Podiplomskašola ZRCSAZU
Novi trg 2
1000 Ljubljana
T: +386 1 470 64 51

UČNI NAČRT PREDMETA / COURSE SYLLABUS	
Predmet:	Filozofija umetne inteligence: človeški stroji, tehnološka singularnost in superinteligenca
Course title:	The philosophy of artificial intelligence: human machines, technological singularity and superintelligence

Študijski program in stopnja **Study programme and level**	**Študijska smer oz. modul** **Study field or module**	**Letnik** **Academic year**	**Semester** **Semester**
Primerjalni študij idej in kultur, doktorski študij 3. stopnje	Interdisciplinarni študij institucij in družbe 21. stoletja - politike, ekonomije, tehnologije, epistemologije		
Comparative studies of ideas and cultures, doctoral study 3rd cycle	Interdisciplinary study of institutions and society in the 21st century – politics, economics, technology, epistemology		

Vrsta predmeta / Course type	Splošni izbirni predmet
Univerzitetna koda predmeta / University course code:	101

Predavanja **Lectures**	**Seminar** **Seminar**	**Vaje** **Tutorial**	**Lab. vaje** **Laboratory work**	**Teren. vaje** **Field work**	**Samost. delo** **Individ. work**	**ECTS**
60	30				90	6

Nosilec predmeta / Lecturer:	doc. dr. Gregor Strle

Jeziki / slo/eng **Languages:**	**Predavanja / Lectures:**	Slovenščina/English
	Vaje / Tutorial:	Slovenščina/English

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:	**Prerequisites:**
Vpis v doktorski študij	Enrollment in doctoral studies

Vsebina:	**Content (Syllabus outline):**
Filozofija AI: • Zgodovisnki pregled: od automate do AI; • Koncept naravne in umetne inteligence; • Umetna splošna inteligenca (AGI - Artificial General Intelligence); • Eksistenčna tveganja, "tehnološka singularnost" in "superinteligenca". Kognitivni mehanizmi: • Kognitivne arhitekture AI;	Philosophy of AI: • Historical overview: from automata to AI; • The concept of natural and artificial intelligence; • Artificial General Intelligence (AGI); • Existential risks, “technological singularity”, and “superintelligence”. Cognitive mechanisms: • Cognitive architectures of AI;

• Teorija uma (Theory of Mind); • Razumevanje in simuliranje človeškega razmišljanja in vedenja; • Od umetne inteligence do umetne zavesti. Etika AI: • Možnost etike umetne inteligence; • (Ne)pristranski algoritmi; • Privatnost in demokratičnost v dobi AI; • Distopije in utopije uporabe AI in AGI; • Ravnovesje med inovacijami in varnostjo.	• Theory of mind; • Understanding and simulating human thinking and behaviour; • From artificial intelligence to artificial "consciousness". Ethics of AI: • Ethics for artificial intelligence; • (Un)biased algorithms; • Privacy and democracy in the age of AI; • Dystopias and utopias of AI and AGI; • The balance between innovation and safety.

Temeljni literatura in viri / Readings:

- Boden, M. A. 2006. *Mind as machine: A history of cognitive science*. Oxford: Clarendon Press.
- Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies* (1st ed.). Oxford: Oxford University Press, Inc.
- Brockman, John. 2019. *Possible Minds: Twenty-Five Ways of Looking at AI*. New York, NY: Penguin Press.
- Brooks, R. A. 1990. "Elephants don't play chess", *Robotics and autonomous systems*, 6(1-2), pp. 3-15.
- Churchland, P. M. 1995. *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.
- Coeckelbergh, Mark. 2020. *AI Ethics*. Cambridge: The MIT Press.
- Deacon, T. W. 1997. *The symbolic species: the co-evolution of language and the brain*. New York: W.W. Norton and Company.
- Deleuze, Gilles. 1992. "Postscript on the Societies of Control", *October*, 59, pp. 3–7.
- Dennett, Daniel C. 1991. *Consciousness Explained*. London: Penguin Books.
- Dreyfus, H. 1972. *What Computers Can't Do*. New York: Harper and Row.
- Gärdenfors, P. 2000. *Conceptual Spaces: The Geometry of Thought*. Cambridge, MA: MIT Press.
- Haugeland, J. 2000. *Having thought: essays in metaphysics of mind*. Cambridge, MA: Harvard University Press.
- Hinton, G. E. 1989. "Connectionist Learning Procedures", *Artificial Intelligence*, 40: 185–234.
- Heidegger, Martin. 1977. *The Question Concerning Technology and Other Essays*, trans. William Lovitt, 1st Harper pbk. ed. New York: HarperPerennial.
- Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: an Eternal Golden Braid*. New York: Basic Books.
- Humphreys, Paul. 2004. *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.
- Kurzweil, R. 1992. *The Age of Intelligent Machines*. Cambridge, MA: The MIT Press.
- Malabou, C., and C. Shread. 2019. *Morphing Intelligence: From IQ Measurement to Artificial Brains*. New York: Columbia University Press.
- Massimi, Michela. 2011. "From Data to Phenomena: A Kantian Stance", Synthese, 182(1): 101–116.
- Mitchell, Melanie. 2021. "Abstraction and analogy-making in artificial intelligence", Annals of the New York Academy of Sciences, 1505 (1). https://doi.org/10.1111/nyas.14619.

- McCarthy, John, Marvin Minsky, Nathan Rochester, Claude Shannon. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*, http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html.
- McCarthy, J., and P. J. Hayes 1969. "Some Philosophical Problems from the Standpoint of Artificial Intelligence", in B. Meltzer and D. Michie (eds.), *Machine Intelligence* 4. Edinburgh, Edinburgh University Press, pp. 463-502.
- Minsky, Marvin L. 1986. *The Society Of Mind*. New York: Simon & Schuster.
- Newell, A., and H.A. Simon. 1959. *The simulation of human thought*. Santa Monica, CA: Rand Corp.
- Russell, S., and P. Norvig. 1995. *Artificial Intelligence: A Modern Approach*. Englewood Cliffs, NJ: Prentice-Hall.
- Searle, J. 1989. "Artificial Intelligence and the Chinese Room: An Exchange", New York Review of Books, 36: 2.
- Turing, A.M. 1950. "Computing machinery and intelligence", Mind 59, pp. 433–460.
- Varela, F. J., E. Thompson, and E. Rosch. 2016. The Embodied Mind. Cognitive Science and Human Experience (Revised). Cambridge, MA: MIT Press.
- Winograd, T., and F. Flores 1987. *Understanding Computers and Cognition: A New Foundation for Design*. Norwood, NJ: Addison-Wesley.

Cilji in kompetence:

- Zgodovinska umestitev in razumevanje AI in AGI (Artificial General Intelligence);
- Vzpostavitev filozofskega konteksta AGI: razmerje človek-stroj, "tehnološka singularnost" in "superinteligenca";
- Poznavanje temeljnih mehanizmov AI: reprezentacije, kognitivni modeli, strojno učenje;
- Razumevanje epistemoloških izzivov AGI, tudi z vidika eksistenčnih tveganj;
- Sposobnost kritično artikulirati prihajajoče transformacije AI in AGI za posameznika in družbo.

Objectives and competences:

- The ability to provide historical context in the evolution of AI;
- The ability to gain knowledge of the core mechanisms of AI: representation, cognitive modelling, machine-learning, etc.;
- The ability to extricate a philosophical context for AGI: the human-machine relationship, "technological singularity", and "superintelligence";
- The ability to engage in the epistemological challenges of AGI, including an assessment of existential risks;
- The ability to critically articulate the coming transformations of AI and AGI in relation to the individual and society.

Predvideni študijski rezultati:

Slušateljice in slušatelji se bodo soočili z nujnostjo same filozofske utemeljitve predpostavljenih pojmov kot tudi vzporedno določitev kognitivnih, etičnih, ekonomskih, socialnih in političnih razsežnosti umetnih inteligenc.

Intended learning outcomes:

Students will comprehend the necessity of a new philosophical grounding of the notions assumed, while in parallel also seeking determination of cognitive, ethical, economic, social and political dimensions of AGI.

Metode poučevanja in učenja:	Learning and teaching methods:
Oblike dela:	Types of learning/teaching:
☒ Frontalna oblika poučevanja	☒ Frontal teaching
☒ Delo v manjših skupinah oz. v dvojicah	☒ Work in smaller groups or pair work
☒ Samostojno delo študentov	☒ Independent students work
☒ e-izobraževanje	☒ e-learning
Metode (načini) dela:	Teaching methods:
☒ Razlaga	☒ Explanation
☒ Razgovor/ diskusija/debata	☒ Conversation/discussion/debate
☒ Delo z besedilom	☒ Work with texts
☒ Proučevanje primera	☒ Case studies
☐ Igra vlog	☐ Roleplay
☒ Druge vrste nastopov študentov	☒ Different presentation
☒ Reševanje nalog	☒ Solving exercises
☒ »Terenske vaje« (npr. obiski podjetij)	☒ Field work (e.g. company visits)
☒ Vključevanje gostov iz prakse	☒ Inviting guests from companies

Načini ocenjevanja:	Delež (v %) / Weight (in %)	Assessment
Krajši pisni izdelki	20	Short written assignments
Daljši pisni izdelki	80	Long written assignments

Reference nosilca / Lecturer's references:

- Strle, Gregor, Košir, Andrej, Stojmenova Pečečnik, Kristina, Sodnik, Jaka. 2022. The effects of driving disengagement on response time in transition to manual driving mode. V: Stojmenova Pečečnik, Kristina (Ur.), Jakus, Grega (Ur.). HCI SI 2022 : 7th Human-Computer Interaction Slovenia Conference 2022 : proceedings of the 7th Human-Computer Interaction Slovenia conference 2022 : Ljubljana, Slovenia, November 29, 2022. [Aachen]: CEUR-WS, cop. 2022. Str. 1-10, ilustr. CEUR workshop proceedings, vol. 3300. ISSN 1613-0073
- Strle, Gregor, Košir, Andrej, Oğuz, Evin Aslan, Burnik, Urban. 2022. Predicting user engagement in video advertisement : insights from pupillary response and heart rate. V: STOJMENOVA PEČEČNIK, Kristina (Ur.), JAKUS, Grega (Ur.). HCI SI 2022 : 7th Human-Computer Interaction Slovenia Conference 2022 : proceedings of the 7th Human-Computer Interaction Slovenia conference 2022 : Ljubljana, Slovenia, November 29, 2022. [Aachen]: CEUR-WS, cop. 2022. Str. 1-10, ilustr. CEUR workshop proceedings, vol. 3300. ISSN 1613-0073.
- Strle, Gregor, Xing, Yilun, Miller, Erika E., Boyle, Linda Ng, Sodnik, Jaka. 2021. Take-over time : a cross-cultural study of take-over responses in highly automated driving. Applied sciences. Sep.-1 2021, no. 17, 7959, str. 1-10, ilustr. ISSN 2076-3417.
- Strle, Gregor. 2021. Realist and cognitive perspectives on meaning and semantics. Traditiones. 2021, letn. 50, št. 2, str. 17-34. ISSN 0352-0447.
- Meža, Marko, Košir, Andrej, Strle, Gregor. 2020. Measuring the induction of affect using facial expression analysis technology : a pilot study. V: Žemva, Andrej (ur.), Trost, Andrej (ur.). Proceedings of the Twenty-ninth International Electrotechnical and Computer Science Conference ERK 2020. ERK 2020, Portorož, Slovenija, 21.-22. september 2020. Ljubljana: Slovenian Section IEEE, 2020. Str. 348-350, ilustr.