Magíster en Pensamiento Contemporáneo. Filosofía y pensamiento político; Doctorado en Filosofía

Problemas políticos de la inteligencia artificial

Periodo Académico	: Segundo semestre de 2022
Créditos	: 6
Requisito	: –
Horario	: Lunes, miércoles y jueves, 18:30-21:30 horas
Fechas	: Del 1 al 25 de agosto
Horario Atención	: Jueves, 15:00-17:00 horas
Profesor	: Wolfhart Totschnig <wolfhart.totschnig@mail.udp.cl>

I. DESCRIPCIÓN

¿Qué pasará cuando las máquinas se vuelvan tan inteligentes como nosotros, como los seres humanos? ¿Y cómo debemos prepararnos hoy para esta eventualidad?

Hasta hace relativamente poco, se consideraba que estas cuestiones pertenecían al ámbito de la ficción: entretenidas de imaginar quizá, pero no dignas de ser consideradas seriamente. Sin embargo, desde el ensayo seminal de Vernor Vinge "The coming technological Singularity" (1993), un número creciente de filósofos, informáticos y otros pensadores han comenzado a plantearlas y debatirlas.

Este curso pretende indagar en el debate sobre las posibles consecuencias del advenimiento de una inteligencia artificial (IA) equivalente a la humana. En particular, pretende poner de manifiesto y analizar los problemas *políticos* que este acontecimiento podría provocar. Así, la orientación del curso se opone a la actitud predominante en el debate, que consiste en ver la situación no como un problema político, sino como uno *tecnológico*, un problema relativo a cómo mantener la situación *bajo control* (Yudkowsky, Bostrom, Omohundro, Domingos).

Concretamente, algunos de los problemas políticos que discutiremos son:

1) los conflictos entre los seres humanos sobre el acceso a los beneficios de las nuevas tecnologías (en general) y la inteligencia artificial (en particular);
2) los conflictos entre los seres humanos sobre cómo reaccionar ante una inteligencia artificial equivalente a la humana, si es que llega a existir tal IA;
3) el problema de cómo establecer una forma de coexistencia pacífica con una IA equivalente a la humana.

II. OBJETIVOS

Los objetivos principales del curso son

- enfrentar a los/las estudiantes a una temática filosófica importante y de actualidad, a saber, los problemas políticos de la inteligencia artificial;
- darles a conocer textos clásicos y contemporáneos que son relevantes para abordar esta temática;

- introducirlos a la metodología de la filosofía mediante la lectura y discusión de estos textos;
- específicamente, enseñarles a analizar un texto filosófico con respecto a los argumentos que sostiene y luego a evaluar estos argumentos;
- guiarlos en la escritura de un ensayo pertinente y claro sobre un aspecto específico de la temática;
- y, en general, abrir así un espacio para la reflexión conjunta sobre la temática.

III. METODOLOGÍA DE TRABAJO

El curso comprende 11 sesiones que se extienden desde el 1 hasta el 25 de agosto. Las sesiones comenzarán con una presentación introductoria por parte del profesor, en la que se expondrá el contexto histórico y filosófico del texto a tratar. Esta presentación dará lugar a una discusión común sobre las ideas y argumentos centrales del texto, dirigida por el profesor mediante preguntas conductoras. Además, la discusión será impulsada por breves exposiciones (10 minutos como máximo) de parte de los/las estudiantes sobre un aspecto del texto que les parezca especialmente interesante y discutible.

Los estudiantes practicarán la escritura académica a lo largo del curso: por un lado, a través de tres respuestas cortas (una por semana) a una cuestión planteada por el profesor y, por otro, mediante un ensayo final más extendido sobre un tema propuesto por el estudiante (ver abajo sección "Evaluación").

En clase regirá esta regla: los ordenadores portátiles y otros dispositivos electrónicos deben utilizarse solo para fines relacionados con el curso.

El profesor estará disponible dos horas a la semana para consultas individuales, los días jueves desde las 15:00 hasta las 17:00 horas.

IV. CONTENIDOS

En este curso, revisaremos y discutiremos las principales posturas con respecto al posible advenimiento de una inteligencia artificial equivalente a la humana:

1) la idea de una "Singularidad tecnológica" por venir (Vinge);

2) la fusión entre biología y tecnología como apoteosis del ser humano (Kurzweil);

3) el cíborg como transgresión de las dicotomías tradicionales (Haraway);

4) la inteligencia artificial como riesgo existencial para toda la humanidad (Bostrom).

V. EVALUACIÓN

Se requerirán de los estudiantes los trabajos siguientes:

- Una breve exposición (10 minutos como máximo) sobre algún texto del curso. El propósito de la exposición es que el estudiante plantee un comentario, duda u objeción acerca de un aspecto específico del texto y así impulse la discusión. La exposición tendrá un valor de 20% de la nota final.
- Tres respuestas escritas cortas (una página o 400 palabras como máximo), una por semana, a una cuestión planteada por el profesor sobre la lectura de las sesiones precedentes. Estas respuestas tendrán un valor de 30% (cada una 10%) de la nota final.

- Un ensayo final de 8 páginas o 3200 palabras como máximo, sobre algún tema relacionado con el curso. Antes de escribir el ensayo, los estudiantes deberán presentar al profesor –durante las horas de consulta o por correo electrónico– el tema elegido para que el profesor lo apruebe. El ensayo final tendrá un valor de 40% de la nota final.
- El 10% restante de la nota final corresponde a la participación del estudiante en las discusiones en clase. Se considerará tanto la participación pasiva (asistencia) como la participación activa (intervenciones).

Los criterios de evaluación para los trabajos escritos (respuestas cortas/reflexiones libres y ensayo final) serán los siguientes, en orden de prioridad:

1) Relevancia: Lo escrito debe ser relevante para el tema planteado, y debe ser evidente cómo es relevante.
2) Claridad: Lo escrito debe poder entenderse por alguien que no está familiarizado con el tema y los textos en cuestión.
3) Precisión: La explicación de los conceptos y argumentos en cuestión debe ser fiel a los textos y hechos a que se refiere.

Además, el ensayo final (pero no las respuestas cortas) debe ser más que un mero resumen del/de los texto(s) en cuestión. Es decir, debe plantear alguna duda, objeción, extensión, aplicación, comparación, etcétera, con respecto al/a los texto(s) que discute.

Para estudiantes de doctorado, las tres respuestas cortas se sustituyen por tres reflexiones libres sobre algún texto de las sesiones precedentes (una página o 400 palabras como máximo). Las demás exigencias son las mismas que para estudiantes de magíster.

Los estudiantes de doctorado pueden presentar los trabajos escritos en inglés si desean practicar este idioma.

VI. **CRONOGRAMA**

1. Lunes 1 de agosto	Tema: Conferencia (clase magistral) en el marco de la Cátedra de Filosofía Jorge Eugenio Dotti Lectura: –
2. Miércoles 3 de agosto	Tema: Introducción (breve historia de la inteligencia artificial) y organización del curso Lectura: Pedro Domingos, *The master algorithm*, capítulo 1
3. Jueves 4 de agosto	Tema: Turing: el "juego de la imitación" como prueba de inteligencia artificial Lectura: Alan Turing, "Maquinaria computadora e inteligencia"

4. Lunes 8 de agosto	Tema: Vinge: la idea de una "singularidad tecnológica" por venir Lectura: Vernor Vinge, "The coming technological Singularity: How to survive in the post-human era" **Entrega de la primera respuesta corta**
5. Miércoles 10 de agosto	Tema: Kurzweil: la inteligencia artificial es una extensión de nosotros mismos Lectura: Ray Kurzweil, *La Singularidad está cerca*, capítulo 1
6. Jueves 11 de agosto	Tema: Kurzweil: la idea de que, gracias a la IA, nos transformaremos en dioses Lectura: Ray Kurzweil, *La Singularidad está cerca*, capítulo 6
Lunes 15 de agosto	**Feriado**
7. Miércoles 17 de agosto	Tema: Haraway: el cíborg como transgresión de las dicotomías tradicionales Lectura: Donna Haraway, "Manifiesto para *cyborgs*" **Entrega de la segunda respuesta corta**
8. Jueves 18 de agosto	Tema: Bostrom: la inteligencia artificial representa un riesgo existencial para la humanidad entera Lectura: Nick Bostrom, *Superinteligencia*, prefacio, capítulos 7 y 8
9. Lunes 22 de agosto	Tema: Bostrom: ¿cómo solucionar el problema de controlar la inteligencia artificial? Lectura: Nick Bostrom, *Superinteligencia*, capítulos 9 y 15
10. Miércoles 24 de agosto	Tema: Domingos: nuestro futuro con la inteligencia artificial Lectura: Pedro Domingos, *The master algorithm*, capítulo 10 **Entrega de la tercera respuesta corta**
11. Jueves 25 de agosto	Tema: repaso y conclusión Lectura: –
Domingo 11 de sept.	**Entrega del ensayo final**

VII. NORMAS ADMINISTRATIVAS Y PEDAGÓGICAS DEL CURSO

1. ASISTENCIA:

El porcentaje mínimo de asistencia para este curso es de **82%** (9 sesiones).

2. SOBRE LA NO ENTREGA DE EVALUACIONES:

En caso de la no entrega a tiempo de los trabajos finales, se aceptará como única justificación un certificado médico o la acreditación de una razón de fuerza mayor. En ambos casos los antecedentes serán recibidos por el/la coordinador/a académico/a, solo hasta una semana después de la fecha del plazo de entrega, y será responsabilidad del estudiante hacerla llegar dentro del plazo indicado.

Existirá una nota "P" (Pendiente), calificación que se aplicará al estudiante que, por motivos justificados o por razones de fuerza mayor debidamente acreditadas, no haya podido cumplir con las evaluaciones finales del curso o actividad en que se ha inscrito. Dicha calificación permitirá al estudiante inscribirse en cursos para los cuales constituye requisito aquel cuya calificación hubiere quedado pendiente. La nota "P" (Pendiente) deberá ser autorizada por el comité académico o, en su defecto, por el director del programa, debiendo el académico responsable del curso o actividad fijar al estudiante las exigencias que deberá cumplir para obtener la calificación definitiva. Si el estudiante no diere cumplimiento a lo señalado anteriormente, en el plazo que se fije que no podrá ser superior a un semestre, será calificado en el respectivo ramo con nota final uno (1,0).

3. INTEGRIDAD ACADÉMICA:

El Reglamento del Estudiante de la UDP establece severas sanciones para casos de plagio, copia, falsificación y uso indebido de documentos, que van desde la nota mínima en la evaluación hasta la expulsión de la Universidad.

VIII. BIBLIOGRAFÍA

1. LITERATURA OBLIGATORIA

Bostrom, Nick. *Superinteligencia: caminos, peligros, estrategias*. Traducido por Marcos Alonso. Zaragoza: Teell, 2016.

> Original en inglés: Bostrom, Nick. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press, 2014.

Domingos, Pedro. *The master algorithm: How the quest for the ultimate learning machine will remake our world*. New York: Basic Books, 2015.

Haraway, Donna J. "Manifiesto para *cyborgs*: ciencia, tecnología y feminismo socialista a finales del siglo XX". En *Ciencia,* cyborgs *y mujeres: la reinvención de la naturaleza*, 251-311. Madrid: Ediciones Cátedra, 1995.

> Original en inglés: Haraway, Donna J. "A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century". En *Simians, cyborgs, and women: The reinvention of nature*, 149-181. New York: Routledge, 1991.

Kurzweil, Ray. *La Singularidad está cerca: cuando los humanos transcendamos la biología*. Traducido por Carlos García Hernández. Berlin: Lola Books, 2012.

> Original en inglés: Kurzweil, Ray. *The Singularity is near: When humans transcend biology*. New York: Viking, 2005.

Turing, Alan. "Maquinaria computadora e inteligencia". En *Controversia sobre mentes y máquinas*, editado por Alan Ross Anderson. Barcelona: Tusquets, 1984.

Original en inglés: Turing, Alan. "Computing machinery and intelligence". *Mind* 59:236 (oct. 1950), 433-460.

Vinge, Vernor. "The coming technological Singularity: How to survive in the post-human era". 1993. https://www-rohan.sdsu.edu/faculty/vinge/misc/singularity.html

2. LITERATURA COMPLEMENTARIA

Butler, Samuel. *Erewhon, o Al otro lado de las montañas*. Traducido por Andrés Cotarelo Jiménez. Madrid: Akal, 2012. Publicado originalmente en 1872.

Original en inglés: Butler, Samuel. *Erewhon*. New York: Dover Publications, 2002.

Chalmers, David J. "The singularity: A philosophical analysis". *Journal of Consciousness Studies* 17:9-10 (2010), 7-65.

Dreyfus, Hubert L. *What computers still can't do: A critique of artificial reason*. Cambridge, Massachusetts: The MIT Press, 1992.

Geraci, Robert M. "Apocalyptic AI: Religion and the promise of artificial intelligence". *Journal of the American Academy of Religion* 76:1 (2008), 138-166.

Good, Irving John. "Speculations concerning the first ultraintelligent machine". En *Advances in computers*, vol. 6, editado por Franz L. Alt y Morris Rubinoff, 31-88. New York: Academic Press, 1965.

Moravec, Hans. *Robot: Mere machine to transcendent mind*. Oxford: Oxford University Press, 1999.

Omohundro, Stephen M. "The nature of self-improving artificial intelligence." 2008. https://selfawaresystems.files.wordpress.com/2008/01/nature_of_self_improving_ai.pdf.

Penrose, Roger. *La mente nueva del emperador*. Traducido por José Javier García Sanz. México: Fondo de Cultura Económica, 1996.

Original en inglés: Penrose, Roger. *The emperor's new mind*. Oxford: Oxford University Press, 1989.

Searle, John. "Mentes, cerebros y programas". En *Filosofía de la inteligencia artificial*, editado por Margaret Boden, 82-104. México: Fondo de Cultura Económica, 1994.

Original en inglés: Searle, John. "Minds, brains, and programs". *The Behavioral and Brain Sciences* 3:3 (sept. 1980), 417-424.

Totschnig, Wolfhart. "The problem of superintelligence: Political, not technological". *AI & Society* 34:4 (2019), 907-920.

Turing, Alan. "Intelligent machinery, a heretical theory". *Philosophia Mathematica* 4:3 (1996), 256-260.

Vinge, Vernor. *A fire upon the deep*. New York: Tor Books, 1992.

Wiener, Norbert. "Some moral and technical consequences of automation". *Science* 131:3410 (mayo 1960), 1355-1358.

Yudkowsky, Eliezer. *Creating friendly AI 1.0: The analysis and design of benevolent goal architectures*. San Francisco: The Singularity Institute, 2001.

3. PELÍCULAS

2001: A Space Odyssey, dir. Stanley Kubrick, Reino Unido 1968.

Colossus: The Forbin Project, dir. Joseph Sargent, EE. UU. 1970.

Demon Seed, dir. Donald Cammell, EE. UU. 1977.

Blade Runner, dir. Ridley Scott, EE. UU. 1982.

The Terminator, dir. James Cameron, EE. UU. 1984.

Terminator 2: Judgment Day, dir. James Cameron, EE. UU. 1991.

Ghost in the Shell, dir. Mamoru Oshii, Japón 1995.

The Matrix, dir. Andy Wachowski & Lana Wachowski, EE. UU./Australia 1999.

A.I. Artificial Intelligence, dir. Steven Spielberg, EE. UU. 2001.

I, Robot, dir. Alex Proyas, EE. UU. 2004.

Her, dir. Spike Jonze, EE. UU. 2013.

Transcendence, dir. Wally Pfister, EE. UU. 2014.

Ex Machina, dir. Alex Garland, Reino Unido 2015.

San Junipero (episodio de *Black Mirror*), dir. Owen Harris, Reino Unido 2016.

AlphaGo, dir. Greg Kohs, EE. UU. 2017.