CENTRO TÉCNICO CIENTÍFICO	CENTRO DE CIÊNCIAS SOCIAIS	
DEPARTAMENTO DE INFORMÁTICA	DEPARTAMENTO DE CIÊNCIAS SOCIAIS	
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA	PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS SOCIAIS	2026.1

Projeto de Sistemas de IA Confiáveis e Socialmente Responsáveis

INF2921 (3WA) TOP ENGENHARIA DE SOFTWARE V

CIS2114 (2NA) SEMINARIOS ESPECIAIS VI

5ª feira – 16h às 18h + 1h SHF

Carga horária total: 3h Créditos: 3

Profs. Renato Cerqueira e Gabriel Banaggia

EMENTA

Ementa de conteúdo variável.

OBJETIVOS

Estudar técnicas e desafios de como elaborar e executar projetos de sistemas de IA que sejam simultaneamente confiáveis e socialmente responsáveis. Capacitar estudantes a conceber e executar projetos que integrem fundamentos técnicos, metodológicos e éticos no ciclo completo de desenvolvimento de sistemas de IA, desde a concepção e definição de requisitos e a captação de dados até a implementação, avaliação e monitoramento em contexto real. Promover compreensão crítica e aplicação prática em projetos concretos de princípios como transparência, explicabilidade, robustez, segurança, equidade, privacidade e prestação de contas. Ao final do curso, espera-se que a turma seja capaz de entender as principais abordagens e desafios para projetar soluções de IA alinhadas a marcos regulatórios, normas técnicas e demandas sociais, avaliando riscos, impactos e compensações.

Atenção: Esta disciplina aceita estudantes matriculados em qualquer curso de pós-graduação, e os de graduação que tenham cumprido o requisito formal de 60% do total de créditos já cursados. Em ambos os casos, não há exigência de conhecimento prévio de programação, mas experiência com projetos tecnológicos é bem-vinda. A disciplina tem um viés prático, com avaliação baseada principalmente na concepção

e prova de conceito de um sistema de IA. Os projetos terão tarefas distribuídas por perfil (técnico, metodológico, design, especialista do domínio, análise de impacto etc.).

PROGRAMA

Entre os tópicos a serem trabalhados na disciplina, encontram-se:

- Fundamentos de IA e Sistemas baseados em IA
- Formulação de escopo e definição de requisitos
- Design de tecnologias e sistemas de IA
- Ciclo de vida de sistemas de IA
- Partes interessadas e análise sociotécnica
- Captação, curadoria, conformidade e controle de dados
- Transparência, explicabilidade e interpretabilidade
- Robustez, segurança e confiabilidade
- IA socialmente responsável
- Avaliação de impactos, riscos e compensações (*trade-offs*)
- Detecção e mitigação de viés em modelos de IA
- Monitoramento, auditoria e responsabilização
- IA epistêmica e governável

AVALIAÇÃO

A avaliação na disciplina é composta por diferentes componentes, incluindo ao menos:

1. Participação ativa durante sessões e seminários da disciplina
2. Concepção de um projeto de sistema baseado em IA
3. Desenvolvimento e teste de uma prova de conceito (TRL <= 3)

Os projetos serão desenvolvidos em grupos, a serem montados coletivamente nas primeiras semanas de aula, mas os componentes irão receber notas de maneira individual. Haverá tarefas específicas pensadas segundo as habilidades de cada integrante do grupo.

BIBLIOGRAFIA PRINCIPAL

Blackwell, Alan. 2024. *Moral Codes: Designing Alternatives to AI* (capítulos 1, 9 e 14). Cambridge: The MIT Press. Disponível em: https://direct.mit.edu/books/oa-monograph/5814/Moral-CodesDesigning-Alternatives-to-AI

High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines For Trustworthy Ai*. Disponível em: https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Hong, Sun-ha. 2024. "Strategic Misrecognition and Speculative Rituals in Generative AI." *Journal of Digital Social Research*, *6* (4), pp. 92-106. Disponível em: https://publicera.kb.se/jdsr/article/view/40474

Huyen, Chip. 2025. *AI Engineering*. O'Reilly Media, Inc. Disponível em: https://www.oreilly.com/library/view/ai-engineering/9781098166298/

Jannel, R. & J. Tallant. 2026. "Trustability and trustworthiness: conceptual foundations and the case of AI". *AI Ethics*, *6* (13). Disponível em: https://doi.org/10.1007/s43681-025-00839-w

Messeri, Lisa, & M. J. Crockett. 2024. "Artificial Intelligence and Illusions of Understanding in Scientific Research." *Nature*, *627* (8002), pp. 49-58. Disponível em: https://www.nature.com/articles/s41586-024-07146-0

Norman, Don. 2024. *Design for a Better World: Meaningful, Sustainable, Humanity Centered*. MIT Press. Disponível em: https://mitpress.mit.edu/9780262548304/design-for-a-better-world/

BIBLIOGRAFIA COMPLEMENTAR

Chen, Pin-Yu. & Liu, Sijia. 2025. *Introduction to Foundation Models*. Springer. Disponível em: https://doi.org/10.1007/978-3-031-76770-8

Christian, Brian. 2020. *The Alignment Problem*. New York: W. W. Norton & Company. Disponível em: https://brianchristian.org/the-alignment-problem/

Crawford, Kate. 2021. *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven: Yale University Press. Disponível em: https://www.jstor.org/stable/j.ctv1ghv45t

European Association of Research and Technology Organisations. 2014. "The TRL Scale as a Research & Innovation Policy Tool, EARTO Recommendations". *EARTO Publications*. Disponível em: https://www.earto.eu/wp-content/uploads/The_TRL_Scale_as_a_R_I_Policy_Tool_-_EARTO_Recommendations_-_Final.pdf

Fundação Itaú. 2025. *Stanford Social Innovation Review Brasil*. *Número especial inteligência artificial*. Disponível em: https://ssir.com.br/edicao-especial/futuros-possiveis-ia/

Morris, Robert JT. 2025. *Healthcare Transformation using Artificial Intelligence*. Cambridge: Academic Press.

Vilsmaier, Ulli, & Julie Thompson Klein. 2023. "Boundary Work." in Thorsten Philipp & Tobias Schmohl (eds.). *Handbook Transdisciplinary Learning*. Bielefeld: Transcript Verlag. Disponível em: https://www.transcript-publishing.com/978-3-8376-6347-1/handbook-transdisciplinary-learning/