



**UNIVERSIDAD NACIONAL AUTÓNOMA DE  
MÉXICO  
FACULTAD DE FILOSOFÍA Y LETRAS**



**LICENCIATURA EN FILOSOFÍA**

**ASIGNATURA:**

PROBLEMAS DE TEORÍA DEL CONOCIMIENTO—FILOSOFÍA DE LA CIENCIA

**Filosofía de la Inteligencia Artificial**

Imparte: Maestro/Candidato a Doctor. Rodolfo Carlos Prieto Mendoza

Correo: [carlosprietomendoza@gmail.com](mailto:carlosprietomendoza@gmail.com)/[carlosprietofilos.unam.mx](mailto:carlosprietofilos.unam.mx)

**PRESENTACIÓN**

La inteligencia artificial (IA) es una tecnología abrumadoramente transformadora que no solo está remodelando nuestras herramientas y servicios, sino también algunas de nuestras nociones filosóficas más fundamentales. Este curso ofrece un análisis introductorio de las intersecciones entre la filosofía y la IA. En particular, exploraremos cuatro dimensiones filosóficas que, a la luz del advenimiento de la IA, merecen nuevas reflexiones: (I) Consciencia, (II) Realidad, (III) Ética y (IV) Coexistencia Humano-Máquina. Este curso no sólo busca ofrecer una comprensión más profunda de la IA, sino también fomentar un diálogo crítico sobre el papel de esta tecnología en una posible reconceptualización de nuestro rol en el mundo, como también un nuevo sentido de ser humano.

**OBJETIVOS GENERALES**

Al finalizar este curso los estudiantes deberán ser capaces de:

- Articular las preguntas filosóficas centrales sobre la inteligencia artificial, incluyendo cuestiones de consciencia, realidad, ética y coexistencia humano-máquina.
- Evaluar diferentes perspectivas filosóficas sobre la naturaleza de la IA y sus capacidades.
- Diferenciar entre problemas fidedignos y no fidedignos con respecto a la IA.
- Construir argumentos razonados sobre el estatus filosófico de la IA.

## FORMA DE EVALUACIÓN

- 80% trabajo final
- 20% participación en clase

## CONTENIDO

NÚM. DE HRS. POR UNIDAD	TEMARIO
8	<p><b>0 Introducción</b></p> <p><b>0.1 Historia y conceptos de la IA 1:</b> GOFAI, redes neuronales artificiales y machine learning.</p> <p><b>0.2 Historia y conceptos de la IA 2:</b> Deep learning y modelos transformer.</p> <p><b>0.3 Mitos y realidades de la IA:</b> Errores comunes y miedos no fundamentados.</p> <p><b>0.4 La intersección de la filosofía y la IA:</b> Preguntas y temas filosóficos clave de la IA para el siglo XXI.</p>
6	<p><b>1. Consciencia</b></p> <p><b>1.1 IA y experiencia subjetiva:</b> <i>Qualia</i> e intencionalidad en la IA.</p> <p><b>1.2 ¿Pueden pensar las máquinas?</b> El test de Turing y otros tests.</p> <p><b>1.3 ¿Pueden tener consciencia las máquinas?</b> Naturaleza de la consciencia y su relación con las mentes y las máquinas.</p>

<b>8</b>	<p><b>2 Realidad</b></p> <p><b>2.1 Diferencias y Similitudes entre ontologías computacionales y filosóficas:</b> Ontologías computacionales como Siri y Alexa vs. ontologías filosóficas tradicionales.</p> <p><b>2.2 IA y la re-ontologización del mundo:</b> El desvanecimiento de la división natural/virtual.</p> <p><b>2.3 Teorías de simulación y sentido de la vida:</b> Sentido y significación a la luz de un potencial nihilismo digital.</p> <p><b>2.4 Riesgos existenciales:</b> Problemas de control y la amenaza de una IA con objetivos mal alineados.</p>
<b>6</b>	<p><b>3 Ética</b></p> <p><b>3.1 La posibilidad de agencia de la IA:</b> Consideraciones morales básicas de la IA.</p> <p><b>3.2 Ética de la colaboración entre humanos y máquinas en la creatividad:</b> Entre creatividad, innovación y plagio.</p> <p><b>3.3 Desafíos éticos de la IA en decisiones autónomas:</b> Sesgos e injusticia algorítmica.</p>
<b>4</b>	<p><b>4 Coexistencia Humano-Máquina</b></p> <p><b>4.1 ¿Desempleo o reconfiguración del mercado en la era de la IA?</b> Retos políticos, jurídicos y sociales para un cambio inevitable.</p> <p><b>4.2 Posthumanismo e IA:</b> Identidad y corporalidad en el mundo de la IA.</p>

### LECTURAS

(Utilizaremos un traductor digital para que todas las lecturas puedan ser accesibles en español. Las lecturas en negrita son obligatorias.)

#### Unidad 0: Introducción

- Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford, UK: Oxford University Press.
- Christian, B. (2011). The Most Human Human: What Artificial Intelligence Teaches Us About Being Alive. New York, NY: Doubleday.
- Crawford, K. (2021). Atlas of AI. Capítulos 1 ("Earth") 2 ("Labor"), 3 ("Data") y 4 ("Classification").
- **Ford, M. (2018). Architects of Intelligence: The truth about AI from the people building it. Birmingham, UK: Packt Publishing.**
- Goldstein, S., & Kirk-Giannini, C. D. (2023). AI Wellbeing.
- **Kaplan, J. (2016). Artificial Intelligence: What Everyone Needs to Know. New York, NY: Oxford University Press.**
- **Levinstein, B. (2023). A Conceptual Guide to Transformers, Part 1.**
- McCorduck, P. (2004). Machines who think: A personal inquiry into the history and prospects of artificial intelligence. A.K. Peters/CRC Press.

- **Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. New York, NY: Farrar, Straus and Giroux.**
- Russell, S., & Norvig, P. (2016). Artificial intelligence: A modern approach (3ra ed.). Pearson.

#### Unidad 1: Consciencia

- Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18(2), 227-287.
- **Boden, M. A. (2006). Mind as machine: A history of cognitive science. Oxford University Press.**
- **Butlin, P., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness.**
- Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory.* Oxford University Press.
- Dennett, D. C. (1991). *Consciousness Explained.* Boston: Little, Brown and Co.
- Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417-457.
- **Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433-460.**

#### Unidad 2: Realidad

- Bostrom, N. (2003). Are You Living in a Computer Simulation? *Philosophical Quarterly*, 53(211), 243-255. <https://doi.org/10.1111/1467-9213.00309>.
- Chalmers, D. J. (2004). The matrix as metaphysics. En *Philosophers explore the matrix* (pp. 132-176). Oxford University Press.
- **Chalmers, D. J. (2016). The Virtual and the Real. *Disputatio*, 9(46), 309-352. <https://doi.org/10.2478/disp-2018-0019>.**
- Chalmers, D. (2023). Reality+, Capítulo 10 ("Do Virtual Reality Headsets Create Reality?") y Capítulo 17 ("Can You Lead a Good Life in a Virtual World?")
- Floridi, L. (2008). Ontological, Epistemological, and Ethical Perspectives of Artificial Agents: Towards an Ecological Naturalism. En S. J. Sullivant (Ed.), *Philosophy of Computer Science* (pp. 97-108). New York: Wiley.
- **Müller, V. C., & Cannon, M. (2022). Existential risk from AI and orthogonality: Can we have it both ways? *Ratio*, 35(1), 25-361.**
- **Nozick, R. (1989). Philosophical explanations. Harvard University Press. (Sección sobre la "Experiencia simulada")**
- **Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.**
- Simondon, G. (1980). On the mode of existence of technical objects. University of Western Ontario. (Originalmente publicado en 1958)

#### Unidad 3: Ética

- Anderson, S. L., & Anderson, M. (Eds.). (2011). *Machine Ethics*. Cambridge.
- Asimov, I. (1950). *I, Robot*. Gnome Press.
- Boden, M. A. (2004). *The Creative Mind: Myths and Mechanisms* (2nd ed.). Routledge.
- **Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press.**
- **Gaffar, H., & Albarashdi, S. (2024). *Copyright Protection for AI-Generated Works: Exploring Originality and Ownership in a Digital Landscape*. *Asian Journal of International Law*.**
- McCormack, J., & d'Inverno, M. (Eds.). (2012). *Computers and Creativity*. Springer.
- Omohundro, S. (2008). *The Basic AI Drives*.
- O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Crown.
- **Sebo, J., & Long, R. (2023). *Moral Consideration for AI Systems by 2030*.**

#### Unidad 4: Coexistencia Humano-Máquina

- Autor, D. (2015). Why are there still so many jobs? The history and future of workplace automation. *Journal of Economic Perspectives*, 29(3), 3-30.
- **Braidotti, R. (2013). *The posthuman. Polity*.**
- **Frey, C. B., & Osborne, M. A. (2017). *The future of employment: How susceptible are jobs to computerisation? Technological Forecasting and Social Change*, 114, 254-280.**
- Goldstein, S., & Kirk-Giannini, C. D. (2023). Language Agents Reduce the Risk of Existential Catastrophe.
- Haraway, D. (1991). A Cyborg Manifesto: Science, Technology, and Socialist-Feminism in the Late Twentieth Century. In *Simians, Cyborgs and Women: The Reinvention of Nature* (pp. 149-181). Routledge.
- Tegmark, M. (2017). *Life 3.0: Being human in the age of artificial intelligence*. Knopf.
- Wiener, N. (1950). *The human use of human beings: Cybernetics and society*. Houghton Mifflin.