Hacia una propuesta de metodología para el desarrollo de proyectos de IA confiables

Carlos Mario Braga
Grupo de investigación Alarcos
Universidad de Castilla-La Mancha
Ciudad Real, España
CarlosMario.Braga1@alu.uclm.es

Manuel A. Serrano
Grupo de investigación Alarcos
Universidad de Castilla-La Mancha
Ciudad Real, España
Manuel.Serrano@uclm.es

Eduardo Fernández-Medina
Grupo de investigación GSyA
Universidad de Castilla-La Mancha
Ciudad Real, España
Eduardo.FdezMedina@uclm.es

***Resumen*—En los últimos meses hemos vivido la irrupción y rápida adopción de la Inteligencia Artificial Generativa (IAG), lo que ha provocado un claro aumento de riesgos en el uso de este tipo de sistemas. En paralelo, han aparecido los primeros acuerdos de organismos legisladores para regular el uso de la Inteligencia Artificial (IA). Estos hechos han evidenciado la necesidad de un refuerzo en la búsqueda de soluciones que ayuden a implementar esta familia de sistemas garantizando que sean confiables. Pero la confiabilidad implica conceptos variados que van desde la ciberseguridad hasta la ética. Por ello, en este artículo, comenzamos un camino hacia la construcción de una propuesta metodológica para el desarrollo de sistemas de IA confiables, arrancando con el análisis del propio concepto de confiabilidad (trustworthy en inglés) en este contexto, y construyendo una taxonomía de atributos que conforman este concepto, cuya definición será clave para el trabajo posterior.**

***Index Terms*—Inteligencia Artificial, IA, Desarrollo de proyectos de IA, Confiabilidad, Ciberseguridad, Ética, Fiabilidad, Inteligencia Artificial Generativa, IAG**

Tipo de contribución: *Investigación original*

I. Introducción

Resultaría extenso documentar de forma exhaustiva todos los estudios relevantes que confirman qué sistemas de IA causaron perjuicios o discriminación a grupos de personas, ya sea en el ámbito de la contratación laboral [1], en sistemas de reconocimiento facial [2], o en sistemas de reconocimiento de voz [3], por citar algunos de ellos. Una lectura del libro "Armas de Destrucción Matemática" de Cathy O'neill [4], ofrece una posibilidad de entender de forma amena la magnitud del problema. No obstante esta pequeña muestra de casos expuestos se caracteriza por estar formada por sistemas de IA que fueron diseñados de forma bienintencionada. Por tanto, el problema se agrava si incluimos en la ecuación el ciberriesgo y permitimos que tomen control sobre los sistemas de IA individuos con intenciones espurias. Es por tanto notoria la importancia y necesidad de disponer de una metodología para construir sistemas de IA confiables [5], y el hecho de que la ciberseguridad ha de tomar una papel relevante en dicha metodología [6], [7].

La aparición de problemas como los anteriormente mencionados y su impacto en derechos fundamentales; como el que supone por ejemplo la discriminación frente al derecho de igualdad; animó a distintas organizaciones nacionales e internacionales de los ámbitos publico, privado y gubernamental a crear grupos de trabajo compuestos por profesionales expertos, que generaron decenas de documentos definiendo los principios éticos en que debería basarse la IA. Hoy, en marzo de 2024, es posible encontrar 167 guías éticas en la web (https://inventory.algorithmwatch.org/), que surge de un proyecto de la institución Algorithm Watch, con objetivo de completar un inventario global de guías que tratan de establecer los principios éticos que deben regir los sistemas de toma de decisión automatizada. Este trabajo ya partía de una base tras la irrupción del Big Data y la consecuente acumulación y manipulación de grandes conjuntos de datos, cuando se había evidenciado la necesidad del desarrollo de una "Ética del Big Data" acorde a cuatro principios de alto nivel: privacidad, confidencialidad, transparencia e identidad (entendida como la capacidad de los individuos para definir quiénes son) [8].

Dado que en el desarrollo de la IA están involucrados muchos profesionales con distinta idiosincrasia cultural y de distintos ámbitos o dominios profesionales, cabe suponer la existencia de cierta dificultad para acordar unos principios éticos generales y aceptables para todos [9]. Pese a ello, han sido varios los intentos exitosos de analizar las principales guías éticas publicadas, extraer conceptos, ordenarlos, y proponer un marco ético global para la IA [10], [11], [12].

Estos trabajos de análisis de las guías éticas, y propuestas de marcos éticos generales mediante convergencia de principios de las guías específicas, impulsaron la calificación de la IA con adjetivos que trataban de describir su vinculación a estos principios éticos, aflorando conceptos como, por ejemplo, IA ética (Ethical AI) [13], [14], IA confiable (Trustworthy AI) [12], IA buena (Good AI) [10], o IA robusta y beneficiosa (Robust and beneficial AI) [13].

Estas distintas calificaciones y definiciones de IA hacen notar falta de consenso, que se puede confirmar con la revisión de uno sólo de estos conceptos, concretamente el de Trustworthy AI; pues es definida como "esencial para alcanzar todo el potencial de la IA y que ha de basarse en cinco principios fundamentales: Beneficencia, No-Maleficencia, Autonomía, Equidad y Explicabilidad" por Thieves et al [12]; como "válida y confiable, segura, resiliente, responsable y trasparente, explicable e interpretable, con gestión justa de sesgos nocivos y con acento en la privacidad" (NIST); o como "Equitativa, Robusta, con Privacidad, Explicable y Transparente" (IBM); o como "legal, ética y robusta" (European Commission's Ethics Guidelines for Trustworthy AI). Esta evidencia de falta de consenso justifica la necesidad de establecer una taxonomía integradora que defina los conceptos subyacentes de la confiabilidad en la IA.

Es necesario sin embargo tener en cuenta que pese a la bonanza del trabajo realizado con génesis en la ética, surgen

diversos focos de crítica constructiva, tanto desde el mundo legal; desde el que se señala que el enfoque acometido propicia que en ocasiones se mezclen y confundan ética y ley a pesar de ser áreas de conocimiento diferentes y complementarias [15], e incluso que esta mezcla se propicie conscientemente para dar una sensación de cumplimiento que haga parecer no necesaria la regulación (ethics washing) [16]; como desde el mundo tecnológico; en el se destaca la falta de aplicabilidad de los principios éticos y la dificultad para trasladarlos al ciclo de vida de los productos de IA [17], [18], [19]; como también desde el propio mundo de la filosofía, desde el que se detectaban omisiones en los principios como el impacto ecológico del entrenamiento de los algoritmos, el potencial control gubernamental, el potencial abuso político, la desinformación, o la propaganda o información tóxica, entre algunas otras carencias evidenciadas [20].

I-A. Planteamieto del trabajo

Una vez introducido el contexto, incidiremos en el resto del trabajo en crear la taxonomía integradora respecto al concepto de confiabilidad en la IA, que considere los principios establecidos en los principales trabajos elaborados hasta el momento desde los ámbitos ético, legal y tecnológico; prestando especial atención a evaluar con un sencillo ejercicio de análisis que las principales dimensiones de la taxonomía apoyan y promueven la ciberseguridad.

Para ello, organizaremos el resto del artículo del siguiente modo. En la sección II, se revisan y analizan las principales contribuciones que tratan de centrar el concepto de confiabilidad en la IA; en la sección III mostramos la taxonomía de conceptos en el contexto de la confiabilidad, que hemos creado en base a las contribuciones existentes, tratando de dar respuesta a las cuestiones planteadas previamente; la sección IV situa la aportación de este artículo en el contexto de una investigación que se inicia con el objetivo de crear sistemas IA confiables; y finalmente, la sección V muestra las principales conclusiones.

II. Trabajo relacionado

II-A. La propuesta ética

Uno de los primeros marcos globales para una IA buena para la sociedad, fue el propuesto en 2018 por Floridi et al. [10], en el que tras analizar 6 guías éticas que contenían un total de 47 principios, se concluyó que era posible clasificar esos 47 principios en 5 dimensiones principales, 4 de ellas dimensiones clásicas en el contexto de la bioética: Beneficencia (Beneficence), No Maleficencia (Non-Maleficence), Autonomía (Autonomy) y Equidad (Justice); y una nueva dimensión adicional específica de la IA: la Explicabilidad (Explicability). Adicionalmente estableció una definición para cada uno de estos principios, que tendremos oportunidad de revisar en este trabajo. Haremos alusión en adelante a estas 5 dimensiones con el acrónimo B-NM-A-E-E.

Un año después, Jobin et al [11], analizaron un total de 84 guías sobre ética en AI, principalmente generadas por compañías privadas o agencias gubernamentales de Europa, Estados Unidos y Japón; aportando entre otras cosas como resultado de la investigación una escala de las dimensiones principales de acuerdo al número de guías éticas en las que cada una de ellas aparecía; una relación entre conceptos que en distintas guías eran invocados con un nombre similar pero no coincidente; y la conclusión de que había una unanimidad global en las dimensiones principales: Transparencia (Transparency), Equidad (Justice and Fairness), No Maleficencia (Non-Maleficence), Responsabilidad (Responsability), Privacidad (Privacy), Beneficencia (Beneficence) y Libertad y Autonomía (Freedom and Autonomy); que aparecian en 73, 68, 60, 60, 47, 41 y 34 de las 84 guías respectivamente, mientras que otras dimensiones importantes como por ejemplo la Sostenibilidad (Sustainability) únicamente aparecía en 14 de las 84 guías.

Aunque pueda parecer en un primer momento que existen diferencias entre ambas investigaciones respecto a estas dimensiones principales, no es así en realidad, ya que la definición de la dimensión Explicabilidad en el primero contempla tanto el sentido ético (Transparencia o Auditibilidad) como el epistemológico (Explicabilidad o Modelos Interpretables), por lo que en la dimensión Explicabilidad de Floridi et al. tienen cabida las dimensiones Responsabilidad y Transparencia de Jobin et al. Del mismo modo, en la definición de la dimensión No Maleficencia por parte de Floridi et al. tiene cabida la dimensión Privacidad del trabajo de Jobin et al. La consolidación de estos dos primeros trabajos fue realizada por Thieves et al [12] en 2021, concluyendo que las 5 dimensiones presentadas en Floridi et al. (B-NM-A-E-E) son suficientes para clasificar los principios éticos contenidos en las principales guías éticas.

Otro estudio interesante afrontado desde la perspectiva de la ética o la filosofía, es el presentado por B.Green en 2018 [21] que, a diferencia de los tres a los que hemos hecho referencia hasta el momento, no se basa en un análisis de las guías de ética para la IA sino que reflexiona sobre 12 retos éticos que de acuerdo a su criterio hay que afrontar respecto a la IA y su uso. Aunque no forma parte de la conclusión de la investigación resulta sencillo clasificar esos 12 retos en las 5 dimensiones principales en las que convergen los trabajos anteriores(B-NM-A-E-E). Es decir, que en este trabajo el principio de Beneficencia se presenta como "Función", o "Buen Uso" de la IA; el de No-Maleficencia se presenta a través de los términos "Mal uso" o hablando de la importancia de evitar causar "Efectos negativos psicosociológicos" o "Efectos negativos espirituales" a las personas, o "el impacto que puede tener en el desempleo"; sitúa el foco en la "Automatización de decisiones morales", en los "Derechos de los sistemas de IA" y en la "Dependencia que pueden causar" en lo referente al principio de Autonomía; menciona la "Desigualdad" y los "Sesgos incorporados en los algoritmos" que claramente tiene relación con la dimensión Equidad; y finalmente, aborda la dimensión Explicabilidad cuando reflexiona sobre la necesidad de "Asegurar la transparencia en la toma de decisiones".

Continuando con estudios emprendidos con un enfoque orientado desde el ámbito de las humanidades, cabe destacar el trabajo de Samuele Lo Piano, que el año 2020 [14], partiendo de los trabajos anteriores de Floridi et al. y Jovin et al., y asumiendo las 5 dimensiones principales (B-NM-A-E-E), expone un problema que será necesario abordar a la hora de llevar los aspectos éticos a la práctica. Este problema es la existencia de puntos de fricción entre algunas de las

dimensiones, como por ejemplo, el que puede existir entre la transparencia y la privacidad. Justamente, esta fricción fue expuesta con detalle por De Laat en 2018 [22], argumentando que la total transparencia puede ocasionar filtrado de datos sensibles o privados a la luz pública, puede lesionar los derechos de propiedad privada de las compañías y tener por tanto consecuencias negativas para su competitividad y/o para la reputación de sus trabajadores; añadiendo que además que pese a fomentar la transparencia puede resultar complejo incluso para expertos interpretar algunos tipos de algoritmos. Son por tanto estas fricciones un punto que requerirá de una especial atención en ejercicios que pretendan llevar estos principios éticos a la práctica.

Se puede concluir por tanto que hasta el año 2022 las dimensiones B-NM-A-E-E, están asumidas en la literatura científica como los principales principios éticos que tienen la capacidad de clasificar el resto. Sin embargo, la situación de los sistemas de IA dio un vuelco significativo a finales del año 2022 y durante el año 2023 con la proliferación de sistemas de IAG, inspirados en una evolución de la arquitectura de red neuronal propuesta en 2017 [23], haciendo totalmente necesario continuar nuestra exposición evaluando si las dimensiones B-NM-A-E-E son suficientes o no para clasificar los retos que plantean los sistemas de IAG nacidos de la implementación de estas nuevas arquitecturas.

En consecuencia y para concluir con este grupo de trabajos orientados desde las humanidades, vamos a destacar la investigación realizada por Thilo Hagendorff [24], en la que ejecuta una revisión sistemática de la literatura científica relativa a los retos éticos que plantean los sistemas de IAG. Para realizar este estudio seleccionó 179 documentos científicos de entre los 1.120 recuperados en su selección inicial, e identificó 378 problemas en la IAG que clasificó en 19 áreas temáticas de la ética en la IA. Algunos de ellos son claramente trasladables a las 5 dimensiones éticas B-NM-A-E-E, aunque otras requieren una revisión para afrontar su clasificación. De este modo, los retos éticos "Sostenibilidad","Alienación con objetivos" y "Valores humanos (utilidad)" encajan con el principio de Beneficencia; mientras que el de No Maleficencia se manifiesta como "Riesgos de Manipulación", "Contenidos dañinos o tóxicos" o "Protección contra riesgos fortuitos (Safety)"; y "Problemas de privacidad", "Protección de derechos de Copyright" claramente relacionados con la protección de datos. En cuanto al principio de Autonomía, aparece ahora como una preocupación por la "Interacción hombre-máquina"; mientras que habla de "Justicia" y "Sesgos" como retos que pueden surgir en estos sistemas relativos al principio de Equidad; y finalmente de dificultades de "Evaluación y Auditoría" y "Trasparencia y Explicabilidad" en lo que refiere al principio de Explicabilidad.

Sin embargo, aparecen un grupo de retos que requieren sistemas "robustos y con seguridad técnica" como la lucha contra el "Cibercrimen" y la proliferación de "Alucinaciones"; o marcos de trabajo que presenten capacidades de "gobernanza" para mitigar los nuevos riesgos asociados a la IAG. También aparece un reto que directamente alude a la necesidad de "Regulación" para disponer de un marco legal que ayude a proteger a las personas de los riesgos de la IA. También otra serie de retos que abren nuevos debates como son las consecuencias negativas que en la sociedad pueden causar los sistemas de IAG en torno a la "escritura y la investigación", el "arte y la creatividad" o la "educación", por pérdida de capital cultural o pérdida de singularidades culturales.

Como veremos a continuación, los nuevos retos éticos que demandan regulación o sistemas robustos tanto desde un punto de vista técnico como procedimental, no partían únicamente del mundo de las humanidades tras la explosión de la IAG, sino que, ya habían sido apuntados previamente desde los mundos legal y tecnológico en diversas investigaciones en las que se aportaba crítica constructiva para ampliar las dimensiones éticas B-NM-A-E-E con otros principios o dimensiones que permitiesen fortalecer las capacidades de generación de sistemas de IA buenos para la sociedad.

II-B. La crítica constructiva

Tal y como hemos adelantado, las críticas constructivas a la suficiencia de una visión de una IA beneficiosa con una perspectiva exclusivamente humanista partieron de los ámbitos legal y tecnológico, por lo que se hace necesario recurrir a investigaciones de ambos dominios que nos ayuden a complementar lo hasta ahora expuesto.

De la perspectiva legal, M. Robles Carrillo en el año 2020 [15] alertó sobre la necesidad de un enfoque holístico que integrase ética, derecho y tecnología para abordar los desafíos que plantea la IA, pese a reconocer que no es fácil la combinación de las tres disciplinas. También profundiza en los aspectos éticos y pone de manifiesto que aunque puedan cumplir un propósito similar al que cumplen los aspectos legales, adolecen tanto de falta de obligatoriedad de aplicabilidad de las normas, como de falta de los mecanismos necesarios para garantizar su cumplimiento; concluyendo que los aspectos legales deben ser incluidos dentro de los principios fundamentales a seguir para construir una IA buena para la sociedad.

También desde el ámbito legal, en su trabajo del año 2019, E. Magrani [25], reconoce lo importante que es para la comunidad investigadora y académica promover un amplio debate sobre los principios éticos que han de guiar la construcción de IA; pues son esos principios éticos los que posteriormente deberían guiar la construcción del marco regulatorio; por lo que al igual que el trabajo anteriormente comentado, señala que siendo importantes y necesarias las guías éticas, no son suficientes de cara a conseguir el objetivo planteado.

Para finalizar con la las aportaciones provenientes del mundo legal, otra visión de la necesidad de regulación, parte de la reflexión de que la consideración de los principios éticos de forma aislada puede suponer un incentivo para que algunas empresas traten de evitar una regulación más estricta, y apuesten por proyectar autorregulación, desarrollando políticas internas con débil implementación real [26], [27], [20]. Este peligro no es exagerado a la vista de los datos pues en 2022 cuando en el inventario de Algorithms Watch había 173 guías éticas, 115 eran recomendaciones cuyo cumplimiento no era obligatorio, y respecto a las 41 que provenían del sector privado 15 eran recomendaciones y 22 voluntarias, quedando por tanto únicamente cuatro de ellas caracterizadas como acuerdos vinculantes con medios para sancionar el incumplimiento [28].

Girando la vista al mundo tecnológico, cabe destacar el trabajo de Pavaloiu et al. en 2017 [13], en el que reflexiona, entre otros aspectos, acerca de la necesidad de tener entornos de IA amigables para las personas o profesionales; al igual que sobre la necesidad de sistemas robustos, argumentando para ello que la IA puede fallar de maneras inesperadas en tareas ordinarias debido a los diferentes patrones de pensamiento.

También es destacable desde esta perspectiva el trabajo de Zhou [29], en el que apunta la necesidad de integrar los principios éticos a lo largo de todo el ciclo de vida de la IA; desde el diseño y el desarrollo, hasta la implantación y monitorización; a la vez que propone, de manera similar a como se hace en biomedicina, establecer cuestionarios a modo de lista de verificación que puedan servir como punto de referencia y ayuda para la traslación de los principios éticos a los sistemas de IA.

Ahondando en la misma línea, el trabajo de Christoforaki et al. de 2022 [28], subraya la importancia de disponer de marcos éticos centrados en principios como el respeto a la autonomía humana, prevención de daños, justicia o explicabilidad; complementados con otros principios como la agencia y la supervisión humanas, la solidez o robustez técnica, la seguridad, la privacidad, el gobierno, la transparencia, la diversidad, la no discriminación, la equidad, el bienestar ambiental y social y la rendición de cuentas. Es claro que entre todos estos principios la robustez técnica, la seguridad y el gobierno son claramente principios del ámbito técnico, mientras que la rendición de cuentas requiere de un marco regulatorio con capacidad sancionadora, y pertenece en consecuencia al ámbito legal.

Aunque hemos centrado la exposición de la crítica constructiva a las guías éticas desde la perspectiva legal y tecnológica, es justo mencionar que desde el propio ámbito de la ética se apuntaba la dificultad de encontrar formas de poner en práctica los principios éticos de la IA al existir incertidumbre sobre como estos principios deben implementarse [11]; o por la complejidad, variabilidad, subjetividad y falta de estandarización, incluyendo la interpretación de cada uno de los principios éticos [18]; así como la necesidad de establecer un marco ético para la IA que oriente su desarrollo y despliegue [30].

III. Taxonomía de confiabilidad en IA

Como hemos visto hasta el momento, parece necesario para construir una IA confiable combinar las dimensiones éticas relevantes con un marco regulatorio que permita acompañarlas de obligación de aplicabilidad; de los mecanismos necesarios para garantizar su cumplimiento; y con aspectos o herramientas técnicas que garanticen la robustez, el gobierno, y la traslación a la práctica de estos principios a todo el ciclo de vida del desarrollo de la IA.

Por tanto y partiendo de todo el análisis anterior procedemos a presentar una propuesta, en la que partiendo de la visión de convergencia de principios éticos (B-NM-A-E-E), y las aportaciones legal y tecnológica, generemos una taxonomía integradora (Figura 1) que establezca los conceptos fundacionales subyacentes sobre la que posteriormente sea posible basar una metodología confiable de desarrollo de soluciones de IA.

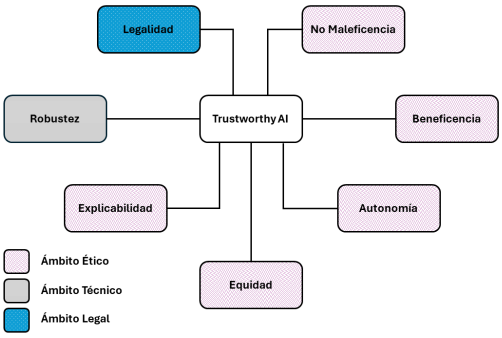

Figura 1. Taxonomía para una AI Confiable

Adicionalmente, seleccionaremos para cada una de las dimensiones una definición, bien sea aportándola de la literatura consultada por poseer toda la carga semántica de la que se quiere dotar a esa dimensión, o bien con una definición propia que posea toda esa carga semántica en caso contrario; y presentaremos una relación entre cada dimensión y los principios equivalentes que aparecen en la literatura revisada identificando el estudio en el que aparece el término que asignamos a la taxonomía y que justifica la asignación, con el objetivo de definir un segundo nivel en la taxonomía que contenga principios convalidables por el principio fundamental del que dependen. Marcaremos además en **negrita** los términos que forman parte del contexto de la ciberseguridad, para ilustrar la variedad de conceptos de ciberseguridad que hay en el contexto de la IA Confiable.

III-A. Legalidad

La legalidad es el principio que refiere al respeto a todas las leyes y regulaciones aplicables.

En la en la Tabla I presentamos un segundo nivel para esta dimensión; presentando en la primera columna otras formas de nombrar a la dimensión Legalidad en la literatura, y en la segunda la referencia que justifica la inclusión de ese término como segundo nivel de la dimensión presentada en la tabla.

Tabla I
Clasificación de segundo nivel: Legalidad.

Concepto	Referencia
Normas y Regulación/Rules and Regulation	[13][24]
Legalmente Responsable/Liability	[13][25]
Acceso a la Justicia/Access to Justice	[31]
Principios fundamentales/Core Values	[31]
Imparcialidad/Balancing Legal Values	[31]

III-B. Beneficencia

La beneficencia es el principio ético que se refiere al desarrollo, implementación y uso de IA de forma beneficiosa para la humanidad y el planeta, en el sentido de que promueva el bienestar de los seres humanos y el medio ambiente, y respete los derechos humanos fundamentales [10].

Presentamos en la Tabla II un segundo nivel para esta dimensión.

Tabla II
CLASIFICACIÓN DE SEGUNDO NIVEL: BENEFICENCIA.

Concepto	Referencia
Armonía-He/Harmony-He	[12]
Benevolencia/Benevolence	[12]
Utilidad/Helpfulness	[12][24]
Beneficio/Benefict	[11]
Bienestar/Well-being	[11][12]
Paz/Peace	[11]
Bien social/Social Good	[11]
Bien común/Common Good	[11]
Buena Función/Good Function	[32][21]
Buen Uso/Good Use	[21]
Sostenibilidad/Sustainability	[21]

III-C. No Maleficencia

El principio ético de No Maleficencia aboga por el desarrollo, implementación y uso de la IA de manera que evite dañar a las personas o las proteja de daños [10].

Se presenta un segundo nivel para esta dimensión en la Tabla III.

Tabla III
CLASIFICACIÓN DE SEGUNDO NIVEL: NO MALEFICENCIA

Concepto	Referencia
Prot. contra riesgos intencionados/Security	[12][11]
Prot. contra riesgos fortuitos/Safety	[12][11][13][24]
Privacidad/Privacy	[13][29][11][12][24]
Prot. contra daños/Harm Protection	[11][24]
Prot. contra Información Tóxica/Toxic Info	[24]
Precaución/Precaution	[11]
Prevención/Preventión	[11]
No subversión/Non-Subversion	[11]
Uso no maligno/Non-evil Use	[21]
Uso no nocivo/Non-harmful use	[26]
Integridad corporal o mental/Integrity	[26][24]
Prot. de datos/Data protection	[26][24][11]
Prot. de la vulnerabilidad/Vulnerability	[32]

III-D. Autonomía

La Autonomía es el principio ético que establece que los humanos conservan el derecho de decisión sobre quién debe decidir. En ocasiones es referenciado como meta-autonomía [10].

Una vez más ofrecemos el segundo nivel de la dimensión en la Tabla IV.

Tabla IV
CLASIFICACIÓN DE SEGUNDO NIVEL: AUTONOMÍA.

Concepto	Referencia
Promoción de la voluntad/Promotion of agency	[12]
Promoción de la supervisión/Promotion of overssight	[12]
Control de la IA/Controlability of AI	[12]
Interacción hombre-máquina/HMI	[13][24]
Libertad y autonomía/Fredom and authonomy	[11]
Consentimiento/Consent	[11]
Elección/Choice	[11]
Autodeterminación/Self-Determination	[11]
Libertad/Liberty	[11]
Empoderamiento/Empowerment	[11]
Dignidad/Dignity	[11]
Evitar la dependencia/Avoid Dependency	[21]

III-E. Equidad

El principio ético de la Equidad es el que impulsa la utilización de la IA para corregir desigualdades pasadas como la discriminación; la justa distribución de los beneficios generados a través de IA; y que impide la afloración de nuevos daños e inequidades en los sistemas de IA [10].

Como con el resto de dimensiones y con el mismo criterio se puede ver en la Tabla V el segundo nivel de la dimensión.

Tabla V
CLASIFICACIÓN DE SEGUNDO NIVEL: EQUIDAD.

Concepto	Referencia
Justicia/Fairness	[12][26][31][11][29][24]
Justicia/Equity	[12]
Beneficio compartido/Shared benefit	[12]
Prosperidad compartida/Shared prosperity	[12]
Sin Sesgos/Unbiased	[31][32][21][11][24]
Igualdad socioeconómica/Status equality	[21][11]
Distribución/Distribution	[11]
Inclusión/Inclusion	[11]
No discriminación/Non-Discrimination	[11]
Diversidad/Diversity	[11]
Pluralidad/Plurality	[11]
Accesibilidad/Accessibility	[11]
Compensación/Redress	[11]
Acceso/Access	[11]
Solidaridad/Solidarity	[11]
Consistencia/Consistency	[11]
Reversibilidad/Reversibility	[11]
Reparación/Remedy	[11]
Objección/Challenge	[11]

III-F. Explicabilidad

La Explicabilidad es el principio ético que defiende la generación de una IA explicable tanto desde un punto de vista epistemológico fomentando la creación de IA explicable mediante la producción de modelos de IA (más) interpretables manteniendo altos niveles de rendimiento y precisión, como desde un punto de vista ético estimulando la creación de IA responsable (accountable) [10].

El detalle del segundo nivel para esta dimensión puede consultarse en la Tabla VI.

Tabla VI
CLASIFICACIÓN DE SEGUNDO NIVEL: EXPLICABILIDAD.

Concepto	Referencia
Transparencia/Transparency	[12][26][31][29][29][21][11][24]
Inteligibilidad/Intelligibility	[12]
Responsabilidad/Accountability	[12][11][24]
Interpretabilidad/Interpretability	[12][32][11]
Confiabilidad/Reliability	[12]
Responsabilidad/Responsibility	[12][29][11]
Explicabilidad/Explainability	[11][29][32]
Comprensibilidad/Understandabilty	[11]
Comunicable/Communicable	[11]
Divulgable/Disclosure	[11]
Honestidad/Integrity	[11]
Evaluación/Evaluation	[24]
Auditoría/Auditing	[24]

III-G. Robustez

La Robustez refiere al principio que aboga por la fortaleza y resiliencia de los sistemas de IA, tanto desde una perspectiva técnica y procedimental, como desde el punto de vista de conciencia social. En cuanto al punto de vista

procedimental promueve herramientas como por ejemplo un marco de gobierno, cuestionarios o listas de verificación a modo de ayuda a la toma de decisiones éticas, o un ciclo de desarrollo de software (SDLC) con las dimensiones incluidas en esta taxonomía integradas. Desde el punto de vista técnico promueve la contemplación de la seguridad desde el diseño como principio fundamental. Al considerar esta dimensión la promoción de herramientas de ayuda a la integración del resto de dimensiones, supone el principio motor en el éxito de la posterior aplicación de la taxonomía a la construcción de una metodología confiable de desarrollo de sistemas de IA.

En la Tabla VII se presenta el segundo nivel para esta dimensión.

Tabla VII
CLASIFICACIÓN DE SEGUNDO NIVEL: ROBUSTEZ.

Concepto	Referencia
Cuestionarios/Verification Lists	[33][29]
Modelo de Gobierno/Governance Model	[15][24]
Robusted técnica/Technical robustness	[24]
SDLC Confiable/Trustworthy SDLC	[17][18][19]
Seguridad Técnica/Tech. Security	[28]

Es importante destacar que hay intersección entre el principio de Legalidad y los 5 principios éticos, pero hemos visto que los principios éticos convergen de un amplio número de guías éticas provenientes de diversas culturas y localizaciones geográficas por lo que se pueden considerar “universales”; mientras que el marco legal es distinto en distintos países o grupos de países, pudiendo ser más riguroso o más laxo dependiendo del regulador de cada zona geográfica. Esto ha sido relevante a la hora de asociar distintos conceptos a la taxonomía propuesta ya que se ha hecho prevalecer la asociación a los principios éticos respecto a la asociación al principio de legalidad para hacer el ejercicio lo mas universal posible. Usamos por tanto el principio de legalidad para referir a marcos regulatorios, rendición de cuentas y aspectos que la visión ética de forma aislada no puede aportar.

III-H. La confiabilidad en la IA

Una vez expuestos y definidos los principios que deben constituir un sistema de IA confiable, podemos identificar los sistemas o metodologías de IA confiables como aquellos construidos y operados bajo las dimensiones de Legalidad, Beneficencia, No-Maleficencia, Autonomía, Equidad, Explicabilidad y Robustez, considerando para cada dimensión la definición anteriormente expuesta.

Dado que el trabajo realizado parte de una visión ética que ya había sido trabajada en la literatura, y la parte legal vendrá de la mano de las autoridades reguladoras en cada área geográfica, es necesario poner el foco en los siguientes pasos en la dimensión de Robustez; para asegurar que trasladamos la necesidad de gobierno [15], [24], [28]; entornos amigables de desarrollo IA para los profesionales [13]; herramientas para conseguir que este dimensión sea el principio motor y que logre integrar el resto de principios en el ciclo de vida de IA, y fomentar los cuestionarios a modo de listas de verificación [29]; y la seguridad [28].

III-I. Conexión con la ciberseguridad

Finalmente, y a modo de prueba de la bondad de la taxonomía propuesta presentamos un breve ejercicio para verificar si adicionalmente a la dimensión de Robustez que promueve la seguridad y resiliencia técnicas, el resto de dimensiones la complementan a la hora de incorporar ciberseguridad a los sistemas construidos con metodologías basadas en esta taxonomía. Para ello, vamos a revisar las 5 riesgos de seguridad más importantes dentro del Top-10 de una de las categorías relativas a los sistemas de IA que publica OWASP (Open Web Application Security Project), ver en qué consisten, cómo se pueden mitigar, qué daños producen si se produce un ataque e identificar si alguna de las dimensiones de la taxonomía propuesta ayuda a mitigar la presencia de estas vulnerabilidades.

El primer potencial ataque es el ML01:2023 Input Manipulation Attack; y se trata de un tipo de ataque en el que un atacante altera deliberadamente los datos de entrada para engañar al modelo. Este ataque se puede prevenir bien utilizando modelos robustos diseñados contra este tipo de ataques, bien implantando validaciones de los datos de entrada al modelo que rechacen los datos de entrada con alta probabilidad de ser maliciosos, o bien entrenando el modelo con este tipo de ejemplos para hacerlo más robusto y reducir las opciones de que el modelo sea engañado. Este tipo de ataque puede provocar el robo de datos, comprometer el sistema u otras formas de daño.

Tener en mente en todo el ciclo de vida la dimensión No Maleficencia; relacionada como hemos visto con la protección de datos, la privacidad, y el no causar perjuicio a las personas (Tabla III); ayuda a establecer controles para prevenir este ataque. Adicionalmente el principio de Robustez, y concretamente la robustez técnica, fomenta el uso de modelos robustos, validación de entrada, y a no limitar las pruebas a casos de uso en los que la entrada sea la esperada; por lo que podemos concluir que no es la Robustez la única dimensión que aporta en la construcción de un modelo resiliente desde el punto de vista de la ciberseguridad.

El segundo de los potenciales ataques en el top-10 es el conocido como ML02:2023 Data Poisoning Attack; que ocurre cuando un atacante manipula los datos de entrenamiento para hacer que el modelo se comporte de manera no deseada. Hay varias formas de mitigación que pasan por la validación de los datos de entrenamiento, el almacenamiento seguro de los datos de entrenamiento y el control a su acceso, hacer predicciones con un modelo construido a partir de varios modelos, o implementar un control de anomalías en los datos de entrenamiento. Este ataque causa que el modelo haga predicciones que pueden ser sesgadas y perjudicar por tanto a las personas.

Tener en mente en todo el ciclo de vida el principio de No maleficencia y el no causar perjuicio a las personas (Tabla III) y el de Equidad y el evitar sesgos (Tabla V de forma adicional al de Robustez ayuda tanto a tomar las medidas mitigadoras, como a tener una monitorización del modelo con el que podamos detectar predicciones sesgadas consecuencia de estos ataques , por lo que podemos concluir que las dimensiones de la taxonomía propuesta colaboran en la implementación de la ciberseguridad.

El tercer ataque a considerar es el ML03:2023 Model Inversion Attack; que ocurre cuando un atacante aplica ingeniería inversa al modelo para extraer información de él. Puede prevenirse limitando el acceso al modelo y sus predicciones, validando las entradas al modelo, haciendo el modelo transparente (capturando logs de inputs y outputs para generar explicaciones del modelo), o monitorizando el comportamiento del modelo y detectando alteraciones en las métricas. Es evidente en este caso que se podría extraer del modelo información personal fiable a partir de las predicciones del modelo por lo que los principios de No Maleficencia y Equidad, adicionalmente al de Robustez, ayudan, igualmente que en el tipo de ataque anterior, a poner las bases para la mitigación de este riesgo o la monitorización para detectarlo, por lo que podemos concluir de nuevo que las dimensiones alternativas a la Robustez aportan positivamente a la ciberseguridad.

Prosiguiendo con el cuarto ataque llamado ML04:2023 Membership Inference Attack; que ocurre cuando un atacante manipula datos de entrenamiento del modelo para hacer que se comporte de una manera que exponga información confidencial y puede prevenirse de forma sencilla incluyendo datos aleatorios entre los datos de entrenamiento de modo que no sea posible identificar para un atacante que filas de datos forman parte de los datos de entrenamiento, utilizando técnicas para proteger datos sensibles como la privacidad diferencial, usando técnicas de regularización o actuando sobre el dataset limitando datos redundantes o altamente correlacionados. De nuevo el riesgo es perder información confidencial y por tanto privacidad (Tabla III), por lo que adicional a la dimensión de Robustez, revisar durante todo el ciclo de vida la dimensión de No Maleficencia impulsa la toma de medidas mitigadoras adecuadas para esta vulnerabilidad.

Para concluir este ejercicio, veamos el quinto ataque, ML05:2023 Model Theft, que ocurre cuando el atacante obtiene acceso a los parámetros del modelo desensamblando el código binario, o accediendo a los datos de entrenamiento y al algoritmo del modelo. Este ataque puede causar el robo del modelo y de los datos de entrenamiento y puede mitigarse encriptando el modelo y los datos de entrenamiento, implementando control de acceso a estos activos con autenticación multi-factor, teniendo un back-up para poder restaurarlo tras el robo, aplicándole alguna protección legal como patente, o monitorizando los accesos para poder detectar cuando un atacante está intentando entrar la modelo. Consecuencia de este ataque de nuevo perdemos datos e incluso el modelo completo, por lo que una vez más adicional a la dimensión de Robustez, la de No Maleficencia (Tabla III) ayuda a aplicar medidas mitigadoras para este tipo de ataques, llegando por tanto a la misma conclusión que en los casos anteriores.

Se concluye por tanto de forma general que si bien una metodología confiable para la generación de sistemas de IA incluye pero no está limitada a la ciberseguridad, el resto de principios propuestos suman junto con la Robustez técnica a la implantación consciente de la ciberseguridad.

IV. Desarrollo de proyectos de IA confiables

Este artículo, y la taxonomía propuesta, tienen su justificación en el contexto de ejecución de una investigación que se inicia con el objetivo de crear una metodología que facilite y simplifique la generación de sistemas de IA confiables, mediante la traslación de los siete principios de la taxonomía a lo largo del ciclo de vida del proyecto tal y como se muestra en la Figura 2.

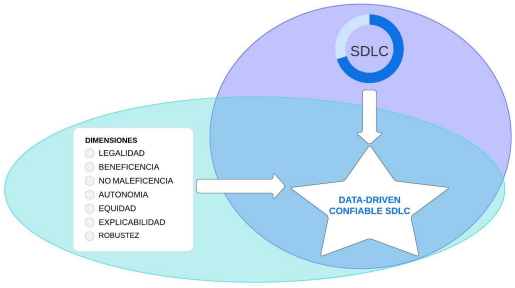

Figura 2. Data-Driven Trustworthy SDLC

Respecto al SDLC al que trasladar los principios fundamentales de la taxonomía, un punto de partida podría ser CRISP-DM [34], pues es de uso común en un amplio número de dominios [35], y flexible y de fácil adopción [36]; y aunque tiene algunos problemas como que no favorece la comunicación y el trabajo en equipo [36], [35], o que fue creado hace más de 20 años en un contexto analítico dominado por el Data Mining y en un contexto tecnológico diferente al actual [34]; ha experimentado a lo largo de los años intentos exitosos de extensión para dotarlo de más capacidades relativas a mejorar el trabajo en equipo y el trabajo iterativo (TSDP - Microsoft 2018), flexibilidad para proyectos de data science (FMDS - IBM-2015), reutilización de modelos y adaptabilidad al contexto [37], incluir aspectos de MLOPS (ASUM-DM - IBM 2015), incluir aspectos de ciberseguridad y adaptación para proyectos real-time [38], o incluir capacidades de gestión de modelos y gobierno [39]. Por supuesto, modelos donde ya se integra la seguridad como parte del ciclo de vida (S-SDLC) también son un punto de partida relevante en la búsqueda del modelo de proceso objetivo.

Finalmente hemos revisado un trabajo [18], que propone una tipología combinando los “principios éticos” con las etapas del ciclo de vida de la IA; concluyendo que la disponibilidad de herramientas para llevar los principios de “IA Ética” a la práctica no está distribuida uniformemente a lo largo del ciclo de vida, existiendo más dificultad para garantizar la alineación con estos principios en la etapa de implantación que en las de diseño o pruebas del modelo; y que hay ausencia de herramientas para trasladar el principio de autonomía. Pese a que este trabajo está limitado a los principios éticos y no al conjunto de los 7 principios o dimensiones que hemos establecido en este artículo, y pese a que el ciclo de vida, al ser este trabajo del año 2020, no está obviamente adaptado a las necesidades de los nuevos sistemas de IAG; este trabajo supone un buen acompañante a la taxonomía y a la idea inicial de extensión de CRISP-DM, para iniciar los primeros pasos en la búsqueda de la metodología que se propone.

V. Conclusiones

Las soluciones y capacidades de las que podemos dotarnos con los sistemas de IA no están exentas de riesgos que pueden perjudicar a personas o grupos sociales incluso en

situaciones en que se diseñan con buenas intenciones. El cibercrimen, con el robo o toma de control de los modelos por ciberdelincuentes, amplía la probabilidad de ocurrencia de estos perjuicios si no se integra la ciberseguridad en estos sistemas.

Este artículo revisa la convergencia de las distintas guías éticas a unos principios éticos fundamentales y los completa con principios que se aportan desde el mundo legal y tecnológico, creando una primera taxonomía que constituya la base de la creación de una metodología confiable de desarrollo de sistemas de IA que no se limita únicamente a dimensiones éticas. Durante la exposición hemos visto que los siete principios fundacionales de esta taxonomía cubren las principales preocupaciones surgidas tras la reciente irrupción de la IAG; y que son principios fuertes que ayudarán, una vez incluidos a lo largo de las fases del ciclo de vida del desarrollo de software, a reforzar las políticas de ciberseguridad y mitigar por tanto las vulnerabilidades que abren las puertas a los principales ataques que se pueden perpetrar sobre estos sistemas.

Agradecimientos

Este trabajo ha sido desarrollado con el apoyo de los siguientes proyectos: Di4SPDS (PCI2023145980-2), financiado por MCIN/AEI/10.13039/501100011033 y por la Unión Europea (programa Chist-Era), AETHER-UCLM (PID2020-112540RB-C42) financiado por MCIN/AEI/10.13039/501100011033, ALBA-UCLM (TED2021-130355B-C31, id.4809130355-130355-28-521), y MESIAS (2022-GRIN-34202) financiado por FEDER.

Referencias

[1] A. Köchling and M. C. Wehner, "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of hr recruitment and hr development," *Business Research*, vol. 13, no. 3, pp. 795–848, 2020.

[2] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[3] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[4] C. O'neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

[5] I. Martinez, E. Viles, and I. G. Olaizola, "Data science methodologies: Current challenges and future approaches," *Big Data Research*, vol. 24, p. 100183, 2021.

[6] J. S. Saltz and I. Shamshurin, "Big data team process methodologies: A literature review and the identification of key factors for a project's success," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 2872–2879.

[7] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of big data challenges and analytical methods," *Journal of business research*, vol. 70, pp. 263–286, 2017.

[8] N. M. Richards and J. H. King, "Big data ethics," *Wake Forest L. Rev.*, vol. 49, p. 393, 2014.

[9] S. Leonelli, "Locating ethics in data science: responsibility and accountability in global and distributed knowledge production systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2083, p. 20160122, 2016.

[10] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi *et al.*, "Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations," *Minds and machines*, vol. 28, pp. 689–707, 2018.

[11] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[12] S. Thiebes, S. Lins, and A. Sunyaev, "Trustworthy artificial intelligence," *Electronic Markets*, vol. 31, pp. 447–464, 2021.

[13] A. Pavaloiu and U. Kose, "Ethical artificial intelligence-an open question," *arXiv preprint arXiv:1706.03021*, 2017.

[14] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–7, 2020.

[15] M. R. Carrillo, "Artificial intelligence: From ethics to law," *Telecommunications Policy*, vol. 44, no. 6, p. 101937, 2020.

[16] E. Bietti, "From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 210–219.

[17] J. Whittlestone, R. Nyrup, A. Alexandrova, and S. Cave, "The role and limits of principles in ai ethics: Towards a focus on tensions," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 2019, pp. 195–200.

[18] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, "From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices," *Science and engineering ethics*, vol. 26, no. 4, pp. 2141–2168, 2020.

[19] C. Stix, "Actionable principles for artificial intelligence policy: three pathways," *Science and Engineering Ethics*, vol. 27, no. 1, p. 15, 2021.

[20] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and machines*, vol. 30, no. 1, pp. 99–120, 2020.

[21] B. P. Green, "Ethical reflections on artificial intelligence," *Scientia et Fides*, 2018.

[22] P. B. De Laat, "Algorithmic decision-making based on machine learning from big data: can transparency restore accountability?" *Philosophy & technology*, vol. 31, no. 4, pp. 525–541, 2018.

[23] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[24] T. Hagendorff, "Mapping the ethics of generative ai: A comprehensive scoping review," *arXiv preprint arXiv:2402.08323*, 2024.

[25] E. Magrani, "New perspectives on ethics and the laws of artificial intelligence," *Internet policy review*, vol. 8, no. 3, 2019.

[26] S. Larsson, "On the governance of artificial intelligence through ethics guidelines," *Asian Journal of Law and Society*, vol. 7, no. 3, pp. 437–451, 2020.

[27] M. Coeckelbergh, "Artificial intelligence: some ethical issues and regulatory challenges," *Technology and regulation*, vol. 2019, pp. 31–34, 2019.

[28] M. Christoforaki and O. Beyan, "Ai ethics—a bird's eye view," *Applied Sciences*, vol. 12, no. 9, p. 4130, 2022.

[29] J. Zhou, F. Chen, A. Berry, M. Reed, S. Zhang, and S. Savage, "A survey on ethical principles of ai and implementations," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 3010–3017.

[30] D. Dawson, E. Schleiger, J. Horton, J. McLaughlin, C. Robinson, G. Quezada, J. Scowcroft, and S. Hajkowicz, "Artificial intelligence: Australia's ethics framework," *Data61 CSIRO, Australia*, 2019.

[31] H. Surden, "The ethics of artificial intelligence in law: Basic questions," *Forthcoming chapter in Oxford Handbook of Ethics of AI*, pp. 19–29, 2020.

[32] S. M. Liao, *Ethics of artificial intelligence*. Oxford University Press, 2020.

[33] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building ethics into artificial intelligence," *arXiv preprint arXiv:1812.02953*, 2018.

[34] C. Shearer, "The crisp-dm model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[35] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying crisp-dm process model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021.

[36] J. S. Saltz, "Crisp-dm for data science: strengths, weaknesses and potential next steps," in *2021 IEEE International Conference on Big Data (Big Data)*. IEEE, 2021, pp. 2337–2344.

[37] F. Martínez-Plumed, L. Contreras-Ochando, C. Ferri, J. Hernández-Orallo, M. Kull, N. Lachiche, M. J. Ramírez-Quintana, and P. Flach, "Crisp-dm twenty years later: From data mining processes to data science trajectories," *IEEE transactions on knowledge and data engineering*, vol. 33, no. 8, pp. 3048–3061, 2019.

[38] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 1, pp. 97–107, 2013.

[39] L. Cao, "Domain-driven data mining: Challenges and prospects," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 6, pp. 755–769, 2010.