SYLLABUS

COURSE: Introduction to Neurophilosophy VIII:

The ethics of artificial intelligence: critical approaches

INSTRUCTORS Professors **Nythamar de Oliveira/Camila Palhares Barbosa**

71535-03 CREDITS: 3.0 45 h YEAR/SEMESTER: 2024/2

Tuesdays at 9 a.m. Room 229 Building 8 (School of Humanities - PUCRS)

COURSE DESCRIPTION:

This course will begin with an introduction to Bioethics, Neuroethics, and AI Ethics, and will proceed to present the contemporary debates on the subject and the various critiques of mainstream AI ethics, specially from racial theory, feminist and LGBTQIA+ demands to a more inclusive and diverse approach. No previous knowledge of neurophilosophy, bioethics, AI or moral epistemology is required. All classes, readings and discussions will be conducted in English, but there is no requirement of proficiency in this language as the seminar will contribute to improve students' functioning in English.

METHODOLOGY:

This class will be run seminar-style. Students will be encouraged to take part in discussions and present papers or serve as commentators. Student presentations will be similar to our workshop and class presentations, in which a mini-paper about the reading may be presented or an original contributing paper. Afterwards, fellow students will comment on the presentation.

OBJECTIVES:

An introduction to neurophilosohy, neurophenomenology, social epistemology, and in-depth discussions on major topics and problems of Social Epistemology, Neuroethics, AI Ethics. Reading assignments from major contributions to this field.

PROGRAMMATIC CONTENTS:

Week 1: Self-introductions and presentation of the syllabus

Week 2: AI Ethics and Critical Theory - Camila
Reading assignment: Waelen, R. Why AI Ethics Is a Critical Theory. Philos. Technol. 35, 9 (2022). https://doi.org/10.1007/s13347-022-00507-5

Week 3: AI Ethics and Critical Theory - Nythamar
Reading assignment: Oliveira, N. A decolonial critical theory of artificial intelligence: intersectional egalitarianism, moral alignment, and AI governance. Dossier • Filos. Unisinos 25 (1), 2024. https://doi.org/10.4013/fsu.2024.251.14

Week 4: Social choice ethics in artificial intelligence
Reading assignment: Baum, S.D. Social choice ethics in artificial intelligence. AI & Soc 35, 165–176 (2020). https://doi.org/10.1007/s00146-017-0760-1.

Week 5: Algorithmic injustice
Reading assignment: Birhane, A. Algorithmic injustice: a relational ethics approach. Perspective| Volume 2, ISSUE 2, 100205, February 12, 2021. https://doi.org/10.1016/j.patter.2021.100205

Week 6: Moral Agency and Robots
Reading assignment: Véliz, C. Moral zombies: why algorithms are not moral agents. AI & Soc 36, 487–497 (2021). https://doi.org/10.1007/s00146-021-01189-x

Week 7: Feminist Critique
Reading assignment: West, Sarah Myers (2020). Redistribution and Rekognition: A Feminist Critique of Algorithm Fairness. Catalyst: Feminism, Theory, Technoscience, 6(2), 1–24. http://www.catalystjournal.org | ISSN: 2380-3312

Week 8: Applied Ethics in AI
Reading assignment: Sterri, A. B., & Earp, B. D. (in press). The ethics of sex robots. In C. Véliz (ed.), The Oxford Handbook of Digital Ethics. Oxford: Oxford University Press. Available online ahead of print at https://www.academia.edu/42871768/The_ethics_of_sex_robots.

Week 9: Algorithm Oppression: proposing a shift in the discussion
No Reading assignment: Presentation of a paper/draft by Camila

Week 10 to Week 15: Discussions and student presentations
Timnit Gebru, Oxford Handbook on AI Ethics Book Chapter on Race and Gender
https://arxiv.org/abs/1908.06165

GRADING POLICY:

Grades are based on point accumulation throughout the fifteen weeks, comprising class participation (by attending classes, raising questions, sharing students' views, insights, comments, and criticisms with classmates), quizzes (multiple-choice or truth/false), short essays and/or a research paper.

AI Ethics, Gender & Race Bibliography

Benjamin, Ruha. *Race After Technology: Abolitionist Tools for the New Jim Code*. John Wiley & Sons, 2019.

Broussard, Meredith. *Artificial unintelligence: how computers misunderstand the world*. MIT Press, 2018.

Buolamwini, Joy, and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification." In *Conference on fairness, accountability and transparency*, pp. 77-91. 2018.

Dubber, Markus D., Frank Pasquale, and Sunit Das. *The Oxford Handbook of Ethics of AI*. Oxford University Press, 2020.

Eubanks, Virginia. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

Hamidi, Foad, Morgan Klaus Scheuerman, and Stacy M. Branham. "Gender recognition or gender reductionism?: The social implications of embedded gender recognition systems." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

Hicks, Marie. *Programmed inequality: How Britain discarded women technologists and lost its edge in computing*. MIT Press, 2017.

Noble, Safiya U. *Algorithms of oppression: How search engines reinforce racism*. NYU Press, 2018.

O'Neil, Cathy. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books, 2016.

Stitzlein, Sarah M. "Replacing the 'View from Nowhere': A Pragmatist-Feminist Science Classroom." Electronic Journal of Science Education (2004).

BASIC BIBLIOGRAPHY (Neuroethics, AI Ethics, Social and Political Emotions):

Alcoff, Linda Martin and Eduardo Mendieta, eds. *Identities: Race, Class, Gender, and Nationality*. Oxford: Blackwell, 2008.
Anderson, Elizabeth. "The Social Epistemology of Morality: Learning from the Forgotten History of the Abolition of Slavery", in *The Epistemic Life of Groups: Essays in the Epistemology of Collectives*, M. Brady and M. Fricker (eds.), Oxford: Oxford University Press, 2014.
Athanasiou, Athena, Pothiti Hantzaroula, and Kostas Yannakopoulos. "Towards a New Epistemology: The Affective Turn." *Historein* 8 (2008).
Bickle, John, Peter Mandik, and Anthony Landreth, "The Philosophy of Neuroscience", The Stanford Encyclopedia of Philosophy (Fall 2019 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2019/entries/neuroscience/>.
Blackburn, S*. Rulling Passions*. Oxford: Oxford University Press, 1998.
Ben-Ze'ev, Aaron. *The Subtlety of Emotions. A Bradford Book.* The MIT Press. Cambridge, Massachusetts. London, England, 2000.
Ahmed, Sara. *Cultural Politics of Emotion*. Routledge, 2004 (2nd edition 2014)
Bringsjord, Selmer and Naveen Sundar Govindarajulu, "Artificial Intelligence", *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/artificial-intelligence/>.
Chalmers, David J. *The character of consciousness*. Oxford University Press, 2010.
Chalmers, David J. A Computational Foundation for the Study of Cognition. 1994. http://consc.net/papers/computation.html
Chalmers, David J. *Reality+ : Virtual Worlds and the Problems of Philosophy*. New York: W.W. Norton & Company, 2022.
Churchland, Patricia. Braintrust: What neuroscience tells us about morality.Princeton U Press, 2011.
Clough, Patricia Ticineto and Jean O'Malley Halley (eds), *The Affective Turn: Theorizing the Social*, Durham: Duke UP, 2007.
Damásio, António. *Self Comes to Mind: Constructing the conscious brain*.Pantheon, 2010.
Damásio, António. *Looking for Spinoza: Joy, sorrow and the feeling brain*. New York: Harcourt, Inc., 2003.
Feenberg, A. 2017. *Technosystem: The social life of reason*. Cambridge, Massachusetts: Harvard University Press.
Gibbard. Allan. *Wise choices, apt feelings: a theory of normative judgement*. Cambridge: Harvard University Press, 1990.

Haraway, Donna (1995). "Saberes localizados: a questão da ciência para o feminismo e o privilégio da perspectiva parcial". *Cadernos Pagu,* Campinas, SP, n. 5, p. 7-4. Disponível em: <https://periodicos.sbu.unicamp.br/ojs/index.php/cadpagu/article/view/1773/1828>
Haraway, Donna. *Primate visions: gender, race, and nature in the world of modern science.* New York, Routledge. 1989.
Harding, S. *The Science Question in Feminism*. Ithaca & Londres: Cornell U Press, 1986.
Jaeggi, Rahel. *Critique of Forms of Life*. Harvard University Press, 2016.
Lutz, C. A. and Lila Abu-Lughod, *Language and the Politics of Emotion*, Cambridge UP, 1990.
Malabou, Catherine. *What Should We Do with Our Brain?* Fordham University Press, 2008. *Que faire de notre cerveau?* Bayard, 2004.
Malabou, Catherine. *Morphing intelligence: From IQ measurement to artificial Brains*. Translated by Carolyn Shread. New York: Columbia University Press, 2019. *Métamorphoses de l'intelligence*: *Que faire de leur cerveau bleu?* Paris: PUF, 2017.
Malabou, Catherine. *Ontologie de l'Accident*: *Essai sur la plasticité destructice*. Paris: Léo Scheer, 2009.
Malabou, Catherine. *Changer de différence*: *Le féminin et la question philosophique*. Paris: Galilée, 2009.
Malabou, Catherine. *Les Nouveaux Blessés*: *De Freud à la neurologie, penser les traumatismes contemporains*. Paris: Bayard, 2007.
Malabou, Catherine. The end of writing? Gramatology and plasticity. *The European Legacy*: T*owards new paradigms*, Cambridge, v. 12, n. 4, p. 431-444, 2007b.
Malabou, Catherine. *La plasticité au soir de l'écriture*: *Dialetique, destruction, déconstruction*. Paris: Léo Scheer, 2005.
Nussbaum, Martha.*Political emotions:Why love matters for justice*.Harvard U Press, 2013.
Nussbaum, M. *Upheavals of Thought: The Intelligence of Emotions.* Cambridge University Press, 2001.
Prinz, Jesse. *Gut Reactions: A Perceptual Theory of Emotion*. Oxford U Press, 2004.
Rescorla, Michael, "The Computational Theory of Mind", *The Stanford Encyclopedia of Philosophy* (Fall 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2020/entries/computational-mind/>.
Russell, Stuart J.; Norvig, Peter. *Artificial Intelligence: A Modern Approach*. 4th Edition. London: Pearson, 2022.
Solomon. R. *Not Passion's Slave: Emotion and Choice.* New York: Oxford University Press, 2003.
Stets, Jan E. and Jonathan H. Turner (eds), *Handbook of the Sociology of Emotions*. New York: Springer, 2006
White, G. M. "Emotions Inside Out: The Anthropology of Affect," in M. Lewis and J. M. Haviland (eds), *Handbook of Emotions*, New York: Guilford, 1993.
Yates, Candida. *The Play of Political Culture, Emotion and Identity*. New York: Palgrave Macmillan, 2015.

Secondary Bibliography (Neurophilosophy):

Bear, Mark, Barry Connors & Michael Paradiso. 2006. *Neuroscience: Exploring the Brain.* Lippincott Williams & Wilkins. Third edition.
Bickle, John (editor) 2008. *Oxford Handbook of Philosophy and Neuroscience*. Oxford University Press.
Brogaard, Berit (editor). 2014a. *Does Perception Have Content?* Oxford University Press.
Bueno, Otavio and Scott A. Shalkowski. 2015a. Modalism and Theoretical Virtues: Toward an Epistemology of Modality. *Philosophical Studies* Volume 172, Issue 3: 671–689.
Chalmers, David J. 2002. *Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press.
Changeux, Jean-Pierre & Paul Ricoeur. 1998. *Ce qui nous fait penser. La nature et la règle*. Paris: Odile Jacob. ET: *What Makes Us Think. A Neuroscientist and a Philosopher Argue About Ethics, Human Nature, and the Brain* (Princeton U. Press, 2000)
Changeux, Jean-Pierre. 2005. Creation, Art and the Brain. In: Changeux et al. (Eds.) *The Neurobiology of Human Values.* New York: Springer-Verlag, 2005, p. 1-10.
Changeux Jean-Pierre and Dehaene Stanislas (1989), Neuronal Models of Cognitive Function, *Cognition* 33, pp. 63–109.
Chatterjee, Anjan and Martha J. Farah, editors. 2013. *Neuroethics in practice*. Oxford University Press.
Churchland, Patricia S. 1986. *Neurophilosophy: Toward A Unified Science of the Mind-Brain*. Bradford Books. Cambridge, MA: MIT Press.
Churchland, Patricia S. 2002. *Brain-Wise: Studies in neurophilosophy*. NY: Bradford.
Churchland, P. M. 1984. *Matter and Consciousness*. Cambridge, MA: Bradford/MIT Press.
Churchland, P. M. 1994. *The Engine of Reason, The Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge: MIT Press.
Churchland, P.S., and T. J. Sejnowski. 1992. *The Computational Brain: Models and Methods on the Frontiers of Computational Neuroscience*. Cambridge, MA: MIT Press.
Craib, Ian; Benton, Ted (2010). *Philosophy of the Social Sciences: the philosophical foundations of social thought*. Londres: Palgrave Macmillan.
Damásio, António. 1999. *The feeling of what happens: Body and emotion in the making of consciousness*. New York: Harcourt Brace.
Damásio, António. 1994. *Descartes' Error: Emotion, reason, and the human brain*. New York: Putnam.
Dancy, Jonathan. 2000. *Normativity*. Malden, Mass.: Blackwell.
Darwin, Charles. 1872. *The Expression of the Emotions in Man and Animals*. New York: New York Philosophical Library. (Reprint, 1960).
Darwin, Charles. 1980. *Metaphysics, Materialism, and the Evolution of Mind: Early Writings of Charles Darwin*. Transcribed and annotated by Paul H. Barrett with a commentary by Howard E. Gruber. Chicago: The University of Chicago Press.
Darwin, Charles. 1981. *The Descent of Man, and Selection in Relation to Sex*. Princeton: Princeton University Press.
Dehaene, Stanislas. 1997. *The number sense*. New York: Oxford University Press.

Dehaene, Stanislas. 2009. *Reading in the brain*. New York: Penguin.
Dehaene, Stanislas. 2014a. *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts*. Viking Adult.
Dennett, Daniel C. 2006. *Breaking the spell: Religion as a natural phenomenon*. Viking.
Dennett, Daniel C. 1991. *Consciousness Explained*. New York: Little Brown.
Dennett, Daniel C. 1995. *Darwin's dangerous idea: Evolution and the meanings of life*. New York: Simon & Schuster.
De Oliveira, Nythamar. 2016a. On Ritalin, Adderall, and Cognitive Enhancement: Metaethics, Bioethics, Neuroethics. ethic@ 15/2 (2016)
De Oliveira, Nythamar. 2016b. Revisiting the Mind-Brain Reductionisms: Contra Dualism and Eliminativism. Veritas 61/3 (2016)
De Sousa, Ronald. 1999. *The Rationality of Emotion*. Cambridge, MA: MIT Press.
De Waal, Frans. 2006. *Primates and Philosophers: How Morality Evolved*. Edited by Stephen Macedo and Josiah Ober. Princeton: Princeton University Press.
De Waal, Frans. 1996. *Good Natured: The Origins of Right and Wrong in Humans and Other Animals*. Cambridge, MA: Harvard University Press.
Dunbar, Robin I. M. 1998. The Social Brain Hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews* Volume 6, Issue 5 (1998): 178–190.
Ekman, Paul. 1972. *Emotions in the Human Face*. New York: Pergamon Press.
Farah, M.J. 1994. Neuropsychological Inference with an Interactive Brain: A Critique of the Locality Assumption. *Behavioral and Brain Sciences*.
Farah, Martha J. 2012. Neuroethics: The Ethical, Legal, and Societal Impact of Neuroscience. *The Annual Review of Psychology* 63: p. 571-591.
Farah, Martha J.and Todd E. Feinberg, editors. 2005. *Patient-Based Approaches to Cognitive Neuroscience.* 2nd Edition. Cambridge, Mass.: MIT Press.
Gallagher, Shaun and Dan Zahavi. 2008. *The Phenomenological Mind: An Introduction to Philosophy of Mind and Cognitive Science*. Routledge.
Gazzaniga, Michael. 1985. *The social brain*. New York: Basic Books.
Gazzaniga, Michael S. 2005. *The Ethical Brain*. Dana Press.
Gazzaniga, Michael S., ed. 2005. *The Cognitive Neurosciences*. 4th ed. MIT Press.
Giordano, James, and Bert Gordijn, eds. 2010. *Scientific and Philosophical Perspectives in Neuroethics*. Cambridge University Press.
Glannon, Walter. 2011. *Brain, Body, and Mind: Neuroethics with a Human Face*. O.U.P.
Gouveia, Steven. *Philosophy and Neuroscience: A Methodological Analysis*. Palgrave Macmillan (Springer Nature), 2022.
Gouveia, Steven (Editor). *Thinking the New World: Conversations on Artificial Intelligence*, Amazon.
Haddock, Adrian, Alan Millar, and Duncan Pritchard (eds). 2011. *Social Epistemology.* Oxford University Press.
Haidt, Jonathan. 2001. The Emotional dog and its rational tail. *Psychological Review* Vol. 108. No. 4, 814-834.
Harding. S. *Objectivity and Diversity: another logic of scientific research*. Chicago e Londres: The University of Chicago Press, 2015.

Johnson-Laird, Philip. 2008. *How we Reason*. Oxford University Press.
Korsgaard, Christine M. 2010. Reflections on the Evolution of Morality. *The Amherst Lecture in Philosophy* 5 (2010): 1–29.
Moll, Jorge et al. 2002. The neural correlates of moral sensitivity. *Journal of Neuroscience*. Apr 22 (7): 2730-6.
Nadelhoffer, Thomas, Eddy A. Nahmias & Shaun Nichols (eds.) 2010. *Moral Psychology: Historical and Contemporary Readings*. Wiley-Blackwell.
Nadelhoffer, Thomas and Walter Sinnott-Armstrong. 2012. Neurolaw and Neuroprediction: Potential Promises and Perils. *Philosophy Compass* 7/9 (2012): 631–642.
Nichols, Shaun. 2010. Emotions, norms, and the genealogy of fairness. *politics, philosophy & economics* 9 (1): 1-22.
Prinz, Jesse. 2002. *Furnishing the Mind: Concepts and Their Perceptual Basis*. MIT Press.
Prinz, Jesse. 2012. *The Conscious Brain*. Oxford University Press.
Prinz, Jesse. 2004. *The Emotional Construction of Morals*. Oxford University Press.
Putnam, Hilary. 1975. The Meaning of 'Meaning'. In *Mind, Language and Reality. Philosophical Papers*, Volume 2. Cambridge University Press.
Railton, Peter. 2003. *Facts, Values, and Norms: Essays Toward a Morality of Consequence*. Cambridge University Press.
Rose, Nikolas. 2006. *The Politics of Life Itself: Biomedicine, Power, and Subjectivity in the Twenty-First Century*. Princeton, NJ: Princeton University Press.
Roskies, A. 2002. Neuroethics for the new millennium. Neuron 35: 21–3.
Ryle, Gilbert. 1949.*The Concept of Mind*. The University of Chicago Press.
Schaber, Peter (ed.). 2004. *Normativity and Naturalism*. Heusenstamm: Ontos Verlag.
Searle, John. 1984. Can Computers Think? In: *Minds, Brains and Science. Reith Lectures*. Cambridge, Mass.: Harvard University Press.
Searle, John. 1995. *The Construction of Social Reality*. New York: Free Press.
Zimmerman, Aaron. *Moral Epistemology.* 2010. New York: Routledge.