuadamente a tiempo y que las medidas adecuadas para evitar el peligro se vuelvan imposibles. Esto se debe, por un lado, a que los grandes avances en el caso de las IA débiles se han producido predominantemente sólo en los últimos tres años y, por otro, a que, al igual que con otras tecnologías de riesgo (por ejemplo, la energía nuclear, el cambio climático, la ingeniería genética), los posibles peligros de las IA fuertes que aún no existen son abstractos, invisibles y desconocidos. Los datos y los algoritmos despliegan su efecto sólo cuando entran y se manifiestan en la realidad humana de la vida.

Las siguientes tesis iniciales constituyen la base del dictamen:

[16] La fuerte IA logra la adaptación y la mejora de su propio código de programa de tal manera que se puedan realizar más mejoras en la próxima etapa o su próxima generación, lo cual habría sido imposible en las versiones anteriores. Estos ciclos de aprendizaje recursivas se continúan por la fuerte IA hasta que la inteligencia humana es superada y se convierte en super-inteligencia. Un requisito previo para esto es que la fuerte IA "entienda" su propósito y el diseño de su propio código del programa.

1. La investigación, el desarrollo y la aplicación la IA se desarrollan de manera exponencial.[17] Además de EEUU y China, Alemania es líder en la en la investigación básica.[18]
2. Incluso las aplicaciones de IA débiles cambian el comportamiento social y la comunicación[19] y nuestra cultura cotidiana (por ejemplo, la puntuación social) y puede poner en peligro los sistemas sociales (por ejemplo, influir en la elecciones).
3. La creación de la IA fuerte es probable, dada la tasa de crecimiento exponencial de la investigación en IA, aunque el tiempo de realización es difícil de prever y no es exactamente predecible.[20]
4. La transición de débil a fuerte IA es un proceso continuo, que tiene lugar principalmente cuando estén disponibles los recursos necesarios (datos, finanzas, potencia). Los resultados de la investigación y desarrollo (I + D) en esta área no se publican necesariamente debido a los intereses económicos y políticos, lo que hace imposible el control socialmente legítimo o, al menos, considerablemente más difícil.
5. Una IA fuerte puede ser peligrosa para las personas y para la humanidad debido a su actividad autónoma:
    a. Una IA toma deciciones intencionadas por hombres para alcanzar metas o la autopreservación, que como efecto secundario (colateral) perjudican a las personas, lo que también puede conducir al grado de subyugación o

[17] BITKOM (2017); EFI (2018).
[18] EFI (2018), pp 68 y ss.
[19] Eberle (2015), Henk (2014). (Nota del editor: Sin embargo, la influencia de la IA es difícil a medir de forma independiente de otros aspectos de la digitalización).
[20] Bostrom / Müller (2013): Estos lo preveen para 2022 (año optimista mediano), 2040 (año realista mediano) o 2075 (año pesimista mediano). Estimaciones similares también se pueden hacer a base de la Ley de Moore de 1965 extrapolando el avance explosivo de los últimos tres años. La potencia de procesamiento del cerebro humano alcanza, de acuerdo con Raymond Kurzweil, aproximadamente 10.000 teraflops. Esta potencia de cálculo ya han superado significativamente los sistemas informáticos de gran tamaño. Además, se pueden enlacar directamente entre sí en red. Pero hay voces que afirman que el desarrollo de la IA fuerte "en el futuro previsible no se va a poder realizar" (ver EFI (2018), p 69).

aniquilación completa de la humanidad, sin que la IA lo haya "querido deliberadamente".[21]

b. Una IA tiene una intención destructiva (por ejemplo, un sistema autónomo de armas) y aumenta su eficacia de manera que pone en peligro a más personas / la humanidad.

c. Una IA desarrolla competencias que no son intencionadas y persigue objetivos establecidos por sí mismo que ponen en peligro a los individuos o a la humanidad.

6. Desde un punto indefinido en el tiempo, la gente ya no puede seguir el proceso de auto-mejora de una IA fuerte (ni siquiera como colectivo "interconectado"), ya que su proceso de aprendizaje es siempre más lento que el de la IA, por lo que el control /corrección es imposible.

7. Las tesis iniciales 3 a 6 subrayan la necesidad de aplicar el principio de precaución.[22] Con un margen de seguridad de tiempo suficiente antes de la creación de una IA fuerte, todos los pasos necesarios (sobre todo los normativos) deben ser completados con éxito para evitar que una IA fuerte ponga en peligro a los seres humanos individuales y a la humanidad en su conjunto en cualquier forma, en cualquier momento y bajo cualquier circunstancia (goernabilidad anticipada). Para

[21] Un ejemplo de Stephen Hawking complementado: Gente responsable planifica y construye una central hidroeléctrica. Han hecho una evaluación extensiva de los efectos medio-ambientales y sociales. Mamíferos y nidos de pajaros han sido evacuados y se daba una recompensación generosa para el reasentamiento antés de inundar el embalse. Pero nadie se ha preocupado de las 50.000 colonias de hormigas que ahora están ahogadas en el fondo del embalse: 2.500.000.000.000 hormigas matadas son el resultado no intencionado. Hawking opina que hay que emapezar a actuar hoy mismo para evitar que nosotros seamos un día en la situación de las hormigas.

[22] De acuerdo con la Agenda 21 (Capítulo 35, párrafo 3), aprobada en la Conferencia de Naciones Unidas sobre el Medio Ambiente y el Desarrollo en Río de Janeiro (1992), el principio de precaución se puede definir como una llamada a la acción imprescindible a tomar las medidas de protección preventivas para evitar los posibles riesgos / peligros de las cuales la ocurrencia no (exactamente) se puede predecir y que pueden conducir a (exactamente) impredecibles consecuencias negativas para las personas (especialmente de vida y salud) y el medio ambiente, ahora o en el futuro. Esto también se vincula con la ética de la responsabilidad de Hans Jonas.

lograrlo, deben adoptarse inmediatamente decisiones específicas y medidas apropiadas.

8. El hombre y la máquina comienzan a separarse en tres formas:
    a. La tecnología actúa de forma cada vez más independiente y, por lo tanto, asume funciones y tareas de las personas;
    b. los componentes neurológicos-cognitivos están incorporados en los seres humanos (las llamadas "mejoras");[23]
    c. Acoplamiento de los cerebros humanos y sistemas de IA

Se puede hablar de una disolución progresiva de las fronteras, que tiene fuertes implicaciones éticas. Las cuestiones relacionadas con el transhumanismo sólo se abordan de forma rudimentaria en el presente dictamen.

3. Reflexión crítica de los principios Asilomar

3.1 Introducción

Desde marzo de 2014, el Future of Life Institute (FLI), fundado en Boston (EE.UU.), se ocupa de los posibles peligros existenciales de un mayor desarrollo tecnológico para la humanidad. El trabajo sobre la reducción del riesgo de IA es explícitamente el foco del trabajo del instituto.[24] En enero de 2017, el FLI organizó la conferencia "Beneficial AI" en Asilomar (en la costa de California) con casi 1.000 participantes, entre ellos más de 100 de los principales investigadores y empresarios en IA en el mundo, con el fin de analizar los efectos de la IA. Los

[23] TAB (2016) y Vorwinkel (2017).

[24] Los fundadores y promotores fuertes del instituto son Stephen Hawkin, Elon Musk, Max Tegmark (MIT), Jaan Tallinn (inventor de Skype), Stuart J. Russell (ciencias de computación), George Church (biología), Saul Perlmutter y Frank Wilczek (físicas) así cómo Alan Alda y Morgan Freeman (actores). En enero de 2015, Elon Musk dio un fondo de 10 milliones de dolares para un programa de investigación sobre IA para el FLI que está enfocando problemas de seguridad y el desarrollo de una IA útil. El fondo se ha utilizado para financiar 37 proyectos de investigación.

Pprincipios de IA de Asilomar" son el resultado de esta conferencia. El comité de programa estaba formado por personas directa o indirectamente implicadas profesionalmente en la IA[25] Los 23 principios adoptados y firmados por numerosos científicos representan una propuesta de compromiso voluntario para la investigación, desarrollo y aplicación de la IA. Han de ser visto como una reacción al acelerado desarrollo tecnológico en esta área, que los organizadores de la FLI con razón llaman "cambio importante [...] en todos los segmentos de la sociedad".[26] Dentro de las primeras seis semanas, los principios han sido firmados por más de 1000 científicos directamente involucrados en la investigación y el desarrollo de IA, así como por otras 2000 personas.

Ya han mostrado los primeros efectos.

- La Universidad de Montreal comenzó un proceso abierto en el año 2017 para desarrollar los Principios IA Responsable de Montreal, que debe completarse en 2018.[27]
- Algunos 60 científicos de 30 países están llamando a un boicot del Instituto KAIST, Corea del Sur, ya que colabora con el grupo de armamento de Corea del Sur Hanwha.[28]

El lema de la Conferencia “IA Beneficiosa" ya ha dado la dirección de la conferencia, lo que también se refleja en el breve preámbulo de los principios: "La inteligencia artificial ya ha proporcionado herramientas beneficiosas que se utilizan todos los días por personas de todo el mundo. Su continuo desarrollo, guiado por los siguientes principios, ofrecerá increíbles oportunidades para ayudar y empoderar a las personas en las décadas y siglos venideros".[29] Con este punto de vista puramente positivista y utilitaria de la utilización de las tecnologías

[25] Future of Life Institute (2017).
[26] Ibidem.
[27] Universidad de Montreal (2018).
[28] En la creación de un “Centro de Investigación para la Convergencia de la Defensa Nacional e Inteligencia Artificial”. Centro sobre el Impacto de la IA y Robótica (2018).
[29] Future of Life Institute (2017).

de inteligencia artificial, la gran pregunta - y quizá decisiva - permanece abierta (entre otros): ¿Cómo abordar los acontecimientos no beneficiosos (o no beneficiosos para todos) y, sobre todo, cómo abordar las amenazas derivadas de estos acontecimientos?

Las evaluaciones del presente dictamen se basan en el postulado de que el cumplimiento del Convenio de las Naciones Unidas para la Protección de los Derechos Humanos y de las Libertades Fundamentales es una condición absoluta, aunque insuficiente y mínima, y toman la comprensión jurídica de la UE del principio de precaución como un conjunto de valores y normas que deben respetarse. Además, las consideraciones se refieren a la ética de la responsabilidad de Hans Jonas como referencia filosófica, que también subyace al principio de precaución.[30]

Según el enfoque "heurístico del miedo" de Jonas, cualquier decisión humana debe basarse en primer lugar en las consecuencias potenciales para el futuro que podría conllevar dicha decisión. El motivo de Jonás "preservar la integridad de su mundo [del hombre] y de su ser contra las invasiones de su poder"[31] y su imperativo "Actúa de tal manera que los efectos de tu acción sean compatibles con la permanencia de la vida humana real en la tierra"[32] son un criterio útil también para la evaluación de IA. Puesto que desde este punto de vista sólo se puede deducir cómo no queremos vivir, sin embargo, también se deben hacer sugerencias positivas-normativas acerca de cómo se debe estructurar el manejo de la IA.

Entonces, ¿qué es el mundo en el que queremos vivir en el futuro? Los principios de Asilomar no responden a esta pregunta. Asumen un acuerdo generalmente válido y ampliamente aceptado sobre una concepción técnicamente optimista del futuro, que por un lado sigue siendo indeterminada y, debido a la falta de análisis socioeconómico, corre el riesgo de que sólo unos pocos la determinen, pero por otro lado, con la continuación de la búsqueda del

[30] Jonas (1979).
[31] Ibidem, p. 9.
[32] Ibidem, p. 35.

camino tecnológico, acepta la IA como un destino inevitable. Tomados en conjunto, los dos constituyen ya una amenaza social para la democracia, el estado de derecho y los derechos humanos.

- ¿Quién determina lo que es "bueno" cuando la tecnología lo abarca todo y afecta a todo el mundo, no sólo a aquellos que utilizan un producto basado en IA en particular?
- ¿Tenemos un consenso sobre los peligros/riesgos que estamos dispuestos a aceptar para beneficiarnos de la IA?
- ¿Cómo puede alcanzarse un consenso de este tipo a nivel mundial?
- ¿Seguirá existiendo este consenso, una vez establecido, en el futuro, cuando las cosas que resultan del consenso ya no puedan revertirse?

Si se estableciera un conjunto de reglas para el control de una IA fuerte, ¿quién garantizaría que la IA fuerte no anularía este conjunto de reglas y establecería sus propias normas (incluidas las éticas), y que lo haría a una velocidad de acción que imposibilitaría las reacciones humanas (contrarias) eficaces?

Los Principios Asilomar abordan una serie de cuestiones éticas relacionadas con la IA y describen las mejores prácticas moralmente derivadas relacionadas con la investigación y el desarrollo (I+D) de IA, lo que permite un amplio margen de interpretación. Los principios utilizan numerosos términos jurídicos indeterminados que tendrían que definirse para que puedan convertirse en un instrumento manejable. Esto plantea la cuestión del derecho de la definición.

Muchas formulaciones elegidas parecen asumir que los científicos encargados de I + D por IA están trabajando juntos en una manera no jérarquica y cooperativa, o que esto es posible, siempre que los investigadores tengan la buena voluntad para hacerlo. Ya aquí es preguntar:

- ¿Es este un punto de partida realista?

- ¿ En qué medida debe tenerse en cuenta desde el principio que la I+D también tiene lugar en contextos en los que las formulaciones externas de los objetivos (por ejemplo, el beneficio empresarial, la seguridad nacional) tienen al menos un impacto significativo, o determinan esencialmente la agenda de investigación?
- ¿Es realista que el uso de la IA puede ser controlada únicamente por los acuerdos voluntarios entre los investigadores sin la participación formal de las estructuras institucionales y los procesos del espacio político democráticamente constituido existente?

Los Principios Asilomar se dividen en secciones sobre temas de investigación, ética y valores, y temas a largo plazo.

3.2 Preguntas de investigación

Un punto de partida esencial de los principios es la exigencia de que la I + D sobre IA se enfoca de tal manera que sólo se crea IA "útil" . La cuestión de quién define cómo, si y cuando una IA es útil, sigue sin respuesta. La conexión entre "dirigido" y "beneficioso", que se construye en el Principio No. 1, es ilógica, ya que ambas son categorías diferentes.[33] Algo puede ser incontrolado, pero aún así útil, así como hay muchas cosas que surgen de una manera controlada, pero que son inútiles o dañinas. El objetivo de la investigación de IA debe ser, por lo tanto, tanto el control de la "inteligencia" como el de asegurar que ésta haga cosas ecológica y socialmente sostenibles y significativas en todo lugar y en todo momento.

Sobre esta base, habría que discutir los problemas fundamentales del utilitarismo, ya que el fin no justifica todos los medios, o sólo los justifica en situaciones de peligro extremo, especialmente si no hay (todavía) una comprensión básica común del fin, pero el propósito está quizás dominado sólo por unos pocos (por ejemplo, desde el punto de vista de los

[33] “The goal of AI research should be to create not undirected intelligence, but beneficial intelligence”.

productores de sistemas de armas autónomos mortales, estos últimos son ciertamente muy "útiles").

¿Quién determina entonces para qué personas, grupos o instituciones debe ser útil la IA para que se considere útil en general? Por lo tanto, IA puede conducir a un aumento de la eficiencia económica, pero la pregunta de si queremos que esta eficiencia se incremente socioculturalmente no puede ser respondida sólo por la suma de las decisiones de los consumidores. Hemos llegado a un punto de desarrollo tecnológico en el que es cuestionable en casos individuales si una mayor comodidad no exige un precio demasiado alto en términos de pérdida de capacidades y habilidades humanas.[34] Si decidimos que las ganancias de eficiencia son "netas", el segundo paso es tomar una decisión social sobre su distribución. Estas cuestiones deben ser tratadas principalmente por las disciplinas culturales, sociales y económicas.[35]

En general, se necesita una política de investigación legítima que defina de manera comprensible lo que significa en detalle la innovación éticamente responsable en el ámbito de la IA.

El Principio No. 2 trata de la necesidad de acompañar la investigación. Al igual que otros desarrollos técnicos, los sistemas de IA tienen un impacto en los procesos de desarrollo social y tienen una influencia fundamental en ellos. Por esta razón, es necesario incluir en la investigación cuestiones de irreversibilidad y de evaluación del impacto de los riesgos, a fin de comprender y evaluar mejor todos los desafíos éticos y sociales pertinentes. Esto también debe hacerse en el caso de las IA débiles y, en particular, antes de que se desarrolle una IA fuerte, y la investigación en este ámbito debe contar con suficientes recursos financieros. La

[34] Un ejemplo bien conocido de este fenómeno es la “auto-tortura” socialmente reconocida en el gimnasio, porque nuestros cuerpos no son lo suficientemente reivindicados fisicamente.
[35] Hasta ahora no hay evidencia de que el uso de IA va a llevar a menos desigualdad social. Al contrario, hay evidencia inicial de que va a llevar exactamente a un camino opuesto. Eso se está evidenciando generalmente en la digitalización de procesos económicos y entonces también en IA.

UE está presentando actualmente propuestas iniciales a este respecto, que deberán examinarse detenidamente.[36]

El Principio No. 3 habla de un intercambio "constructivo" y "saludable" entre los investigadores de IA y los responsables políticos.[37] Por el contrario, el principio de la democracia estipula que los poderes legislativo y ejecutivo no están en pie de igualdad con los grupos de representantes de la industria o de académicos. Más bien, los objetivos deben ser legitimados democráticamente y las violaciones de las reglas operativas para la implementación de estos objetivos deben poder ser sancionadas por el Estado y, en el caso de los aspectos transfronterizos, también multilateralmente.[38]

Los Principios 4 y 5 asumen que es posible crear una cultura de cooperación, confianza y transparencia entre los investigadores y desarrolladores de IA.[39] En esta formulación general, esto es realista sólo a nivel personal. Sólo en unas pocas instituciones (principalmente públicas) se paga a los investigadores para que trabajen juntos sobre la base de la confianza. De lo contrario, esto sólo se sanciona positivamente en la medida en que sirva para alcanzar los objetivos institucionales. En caso de conflicto, si chocan intereses institucionales que compiten o incluso se contraponen (por ejemplo, beneficios empresariales, seguridad nacional[40]), los investigadores tendrían que estar dispuestos a soportar sanciones negativas (pérdida de empleo y estatus, prisión) para poder cumplir con este principio. En el sensacional estudio "The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation" (El Uso Malicioso de la Inteligencia Artificial: Pronóstico, Prevención y

[36] Como parte de la estrategia exhaustiva de la UE sobre IA, que se ha publicado el 25 de abril de 2018. Comision de la UE (2018) y Krempl (2018).
[37] Vease también: Brundage et al (2018).
[38] La mencionada estrategia de la UE asume que progresos en el desarrollo de AI deben de ser controlados por un observatorio de IA. Comision de la UE (2018).
[39] Con similar tendencia: Executive Office of the President (2016), p 42, recomendación 19 y21.
[40] NSTC (2016), p. 3.

Mitigación[41]), 26 destacados desarrolladores de IA hacen un llamamiento a los investigadores y desarrolladores no sólo para que piensen en las consecuencias de su trabajo, sino también para que adviertan activamente a todos los actores relevantes de las consecuencias negativas y para que amplíen constantemente el círculo de aquellos que deberían participar y tomar decisiones informadas. Un dilema radica en el hecho de que, por un lado, las posibles consecuencias de la I+D y las aplicaciones concretas de la IA deben discutirse abiertamente, formalizarse y legitimarse, pero, por otro lado, la divulgación de los algoritmos subyacentes aumenta el peligro de que se haga un uso indebido.[42]

Los intentos de establecer normas éticas y legales para IA son inútiles mientras no se haya negociado todavía un borrador para el futuro ("cómo queremos vivir") y la visión distópica (teniendo en cuenta la concepción de Jonas del "efecto remoto de la tecnología": daños colaterales, así como los efectos sobre el futuro) no esté clara para todos. El principio es sintomático de esto: "Los equipos que desarrollan sistemas de IA deben cooperar activamente para evitar que se reduzcan las normas de seguridad". Las normas de seguridad (o las regulaciones y limitaciones legales generales) todavía no existen, y será difícil encontrarlas, especialmente mientras el desarrollo tecnológico sea mucho más rápido de lo que la legislación puede reaccionar y establecer marcos (el llamado problema del retraso cultural).[43][44]

[41] Ibidem.
[42] Brundage et al (2018).
[43] Ibidem
[44] A pesar de toda inseguridad: Se ha anunciado el desarrollo de una carta de ética de AI dentro de la iniciativa Europea exhaustiva. Se va a empezar a desarrollar a partir de principios de 2019 a base de un amplio debate.

3.3 Ética y Valores

Los Principios N.° 6 y N.° 7 tratan del requisito de la transparencia de errores y de la seguridad de funcionamiento y exigen una protección de seguridad completa durante todo el período de funcionamiento.[45] Las siguientes preguntas iniciales, por ejemplo, surgen aquí:

1) ¿Cómo puede aplicarse técnica y eficazmente la protección necesaria?
2) ¿Qué hacer en caso de decisiones ambiguas?
3) ¿Qué se debe hacer si las acciones de la IA deben medirse contra valores complejos, numerosos y/o contradictorios, cuya consideración efectiva, si es que la hay, sólo sería posible con la ayuda de una IA?
4) Los sistemas de IA no estarán claramente concentrados localmente, sino dispersos regional/globalmente en varios estados. En estos casos, la supervisión, la seguridad y la verificabilidad sin fisuras son extremadamente difíciles de garantizar.

Para documentar completamente las averías y optimizar las posibilidades de éxito de las reparaciones, los datos de diseño y los códigos fuente deben almacenarse en las instituciones de control gubernamentales para que sean accesibles al público a largo plazo. Esto también incluye datos y criterios de formación. El acceso al código fuente (y a los algoritmos) ha estado hasta ahora insuficientemente regulado. Sabemos que una liberación de código fuente con derechos de autor por parte de actores económicamente poderosos puede ser eludida con relativa facilidad. Por lo tanto, se necesita una legislación clara y rigurosa y una investigación exhaustiva para, al menos, reducir significativamente el riesgo de secreto del código fuente. Por otra parte, en virtud de la legislación actual, la reconstrucción de los códigos fuente a partir de códigos compilados disponibles al público (ingeniería inversa) está generalmente permitida (aunque a menudo, como en la UE, se limita a determinados casos). Esto a su vez facilita el desarrollo incontrolado de IA nuevas o mejoradas fuera de las estructuras conocidas con todas las posibles consecuencias asociadas. Por lo tanto, hay que

[45] Ver también: NSTC (2016).

satisfacer la demanda de introducción de derechos de información, así como las obligaciones de etiquetado y publicación, como las impuestas por la Federación de Organizaciones Alemanas de Consumidores (vzbv), por ejemplo.[46]

Una de las características esenciales de la IA fuerte así como la IA débil más desarrollado es que muestran un "comportamiento" nuevo e impredecible que no puede ser reconstruido a partir de los algoritmos debido a su complejidad. Los sistemas AlphaGo y AlphaGo Zero son ya ejemplos de ello. La exigencia de transparencia de los errores es, por tanto, contradictoria, ya que también dejará algunos errores inexplicables. Esto aumenta las exigencias de las normas de control y responsabilidad

El uso de IA en procedimientos judiciales, que se menciona en el Principio No. 8, tendrá que ser tratado exhaustivamente en el futuro. La motivación para el uso que ya está teniendo lugar en los procesos de toma de decisiones legales es el aumento de eficiencia deseado, especialmente en los procedimientos extensivos, así como el apoyo a las "mejores" decisiones. IA puede ser utilizado directamente en procesos de toma de decisiones legales para estructurar los hechos del caso, para preparar una propuesta de decisión, para dar reincidencia y pronósticos sociales o en casos extremos para tomar una decisión de forma autónoma. En la situación legal actual, esto sólo es posible dentro de unos límites muy estrechos en la mayoría de los países.

Consideramos problemático el uso de IA en la evaluación y la decisión en los procedimientos judiciales.[47] En particular, la demarcación de límites en el supuesto de relaciones de causalidad, así como la definición de los límites del libre albedrío, son un desafío. En los EEUU

[46] Federación Alemana de Organizaciones de Consumidores (2017).

[47] Incluso el uso de IA como fuente de información es objeto de controversia, especialmente en términos de su valor informativo. Por lo tanto, el Parlamento de la UE prveee para robots el estatus legal de una "persona electrónica". Parlamento Europeo (2017). A continuación, más de 200 científicos y empresarios (especialmente expertos en robótica) han respondido en una carta abierta:http://www.robotics-openletter.eu/

ya se está utilizando un software de clasificación para la recaída y el pronóstico social, en el que las evaluaciones iniciales muestran que existe un claro peligro de que conduzca a la discriminación por el color de la piel.[48]

Un ámbito de aplicación muy probable para la IA será la preparación de dictámenes periciales (por ejemplo, sobre la credibilidad de los testigos) o la prueba de la existencia o inexistencia de una reclamación. En todas partes en el área de uso como evidencia, un mínimo de procedimientos científicamente sólidos, transparencia y comprensibilidad deben ser mantenidos y probados por la IA Los medios metodológicos deben corresponder al estado actual de los conocimientos científicos en la materia. En general, la cuestión de la transparencia en lo que respecta a la introducción de datos y a las normas aplicadas es un problema importante. En el marco de un procedimiento, es concebible que la determinación de los hechos del caso pueda estar ya directamente influenciada por las conclusiones obtenidas de una IA.

En casos individuales, es de esperar un esfuerzo considerable en la trazabilidad de los resultados para poder rastrear las pruebas. Por encima de todo, se necesitan normas jurídicas que, en todos los casos de oposición, nombramiento o revisión, prevean la revisión obligatoria por parte de las autoridades superiores, cuyo personal está compuesto exclusivamente por personas. Además, los jueces (así como el personal judicial, abogados, fiscales y notarios) deben ser formados y educados en consecuencia y sensibilizados sobre los peligros de una creencia "ciega" en la tecnología. En la actualidad, tendemos a pedir con más fuerza una prohibición general de la IA en los procedimientos judiciales.

El Principio No. 9 describe las responsabilidades de la IA avanzada. Se afirma que los constructores de IA deben asumir la responsabilidad como "partes interesadas" de las consecuencias previstas y no previstas (implicaciones). La formulación implica que los

[48] Angwin/Larson/Mattu/Kirchner (2016).

sistemas menos avanzados no están sujetos a esta responsabilidad.[49] La cuestión de la rendición de cuentas (es decir, también la causalidad) es difícil en sistemas complejos. Este problema se potencia en los sistemas globalizados. Además, casos análogos (por ejemplo, accidentes en instalaciones industriales) muestran que los grupos de interés financieramente fuertes y poderosos implicados están generalmente comprometidos con la prevención de la ejecución de las reclamaciones de responsabilidad. Las negociaciones multilaterales posteriores a 1992, en particular en el marco del Convenio sobre la Diversidad Biológica (CDB) (por ejemplo, el Protocolo sobre Seguridad de la Biotecnología), han demostrado que las cuestiones de responsabilidad (en particular la regulación de la carga de la prueba) son un instrumento regulador decisivo y pragmático. La cuestión de quién es responsable de qué daños y cuándo tiene repercusiones, por ejemplo, en las estrategias de gestión de riesgos y en los necesarios controles de diligencia debida, hasta el nivel de las decisiones de inversión.

El Principio 12 reclama el derecho a la protección de los datos personales, pero lo limita a los datos generados por los usuarios. Por otra parte, no se incluyen los datos recogidos sobre estos usuarios. Por lo tanto, la mayor parte de los datos queda fuera del control de aquellos sobre los que estos datos dicen algo. Para IA, esto también incluye la información relevante para las personas que la IA genera a través de la asociación/enlace/combinación de datos.[50] Debe garantizarse el derecho de acceso sin restricciones y con barreras mínimas a todos los datos personales que se refieran a uno mismo. Este derecho debe incluir el consentimiento previo explícito para la recogida y el establecimiento de redes, así como el derecho a la

[49] Por el contrario, el servicio científico del Parlamento Europeo ve básicamente siempre la necesidad de definir responsabilidades a priori: EPRS (2016).

[50]Además de los datos quelas personas de manera directa o indirecta revelan (como en las redes sociales, a través del uso de motores de búsqueda y el Internet de las cosas) y contra el abuso cada uno se defender, la importancia de la recogida de datos por sensores (empezando por las cámaras de vigilancia) no debe subestimarse. La evolución actual en la dirección de "Neuro-data" se tiene que observar.

integridad, la transparencia y el derecho a la supresión (el derecho a ser olvidado) si no existen razones jurídicas imperiosas (por ejemplo, la ocultación de un delito).[51]

La protección de la privacidad y de los datos personales está en el centro del Principio No. 13, que establece que el uso de IA puede restringir los derechos humanos, pero no de una manera que se caracterice como "irrazonable". Una posible restricción de la libertad humana individual y colectiva debe evaluarse siempre en el contexto de una posible amenaza a otros valores fundamentales. En principio, la protección de datos y la privacidad son valores básicos esenciales y garantizados de la sociedad, que, especialmente en el ámbito de la IA, deben diseñarse de tal manera que no se limiten ni en parte ni en su totalidad sin necesidad existencial. Los siguientes principios básicos son necesarios para ello:

- Variantes de privacidad por diseño o diseño basados en este concepto, que ya tienen en cuenta la privacidad y la protección de datos en el proceso de diseño técnico y las implementan o habilitan técnicamente.
- En principio, las opciones de exclusión y de inclusión (opt in and opt out) deberían ser obligatorias. Esto también debe incluir la vinculación de los datos.
- La decisión básica sobre el uso de los datos debe recaer en el ser humano individual. Esto requiere una regulación clara que también hace que BigData sea transparente para el individuo.
- La libertad de información debe estar diseñada legalmente de tal manera que los recopiladores de los datos también tengan que proporcionar los datos de enlace, incluidos los algoritmos subyacentes.
- Por medio de la responsabilidad y el derecho penal, el control y la recopilación de datos pueden tener que ser diseñados por terceros de tal manera que el individuo reciba una protección legal efectiva contra ellos.

[51]La demanda general por el derecho a ser olvidado plantea problemas legales. Por ejemplo, ex empleados del servicio de seguridad del estado de la anterior RDA demandaron que sus datos sean borrados con lo que fracasoron delante de un tribunal por razones comprensibles.

El Principio No. 14 quiere empoderar a las personas y afirma el objetivo de que la IA beneficie al mayor número de personas posible. El "empoderamiento" debe entenderse aquí como el fortalecimiento de la capacidad de acción de las personas. En primer lugar, es fundamental que el hombre siga siendo el sujeto de sus acciones, con la máxima autodeterminación, y que su acceso a la participación económica, social y política no se vea debilitado por IA, sino que más bien tienda a ser fortalecido. Para garantizar que el mayor número posible de personas se beneficien de las tecnologías de IA, se requieren opciones y decisiones económicas, sociales y políticas adecuadas, así como culturales.

En vista de las grandes promesas hechas por IA, la demanda política de que IA haga una contribución visible a la reducción de la desigualdad económica y social entre las personas es legítima. Sin embargo, la formulación elegida del principio encierra el peligro de que el bien común sólo se entienda como la suma del interés propio maximizado. En lugar de ello, se debe exigir que IA sirva en principio al bien común.[52]

El Principio No. 15 requiere que la "prosperidad económica creada por la IA" sea ampliamente compartida. Como postulado político, esto presupone en última instancia un sistema económico diferente, que, sin embargo, no se exige en el texto y, por lo tanto, probablemente no se pretende. A pesar de todas las precauciones en las predicciones, los estudios iniciales disponibles (por ejemplo, de la Universidad de Oxford) muestran que la cuarta ola de automatización de la producción que ha sido posible gracias a la creación de redes y a IA llevará a la destrucción "neta" de millones de puestos de trabajo, incluso teniendo en cuenta los numerosos nuevos puestos de trabajo creados.[53] Esto afectará

[52] Como parte de la estragegia eshaustiva de la UE (Comisión de la UE (2018), Heise (2018)), se propone una plataforma "IA a la carta" que adicionalmente a nuevos centros de excellencia y centros de innovación digital, tiene como fin la participación de pequeñas y medianas empresas en el desarrollo.

[53] McIAnsey (2017): En 2030 hay el riesgo de una pérdida bruta alrededor del 15% (máximo hasta el 30%) de todos los puestos de trabajo en todo el mundo (alrededor de unos 400 millones de puestos de trabajo, un máximo de 800 milliones); Frey / Osborne (2013).

principalmente a las personas menos cualificadas y a las mujeres para las que no existen alternativas.[54]

El Principio Nº 18 trata del peligro de una carrera de armamentos de los sistemas letales de armas autónomas. Por lo general, no están prohibidas. Sólo se advierte del peligro de una carrera armamentista. Los Estados Unidos, el Reino Unido, Rusia, China, Israel y Corea del Sur ya están desarrollando sistemas de armas de este tipo y ya están en uso estacionario en la frontera con Corea del Norte. En vista de la dinámica aceleración en el desarrollo de sistemas de armas autónomas letales, faltan declaraciones claras sobre la manera de evitar una carrera de armamentos. El uso de unidades autónomas en situaciones de guerra o combate es un riesgo, también para los usuarios, debido a la falta de comportamiento ético y a posibles desviaciones.[55] Además, la redacción del Principio 18 muestra que incluso las aplicaciones militares de IA no desencadenan un "reflejo de prohibición".[56]

Además, los principales investigadores de IA ven el peligro de ataques agresivos que ya están creciendo rápidamente por debajo del umbral de los sistemas de armamento (por ejemplo, a través del sabotaje de vehículos o de instalaciones a gran escala, así como el debilitamiento selectivo de las democracias y de la autoridad del Estado), ya que los costes de tales misiones están disminuyendo constantemente y el atacante es extremadamente difícil de detectar.[57] [58]

3.4 Problemas a largo plazo

A falta de consenso, el Principio 19 pide que se abandonen las suposiciones sobre si existen y cuáles son los límites técnicos para un mayor desarrollo de la IA y, por lo tanto, evita la

[54]Frey / Osborne (2013) p. 36ss; WEF (2018).
[55] Scott, Ben et al (2018) y Allen Gregory et al (2018).
[56]Sin embargo, el Future of Life Institute se compromete. Durante las negociaciones sobre la Convención sobre Ciertas Armas Convencionales / Grupo de Expertos Gubernamentales sobre Sistemas Autónomas de Armas Letales, en noviembre de 2017 presentó un video de "choque" que ilustra los peligros de los sistemas de armas autónomas de forma espectacular.
[57] Allen, Gregory et al (2018).
[58]El tema se aborda también en: Oficina Ejecutiva del Presidente (2016), p. 42, recomendación 22 y 23.

definición de "líneas rojas" absolutas para un mayor desarrollo de la IA. Por otra parte, el principio de cautela exige precisamente eso. Si no sabemos si surgirá una IA fuerte y cuándo, y tampoco sabemos si esto podría ser peligroso y con qué consecuencias, las sociedades y los Estados deben ponerse de acuerdo sobre líneas rojas de antemano y aplicarlas de manera efectiva.

El Principio No. 20 establece que la IA avanzada puede cambiar profundamente el futuro de la vida en el planeta tierra.[59] A esta posibilidad fundamental se opone la exigencia de planificar y desarrollar "adecuadamente". "Adecuado" es un término legal indefinido - ¿quién determina qué es "adecuado" y cuáles son las normas? El principal problema del principio es que no se pregunta si es permisible en absoluto aceptar cambios profundos en la historia de la vida en la tierra sin presentar razones convincentes, cuya dirección y alcance la gente no conoce y no puede prever. El principio se basa obviamente en el entendimiento de que todo lo que es posible también se hace. Así que si el hombre es capaz de desarrollar una IA fuerte, entonces lo hará también, simplemente porque puede.[60]

La historia, por otra parte, muestra (aunque pocas) excepciones en las que las sociedades, los Estados o las comunidades de Estados han optado por no seguir ciertos caminos tecnológicos porque los riesgos asociados a ellos se consideraban inaceptables (por ejemplo, la fisión nuclear o los combustibles fósiles como fuentes de energía, las armas biológicas y químicas, los CFC, etc., como refrigerantes). Sin embargo, esto sólo sucedió cuando grandes sectores de la población (especialmente las élites) se negaron por una clara mayoría a aceptar (seguir aceptando) ciertos riesgos asociados con el uso de estas tecnologías. Por regla general, esto sólo ocurría después de que se produjeran consecuencias indeseables. También hay ejemplos de que el "abandono tecnológico" de un actor no llevó a otros a seguir el ejemplo, sino que,

[59]La palabra que se usa aquí "historia", es engañosa y no encaja, ya que la historia ya ha pasado. "Curso de la vida" sería más apropiado. Al menos digno de discusión es la idea subyacente teleológica de un destino de la existencia humana.
[60] Ver. Beck (1986).

por el contrario, se benefició de él (actualmente: el uso pacífico de la energía nuclear). Sin embargo, esto no puede justificar una negligencia grave en relación con el principio de precaución.

En el Principio 21 se pide que la planificación y los esfuerzos de mitigación de riesgos sean proporcionales al impacto esperado de los riesgos existenciales y los posibles desastres. Es importante señalar que el principio implica que existen riesgos existenciales y que los desastres son posibles. Estos deben ser atendidos con los mejores "esfuerzos". "Pero "esfuerzos" no son sinónimo de éxito. Es decir, con la formulación del principio se acepta que estos esfuerzos deben ser apropiados en la medida del riesgo, pero que en última instancia pueden fracasar.

En contraste con el Principio 21, la redacción del Principio 22 es apropiada para el tema en cuestión. Sin embargo, también permiten que continúe la investigación y el desarrollo y la aplicación de IA "fuertes", siempre que estén "sujetos a estrictas medidas de seguridad y control". No se considera la posibilidad de una prohibición absoluta, es decir, una prohibición general de IA "fuerte". Incluso si tal prohibición se pudiera hacer cumplir, seguiría existiendo el peligro de que en algún lugar del mundo, en aras de la ventaja, la gente hiciera caso omiso de las prohibiciones, por ejemplo para obtener poder político y/o ideológico/religioso. No sólo el presidente ruso Vladimir Putin ya ha declarado públicamente que el poder que primero tenga una IA "fuerte" dominará el mundo.[61] En particular, hay que tener en cuenta el trabajo en red global, con el que estos sistemas pueden acceder a "medios ejecutivos" como las armas (amenaza directa), a instalaciones que pueden utilizarse como armas, como las centrales nucleares, y a infraestructuras críticas, como las redes eléctricas (amenaza indirecta).

[61] RT.com (2017).

El Principio No. 23 habla directamente de una "superinteligencia" que "debe servir a los ideales éticos generalizados de toda la humanidad y no sólo a un Estado o a una organización". En primer lugar, sin embargo, los seres humanos no vivimos en una sociedad mundial, ideal (paradisíaca), que presumiblemente nunca existirá, debido al cambio permanente de toda la vida. Aquí surge la pregunta: ¿Por qué debemos permitir la creación de la superinteligencia? ¿No nos convertiríamos entonces en creyentes de una religión que entiende al hombre como una etapa intermedia necesaria de la evolución hacia seres más elevados y digitales?[62] No podemos y no debemos querer eso. Una superinteligencia podría posiblemente decidir sobre el futuro destino de la humanidad en el marco de su programación básica desarrollada de forma independiente. Esto no tendría que significar que la humanidad sería destruida o esclavizada, sino que nosotros, los humanos, sólo tendríamos oportunidades de participar "prestadas/otorgadas". Una afirmación violenta de los intereses humanos contra la superinteligencia estaría con toda probabilidad condenada al fracaso.

Al mismo tiempo, queda abierta la cuestión de cuáles son estos "ideales generalizados". Por ejemplo, "ser rico" o "consumir" o "disfrutar" o "divertirse" son ideales muy extendidos en algunas partes del mundo. Pero la aplicación literal de las normas religiosas es también un ideal muy extendido en algunas partes del mundo. Por otra parte, también hay ideales que se excluyen mutuamente y que se extienden por todo el mundo, como "el derecho de autodeterminación de la mujer a abortar" y "la protección de la vida no nacida desde la concepción". Para ambos encontramos más de mil millones de seguidores. Incluso la aplicación consecuente de la Declaración de los Derechos Humanos de la ONU sigue siendo rechazada por un gran número de personas, al menos en parte. Entonces, ¿cuál es la base legitimadora del término "ideales éticos generalizados"?

[62]Ver. Hariri (2017), p. 497ss. y Dotzauer (2017).

3.5 Conclusión

Los autores de los Principios de Asilomar y del Grupo de Estudio de la VDW están de acuerdo en que la IA cambiará profundamente la vida en la Tierra y que es de esperar la creación de una IA fuerte. Al menos hasta cierto punto, también están de acuerdo en que representa un peligro potencial significativo para el futuro de la humanidad. Los Principios Asilomar son un excelente punto de partida para las discusiones sobre cómo explotar el potencial de la inteligencia artificial en los próximos años. Sin embargo, los principios no proporcionan un marco normativo adecuado para la necesaria definición de los límites absolutos de la investigación, el desarrollo y la aplicación de la IA, ni para la aplicación de tales límites por motivos de seguridad.

El grupo de estudio de la VDW considera problemático que los principios se hayan elaborado sobre la base de una premisa (implícita) de utopía exclusiva, que parte del supuesto de que, en principio, es posible el desarrollo o la aplicación ilimitada de la tecnología, lo que sólo requiere una regulación en casos concretos. Así, se requirió una aprobación del 90% en el diseño de los principios de Asilomar para incorporar un principio regulador. Si, por otra parte, una premisa sobre los peligros potenciales correspondiera al principio de precaución de la UE, la aprobación del desarrollo tecnológico o de la aplicación requeriría una aprobación del 90%. El principio de precaución aplicable en la UE debería, por tanto, servir de guía para todos los futuros debates.[63]

[63] Si un riesgo significa que eventos catastróficos o existencialmente amenazantes pueden ocurrir como resultado da la investigación y desarrollo o aplicación de sistemes de IA y estos riesgos son son despreciabes, entonces los riesgos no pueden ser objeto "los esfuerzos de planificación y reducción de riesgos", sino deben encadenar inmediatamente y sin demora todas las medidas de seguridad necesarias.

4. Recomendaciones para la acción

El Grupo de Estudio de la VDW sobre la Evaluación Tecnológica de la Digitalización está convencido de la necesidad de un debate social y científico amplio y específico sobre cómo hacer frente a los retos que plantean la I+D y las aplicaciones de la inteligencia artificial. Existen enfoques iniciales al respecto, pero en nuestra opinión siguen estando demasiado dominados por el entusiasmo técnico y los ya visibles y, en el futuro, claramente crecientes aumentos de la productividad económica. El punto de vista fundamentalmente positivo de muchas voces se apoya a menudo en la convicción de que la única preocupación es evitar o al menos reducir las consecuencias indeseables de la IA, pero no en la cuestión de si, en principio, deberían prohibirse al menos ciertas aplicaciones de la IA débil y la creación de una IA fuerte. En general, prevalece la impresión de que las promesas del IA no deben ser empañadas por consideraciones dudosas o pesimistas.

Sin embargo, las razones de ello también residen en el hecho de que la mayoría de las consecuencias negativas imaginables no se han producido todavía o son conocidas y, en el caso de un IA fuerte, probablemente residan en un futuro que durará años o décadas. Los peligros son abstractos, invisibles y en gran medida desconocidos. Nuestra imaginación de este futuro está formada por la literatura y el cine de ciencia ficción. Por lo tanto, las voces de alerta no se toman en serio en este momento. El futuro está aparentemente muy lejos, demasiado lejos para los debates políticos y sociales que giran en torno a los problemas de hoy y del futuro próximo. También destacaremos los factores de conexión más importantes para lo que hay que hacer.

Discurso científico e investigación sobre la evaluación de tecnologías de IA

Hasta ahora, la investigación básica que aborda las consecuencias tecnológicas y las cuestiones de seguridad, especialmente en lo que se refiere a la interacción hombre-máquina en particular y a la interacción máquina-entorno en general, sólo ha estado en sus inicios. Sin embargo, existe una necesidad considerable de investigación pública; los procesos de

investigación deben diseñarse y promoverse intensamente en lo que respecta a la evaluación de riesgos, las interdependencias sociales y las futuras medidas de desarrollo, de forma que se orienten a la aplicación y acompañen al desarrollo técnico.

La financiación de la investigación (con fondos públicos) necesaria para este fin debe basarse en una investigación básica interdisciplinaria y transdisciplinaria en los ámbitos del derecho, la ética, las ciencias sociales y económicas, la informática, pero también las ciencias de los medios de comunicación, la tecnología y la psicología, en particular en lo que se refiere a los posibles diseños técnicos, tanto de manera específica como global (también para garantizar la protección de los derechos fundamentales).[64] [65] Uno de los objetivos debe ser desarrollar directrices transversales concretas, incluyendo la programación y la tecnología, para el proceso de diseño técnico.

La cooperación y los procesos formales y transdisciplinarios de comunicación científica en la investigación y la enseñanza (por ejemplo, mediante actos y publicaciones conjuntos) promoverán una rápida penetración, la fertilización mutua y la difusión pública de los conocimientos adquiridos.

Discurso social y político

Los ciudadanos, los responsables de la toma de decisiones y los multiplicadores deben estar familiarizados con información pertinente y científicamente sólida sobre la IA. Son necesarios esfuerzos considerables para preparar la información necesaria y debatirla en los foros pertinentes. En particular, los responsables políticos, jurídicos y económicos deben estar en condiciones de tomar decisiones con conocimiento de causa. Esto también se aplica a los

[64] Oficina Ejecutiva del Presidente (2016), p.42; la Recomendación 18 también exprime la responsabilidad de las escuelas.

[65] Por ejemplo se tendrían que realizar proyectos pilotos, que tienen lugar en un área claramente definida y por lo tanto permiten una revisión de los mecanismos de acción, así limitando las consecuencias potencialmente irreversibles, inclueyndo mecanismos independientes de vigilancia.

medios de comunicación y a las instituciones sociales, así como a las áreas subordinadas y relevantes del gobierno (sobre todo la seguridad y la protección del consumidor). Las conferencias y eventos que son adecuados para este propósito también deben ser utilizados para llamar la atención sobre las cuestiones esenciales sobre el tema de IA.

Queremos contribuir a iniciar las discusiones y negociaciones necesarias para prevenir eficazmente los peligros imaginables del IA (especialmente a través de una IA fuerte) o, cuando esto no sea posible/necesario, minimizarlos. A este respecto, hay que tener en cuenta que una gran parte de las actividades de I+D en IA no tienen lugar bajo control estatal, en el marco de la competencia mundial, y que la investigación militar en IA sólo está sujeta, al menos en parte, a un control democrático limitado. En particular, la participación de la sociedad civil y de los individuos en la orientación y configuración de estas importantes cuestiones puede ser, por tanto, un paso crucial para que las acciones necesarias que se exponen a continuación sean un éxito.

Regulación

Como en todas las áreas de investigación, la I+D de IA debe seguir principios normativos (éticos y legales).[66] Desde el punto de vista de la VDW, la base debe ser la Declaración Universal de Derechos Humanos o su codificación en la legislación nacional. Además, las normas jurídicas existentes, como el principio de precaución, deben ampliarse explícitamente a los avances técnicos y hacerse jurídicamente vinculantes (principio de precaución 2.0).[67] En este sentido, las evaluaciones del riesgo y del impacto tecnológico deben ser vinculantes. Se

[66] NSTC (2016).

[67] El principio de precaución debe desarrollarse más. El principio de precaución y la innovación no son mutuamente excluyentes. Por el contrario: Precauciones para el futuro en particular deben ser motivo de innovaciones que promuevan la sostenibilidad. Mantener abiertas y promover alternativas de menor riesgo, y la creación de múltiples opciones para el futuro es demasiado poco. La investigación y la innovación deben ser conscientes de sus responsabilidades aquí y entrar pronto en un diálogo con la sociedad sobre la evaluación de la tecnología. A lo que se refiere a riesgos y oportunidades, la falta de alternativas también tiene que ser considerada.

necesitan normas codificadas que abarquen todas las cuestiones jurídicas pertinentes a todos los niveles nacionales e internacionales necesarios;[68] incluidas las prohibiciones o moratorias. La estandarización debe tener lugar dentro de los respectivos contextos de aplicación, por un lado, y por otro, en la sociedad en su conjunto. Esto incluye también la petición de que el principio de cautela se desarrolle más en términos jurídicos cuando sea necesario en interés de una prevención global de los riesgos. Dado que el desarrollo de sistemas jurídicos internacionalmente válidos equipados con instrumentos de ejecución lleva décadas en lugar de años, la labor debe comenzar de inmediato.

Se necesitan estructuras estatales (y multilaterales) y mecanismos de sanción eficaces para garantizar que la IA sea controlable en todo momento. Esto se aplica a todas las fases de I+D y aplicación. En particular, los lanzamientos al mercado sólo podrán tener lugar después de que se hayan realizado suficientes evaluaciones de seguridad. Esto requiere, por ejemplo, pruebas exhaustivas en escenarios realistas. Se necesitan conocimientos técnicos especializados para prestar apoyo, pero, al igual que los intereses económicos, no deben tener una influencia significativa en las cuestiones de reglamentación.[69] También se necesita un control democrático, técnicamente informado, exhaustivo y global de la investigación y el desarrollo en este ámbito.

La financiación de la investigación descrita anteriormente también debe abordar la identificación y la elaboración científica de modelos jurídicos complementarios, posiblemente necesarios, de regulación eficaz que ayuden a reconocer los efectos de la IA tan pronto como sea posible y a forzar todas las reacciones necesarias con prontitud.

Los comités de expertos y los representantes de la sociedad civil deberían apoyar esta iniciativa. Además, se requieren compromisos éticos para tratar con IA en I+D y aplicación.

[68] Consideraciones iniciales para esto: EPRS (2016).
[69] Menos crítica: Oficina Ejecutiva del Presidente (2016), p.40, recomendación Nº 5 y 6.

La Comisión Europea comienza a publicar su documento al mismo tiempo que su amplia iniciativa europea sobre la IA, que incluye planes para una alianza europea para la IA y un enfoque ambicioso de la IA, que debería convertir a la UE en un "líder de la revolución de la IA"[70] Por otro lado, las indicaciones de la estrategia de posibles distorsiones suenan como concesiones forzadas.[71] También hay demandas iniciales, tanto del Parlamento de la UE (recomendaciones al COM del 27 de enero de 2017) como de la propia Comisión de la UE, de "iniciar inmediatamente todos los trabajos necesarios para desarrollar un marco legal y ético internacionalmente reconocido para el diseño, la producción, el uso y la gobernanza de las IA".[72] Esto también se requiere de la comunidad científica en concreto: Esto también es exigido concretamente por el lado científico: La UE debería elaborar inmediatamente una Carta de la IA y trabajar en la elaboración de una Carta mundial I.[73]

IA - Mecanismos de seguridad inherentes

Además, para la creación de cualquier IA se aplica lo siguiente: el requisito básico es la programación irrevocable de principios éticos que permanezcan funcionales en todos los modos de funcionamiento imaginables. En situaciones en las que esto no está (ya no está) garantizado y se requiere la intervención humana para prevenir daños, todas las acciones necesarias deben ser posibles en todo momento de la manera más oportuna y profiláctica posible. La pauta más importante debe ser que IA no puede dañar a un ser humano bajo ninguna circunstancia concebible. Esto corresponde a las leyes de robot de Asimov y es un prerrequisito absoluto para la "utilidad" de IA. Asimov mismo caracteriza sus leyes como necesarias pero no suficientes.[74] Los principios éticos de los algoritmos también son al menos problemáticos porque no tienen una "personalidad del ego" que pueda tener la experiencia

[70] Comisión de la UE (2018), p. 14
[71] Ibidem
[72] UE DG R + I (2018)
[73] Ver en detalle Metzinger (2018).
[74] Asimov (1950).

del nacimiento, la alegría, el dolor, la enfermedad y la muerte. Si esto ocurriera algún día, nos enfrentaríamos a retos muy diferentes.

Caso especial de los sistemas de armas autónomas letales

Por lo que respecta al uso militar de sistemas de armas autónomas letales, debe basarse en el (no) documento de trabajo franco-alemán sobre las primeras negociaciones formales de las Naciones Unidas sobre sistemas de armas autónomas letales.[75] En ella, ambos Estados miembros de la UE proponen conjuntamente una declaración política a nivel de las Naciones Unidas que establece los primeros pasos hacia un protocolo internacional en el marco de la "Convención sobre la prohibición o restricción del empleo de ciertas armas convencionales que puedan causar sufrimiento excesivo o tener efectos indiscriminados."[76] De hecho, esto significaría un hechizo. También en este caso el éxito dependerá sobre todo de que todos los Estados lleguen a la conclusión de que las armas autónomas letales no pueden ser controladas de forma segura por sus usuarios. Entonces, sólo los productores de sistemas de armas autónomas letales tendrían interés en impedir una prohibición efectiva. El documento del Dr. Alexander Kott, jefe de la División de Ciencias de Redes del Laboratorio de Investigación del Ejército, que hace un llamamiento a los esfuerzos masivos de investigación en los EEUU en el campo del desarrollo de sistemas de armas autónomos, muestra hasta dónde llega este camino, ya que sólo los sistemas de armas autónomos serían capaces de reaccionar adecuadamente a la futura interacción de la "internet autónoma de las cosas de combate" en el campo de batalla.[77]

[75]Las negociaciones se llevaron a cabo en noviembre de 2017 dentro de la Convención sobre Armas Convencionales: Group of Governmental Experts (GGE) on lethal autonomous weapons systems (LAWS). Group of Governmental Experts of the High Contracting Parties (2017) e International Committee of the Red Cross (2004). En cuanto a la posición de China vease Kania (2018).
[76] Naciones Unidas (1980).
[77] Kott (2018).

Invitación al diálogo

Las tareas que tenemos por delante sólo pueden ser dominadas conjuntamente. Por lo tanto, ofrecemos nuestro apoyo a todos aquellos que estén dispuestos a utilizar las oportunidades y posibilidades de IA sólo en la medida en que no ponga en peligro la salud humana, la vida o el medio ambiente y no perjudique el bien común. Por debajo del umbral de los riesgos existenciales, habrá justificadamente un intenso debate social y político sobre lo que beneficia al bien común o lo que al menos no lo perjudica o debe considerarse perjudicial. Esperamos con interés entablar un amplio diálogo sobre esta cuestión.

5. Bibliografía

Angwin, Julia/ Larson, Jeff/ Mattu, Surya/ Kirchner, Lauren (2016): Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks, in ProPublica, May 23, 2016; https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Asimov, Isaac (1950): I, Robot. Gnome Press.

Bahr, Egon/ Lutz, Dieter S. (Hrsg.) (1992): Gemeinsame Sicherheit. Idee und Konzept. Bd. I: Zu den Ausgangsüberlegungen, Grundlagen und Strukturmerkmale Gemeinsamer Sicherheit, Baden-Baden.

Beck, Ulrich (1986): Risikogesellschaft. Auf dem Weg in eine andere Moderne. Erstausgabe, 1. Auflage. Suhrkamp, Frankfurt am Main.

BITKOM (2017): Künstliche Intelligenz verstehen als Automation des Entscheidens – Leitfaden, Berlin.

Brundage, Miles, et al (2018): The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation; https://www.eff.org/files/2018/02/20/malicious_IA_report_final.pdf.

Centre on Impact of IA and Robotics (der UNSW Sydney) (2018): Open Letter to Professor Sung-Chul Shin, president of KIAST from some leading IA researchers in 30 different countries; https://www.cse.unsw.edu.au/~tw/ciIAr/kIAst.html.

Dotzauer, Gregor (2017): Näher, mein Bot, zu dir. In: Tagesspiegel 12.12.2017.

Eberle, Ute (2015): Sprachassistenten verändern unser Leben. In: Wirtschaftswoche 28.7.15.

EFI (Expertenkommission Forschung und Innovation) (2018): Gutachten zu Forschung, Innovation und technologischer Leistungsfähigkeit 2018, Berlin.

EPRS (2016): European Parliamentary Research Service, Scientific Foresight Unit (STOA), PE 563.501: Ethical Aspects of Cyber-Physical Systems, Scientific Foresight study, Brussels.

EU-DG R+I (2018): European Commission, Directorate-General for Research and Innovation; European Group on ethics in Science and New Technologies: Statement on Artificial Intelligence, Robotics and "Autonomous" Systems, Brussels.

EU-KOM (2018): Maximizing the benefits of Artificial Intelligence (Version 15 – 27/02/2018). Unpublished Working Paper, Brussels.

EU-Parliament (2017): European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics; http://www.europarl.europa.eu/sides/getDoc.do?type=TA&reference=P8-TA-2017-0051&language=EN&ring=A8-2017-0005.

Executive Office of the President (2016): Preparing for the Future of Artificial Intelligence. Washington, D.C.

Experts on Lethal Autonomous Weapons Systems (LAWS), Geneva, CCW/GGE.1/2017/WP.4; http://undocs.org/ccw/gge.1/2017/WP.4.

Frey, Carl Benedikt/ Osborne, Michael A. (2013): The Future of Employment: How Susceptible are Jobs to Computerisation, Oxford University.

Future of Life Institute (2017): Asilomar IA Principles; https://futureoflife.org/IA-principles/ und Beneficial IA 2017. Conference Schedule; https://futureoflife.org/bIA-2017/

Göttinger Achtzehn (1957): Göttinger Manifest; https://www.uni-goettingen.de/de/text+des+g%c3%b6ttinger+manifests/54320.html.

Group of Governmental Experts of the High Contracting Parties (2017): To the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. For consideration by the Group of Governmental Experts on Lethal Autonomous Weapons Systems (LAWS). Submitted by France and Germany, CCW/GGE.1/2017/WP.4, Geneva.

Harari, Yuval Noah (2017): Homo Deus – Eine Geschichte von Morgen, München.

Henk, Malte (2014): Jugend ohne Sex. In: Zeit online vom 15.6.14.

Independent Commission on Disarmament and Security Issues (1982): Common security: A blueprint for survival, Simon and Schuster, New York.

International Committee of the Red Cross (2004): Convention on Prohibitions or Restrictions on the Use of CertIAn Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, Geneva.

Jonas, Hans (1979): Das Prinzip Verantwortung. Versuch einer Ethik für die technologische Zivilisation. 1. Auflage, Insel-Verlag. Frankfurt am Main.

Kania, Else (2018): Artificial Intelligence: China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapons Systems. In: Lawfare, April 17, 2018; https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

Kommission der Europäischen Gemeinschaften (2000): Mitteilung der Kommission: die Anwendbarkeit des Vorsorgeprinzips, KOM (2000) 1 endgültig, Brüssel; https://eur-lex.europa.eu/legal-content/DE/TXT/PDF/?uri=CELEX:52000DC0001&from=DE.

Kott, Alexander (2018): Challenges and Characteristics of Intelligent autonomy for Internet of Battle Things in Highly Adversarial environments, Adelphi, MD; https://arxiv.org/ftp/arxiv/papers/1803/1803.11256.pdf.

Krempl, Stefan (2018): Künstliche Intelligenz: EU-Kommission plant umfassende europäische Initiative. In: heise online 26.3.2018.

McKinsey (McKinsey Global Institute), (2017): Jobs Lost, Jobs Gained: Workforce Transitions in a time of Automation.

Metzinger, Thomas (2018): Towards a Global Artificial Intelligence Charter. In: European Parliamentary Research Service: Should we fear artificial intelligence? Brussels.

Müller, Vincent/ Bostrom, Nick (2013): Future progress in artificial intelligence: A Survey of Expert Opinion. In: Vincent C. Müller (ed.): Fundamental Issues of Artificial Intelligence, Berlin.

NSTC (National Science and Technology Council) (2016): The National Artificial Intelligence Research and Development Plan, Washington, D.C.

Oye, Kenneth A., et.al. (2014): Regulating gene drives. Regulatory gaps must be filled before gene drives could be used in the wild, in: Science 17 Jul 2014; http://science.sciencemag.org/content/early/2014/07/16/science.1254287.full.

RT.com (2017): 'Whoever leads in IA will rule the world': Putin to Russian children on Knowledge Day, 1 Sep, 2017 14:08; https://www.rt.com/news/401731-IA-rule-world-putin/.

Schellnhuber, Hans Joachim (2015): Selbstverbrennung: Die fatale Dreiecksbeziehung zwischen Klima, Mensch und Kohlenstoff, C. Bertelsmann Verlag, München.

Scott, Ben/ Heumann, Stefan/Lorenz, Philippe (2018): Artificial Intelligence and Foreign Policy. Stiftung Neue Verantwortung, Berlin.

TAB (Büro für Technikfolgenabschätzung beim Deutschen Bundestag) (2016): Technologien und Visionen der Mensch-Maschine-Entgrenzung. Sachstandbericht zum TA-Projekt „Mensch-Maschine-Entgrenzungen: zwischen künstlicher Intelligenz und Human Enhancements. Arbeitsbericht Nr. 167, Berlin.

United Nations (1948): Universal Declaration of Human Rights, Paris; http://www.un.org/en/universal-declaration-human-rights/.

United Nations (1966): International Covenant on Economic, Social and Cultural Rights. Adopted and opened for signature, ratification and accession by General Assembly resolution 2200A (XXI) of 16 December 1966. Entry into force 3 January 1976, in accordance with article 27; http://www.institut-fuer-menschenrechte.de/fileadmin/user_upload/PDF-Dateien/Pakte_Konventionen/ICESCR/icescr_en.pdf.

United Nations (1980): Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons which may be deemed to be Excessively Injurious or to have Indiscriminate Effects (with Protocols I, II and III), Geneva, 10 October 1980; http://treaties.un.org/Pages/ViewDetIAls.aspx?src=TREATY&mtdsg_no=XXVI-2&chapter=26&lang=en.

United Nations (1992): Agenda 21. Konferenz der Vereinten Nationen für Umwelt und Entwicklung, Rio de Janeiro; http://www.un.org/Depts/german/conf/agenda21/agenda_21.pdf.

United Nations (2015a): Paris Agreement, Paris; https://unfccc.int/files/meetings/paris_nov_2015/application/pdf/paris_agreement_english_.pdf.

United Nations (2015b): Transforming our world: the 2030 Agenda for Sustainable Development. Resolution adopted by the General Assembly on 25 September 2015. A/RES/70/1; http://www.un.org/ga/search/view_doc.asp?symbol=A/RES/70/1&Lang=E.

Université de Montréal (2018): The Declaration: https://www.montrealdeclaration-responsibleIAcom/the-declaration.

Verbraucherzentrale Bundesverband (2017): Algorithmenbasierte Entscheidungsprozesse - Thesenpapier des vzbv, Berlin.

Vowinkel, Bernd (2017): Ist der Mensch eine Maschine; https://transhumanismus.wordpress.com/2017/06/14/ist-der-mensch-eine-maschine/.

Walter, Sven (2014): Situierte Kognition. In: Information Philosophie; Heft 2/2014, S. 28-32,. Lörrach.

Weizsäcker, Ernst Ulrich von/ Wijkman, Anders (2017): Wir sind dran. Club of Rome: Der große Bericht. Was wir ändern müssen, wenn wir bleiben wollen. Eine neue Aufklärung für eine volle Welt, Gütersloh.

WEF (World Economic Forum in collaboration with The Boston Consulting Group), (2018): Towards a Reskilling Revolution: A Future of Jobs for All, Köln/Genf.

6. Apéndice

6.1 Principios Asilomar sobre la Inteligencia Artificial

Preguntas de investigación

1) Meta de la investigación: el objetivo de la investigación de la IA no debería ser crear inteligencia sin dirigir, sino inteligencia beneficiosa.

2) Financiación de la investigación: la inversión en IA debería ir acompañada de fondos para investigar en asegurar su uso beneficioso, incluyendo cuestiones espinosas sobre ciencias de la computación, economía, legislación, ética y estudios sociales.

3) Enlace entre ciencia y política: debería haber un intercambio constructivo y sano entre los investigadores de IA y los legisladores.

4) Cultura de la investigación: una cultura de cooperación, confianza y transparencia debería ser fomentada entre los investigadores y desarrolladores de IA.

5) Evitar las carreras: los equipos que estén desarrollando sistemas de IA deberían cooperar activamente para evitar chapuzas en los estándares de seguridad.

Ética y Valores

6) Seguridad: los sistemas de IA deberían ser seguros a lo largo de su vida operativa, y verificables donde sea aplicable y posible.

7) Transparencia en los fallos: si un sistema de IA causa daño debería ser posible determinar por qué.

8) Transparencia judicial: cualquier intervención de un sistema autónomo en una decisión debería ir acompañada de una explicación satisfactoria y auditable por parte de una autoridad humana competente.

9) Responsabilidad: los diseñadores y desarrolladores de sistemas avanzados de IA son depositarios de las implicaciones morales de su uso, mal uso y acciones, con la responsabilidad y oportunidad de dar forma a dichas implicaciones.

10) Alineación de valores: los sistemas de IA altamente autónomos deberían ser diseñados para que sus metas y comportamientos puedan alinearse con los valores humanos a lo largo de sus operaciones.

11) Valores humanos: los sistemas de IA deberían ser diseñados y operados para que sean compatibles con los ideales de dignidad humana, derechos, libertades y diversidad cultural.

12) Privacidad personal: la gente debería tener el derecho de acceder, gestionar y controlar los datos que generan, dando a los sistemas de IA el poder de analizar y utilizar esa información.

13) Libertad y privacidad: la aplicación de la IA a los datos personales no puede restringir de forma poco razonable la libertad, real o sentida, de las personas.

14) Beneficio compartido: las tecnologías de IA deberían beneficiar y fortalecer a tanta gente como sea posible.

15) Prosperidad compartida: la prosperidad económica creada por la IA debería ser compartida ampliamente, para el beneficio de toda la Humanidad.

16) Control humano: los seres humanos deberían escoger cómo y si delegan decisiones a los sistemas de IA para completar objetivos escogidos previamente.

17) Sin subversión: el poder conferido por el control de sistemas de IA altamente avanzados debería respetar y mejorar, más que subvertir, los procesos sociales y cívicos de los que depende la salud de la sociedad.

18) Carrera armamentística: debería ser evitada cualquier carrera armamentística de armas autónomas letales.

Problemas a largo plazo

19) Capacidad de precaución: al no haber consenso, deberíamos evitar las asunciones sobre los límites superiores de las futuras capacidades de la IA.

20) Importancia: la IA avanzada podría representar un profundo cambio en la historia de la vida en la Tierra, y debería ser planificada y gestionada con el cuidado y los recursos adecuados.

21) Riesgos: los riesgos asociados a los sistemas de IA, especialmente los catastróficos o existenciales, deben estar sujetos a planificación y esfuerzos de mitigación equiparables a su impacto esperado.

22) Automejora recursiva: los sistemas de IA diseñados para automejorarse recursivamente o autorreplicarse de una forma que pudiera llevar al rápido incremento en su calidad o cantidad deben estar sujetos a unas estrictas medidas de control y seguridad.

23) Bien común: la superinteligencia debería ser desarrollada sólo en servicio de unos ideales éticos ampliamente compartidos y para beneficio de toda la Humanidad, más que para un Estado u organización.

6.2 Autores (miembros del grupo de estudio VDW)

Prof. Dr. Ulrich Bartosch (politólogo) es profesor de pedagogía en la Universidad Católica de Eichstätt-Ingolstadt, donde enseña e investiga la teoría pedagógica y la historia de las ideas políticas, la reforma y la educación universitaria, la descripción y el desarrollo de las competencias, la inclusión, la participación, el trabajo social en la escuela y la política interna mundial. De 2005 a 2011 fue presidente de la asociación alemana de facultades de trabajo social (Fachbereichstag Soziale Arbeit). Es miembro del grupo de trabajo especial de "validación y el reconocimiento de los formatos digitales de aprendizaje" del foro de la digitalización de la conferencia de rectores universitarios (HRK) y responsable del proyecto de cooperación de la Universidad Católica Eichstätt-Ingolstadt (KU) y de la VDW "Laudato Si - Die päpstliche Enzyklika im Diskurs der Großen Transformation". De 2009 a 2015 fue Presidente de la Asociación de Científicos Alemanes y desde entonces ha sido Presidente del Consejo Asesor de la VDW.

Prof. Dr. Stefan Bauberger SJ (físico, filósofo) es profesor de Filosofía Natural y Filosofía de la Ciencia en la Universidad de Filosofía de Munich. Es miembro de la Orden de los Jesuitas y teólogo. Tiene un doctorado en física teórica y una habilitación en filosofía. Ha trabajado durante varios años en física teórica de partículas. Investiga y enseña sobre temas fronterizos entre la filosofía y las ciencias naturales, especialmente la física, en el campo del diálogo entre la ciencia natural y la religión, así como sobre la filosofía del budismo, y en los campos de la filosofía de la tecnología y la teoría de la ciencia Anteriormente fue director de formación de la Orden de los Jesuitas en Alemania. Es un Maestro ZEN y dirige un centro de meditación.

Tile von Damm (politólogo) es director del instituto de investigación urbana MOD e investigador asociado de la Universidad Técnica de Berlín. Es experto en urbanismo en el proyecto de la UE "Orfeo&Majnun". Además, es cofundador y director general de DiMed para asegurar la atención médica básica en las zonas rurales. Su investigación y trabajo se centra en el desarrollo rural y urbano inclusivo, la participación y la gobernanza mundial, el código abierto y los datos abiertos, y el desarrollo y la transferencia de investigación. Es miembro de la Red Europea de Cultura e Industrias Creativas. Fue director de investigación del Centro de Investigaciones Literarias y Culturales (ZfL), director del instituto de investigación PerGlobal y coordinador de la iniciativa de excellencia de la Universidad Humboldt de Berlín. En la Cumbre Mundial de las Naciones Unidas sobre el Desarrollo Sostenible y en las negociaciones de las Naciones Unidas sobre la sociedad de la información, formó parte de una delegación de negociadores de la sociedad civil.

Dr. Rainer Engels (agrónomo) es experto en política económica en la cooperación para el desarrollo y se ocupa desde hace muchos años de cuestiones económicas relacionadas con el desarrollo. Se centra en el desarrollo del comercio, la inversión y la política industrial, en particular los derechos de propiedad intelectual y las normas técnicas. Desde 2015 trabaja en la automatización y digitalización de la producción industrial (industria 4.0) y la electromovilidad. Es consultor de política económica sostenible y desarrollo del sector privado en la agencia alemana de cooperación internacional (GIZ). Anteriormente, fue Director General de Germanwatch durante muchos años.

Prof. Dr. Malte Rehbein (historiador) es titular de la Cátedra de Humanidades Digitales de la Universidad de Passau, donde investiga y enseña métodos formales e informáticos, incluidos los métodos basados en IA y sus aplicaciones a las tareas y cuestiones de humanidades y estudios culturales, con especial atención a las ciencias históricas. Publica sobre estudios de datos históricos, digitalización de bienes culturales y modelización de datos, así como sobre cuestiones de ética y crítica de la ciencia y la sociedad. Su experiencia profesional incluye la industria de la informática y la consultoría; las estaciones académicas fueron Göttingen, Würzburg, Galway/Irlanda, Victoria/Canadá y Lincoln-NE/USA. Es miembro de la Comisión Histórica de la Academia Bávara de Ciencias. Es miembro de la Asociación de Científicos Alemanes desde 2017.

Frank Schmiedchen (economista y MBA) es Director del Gobierno del Ministerio Federal de Cooperación Económica y del Desarrollo de Alemania (BMZ) siendo el Economista del ministerio. Anteriormente, fue responsable de biodiversidad y bioseguridad, política exterior y de seguridad en África, grupo de países ACP (África, Caribe y Pacífico), política industrial, ONUDI, derechos de propiedad intelectual y el desarrollo de la producción farmacéutica local en el BMZ y la Representación Permanente de la República Federal de Alemania ante la UE. Anteriormente fue titular de la Cátedra de Economia International y Decano del Departamento de Gestión de PYMES de la Pontificia Universidad Católica del Ecuador y coordinó los departamentos pertinentes de la Asociación de Universidades Jesuitas de América Latina (AUSJAL). 2002-2009 y desde 2016 es miembro del consejo asesor de la Asociación de Científicos Alemanes. Desde octubre de 2017 dirige el Grupo de Estudio de Evaluación de la Tecnología de Digitalización del VDW.

Prof. Dr. Heinz Stapf-Finé (sociólogo, economista) es profesor de Política Social en la Universidad Alice Salomon de Berlín y director académico de la Paritätische Akademie Berlin. Doctor en "La previsión para la vejez en España", es un experto internacional en el ámbito de la política laboral y social. Antes de su nombramiento como profesor, fue Jefe de Política Social en el Comité Ejecutivo Federal de la Federación Sindical Alemana (DGB). Obtuvo su primera experiencia profesional como Director de Operaciones del Luxembourg Income Study y como asistente de investigación en el Institute for Health and Social Research (IGES) Berlin. Luego trabajó como oficial de políticas para la Sociedad Hospitalaria Alemana. Es miembro de la Asociación de Científicos Alemanes desde 2006.

Angelika Sülzen (economista de empresa) es Directora del Gobierno del Ministerio Federal de Cooperación Económica y del Desarrollo de Alemania (BMZ), donde actualmente es responsable de las áreas de igualdad de género, conciliación y gestión de la salud. Anteriormente, fue la Oficial de País para África Central y fue responsable de la cooperación bilateral de la República Federal de Alemania con Burundi y la República Centroafricana. Anteriormente, fue responsable de asuntos presupuestarios y financieros y dirigió un gran proyecto de tecnología de información en el BMZ. De 2003 a 2007 trabajó para el Servicio Alemán de Desarrollo en Sudáfrica y Lesotho.

Les agradecemos sus contribuciones y sugerencias:

Lucas Bartosch, Judith Buttenmüller, Prof. Dr. Hartmut Graßl, Prof. Dr. Regine Kollek, Dr. Hans-Jochen Luhmann, Dr. Michael Marhöfer y Christine von Weizsäcker.