PERÍODO: 2018.2		
CENTRO UNIVERSITÁRIO CTC DEPARTAMENTO DE INFORMÁTOICA		
INF2791	Profª. Clarisse Sieckenius de Souza	
Tópicos em IHC II	CARGA HORÁRIA TOTAL: 45	CRÉDITOS: 3
Explainable AI IA Explicável	Pré-requisito: Formal Nenhum (Recomenda-se algum conhecimento sobre Inteligência Artificial e Redes Neurais)	

OBJETIVOS	1) Introduzir a questão dos 'significados humanos' embutidos em – e atribuídos a – aplicações contemporâneas de Inteligência Artificial 2) Promover uma reflexão especificamente voltada para responsabilidade social e conduta ética no desenvolvimento de sistemas com aprendizado automático (Machine Learning), redes neurais (Neural Nets) e certos tipos de sistemas de apoio a decisão baseados em grandes volumes de dados (Data-Driven Decision Support Systems).
EMENTA	• *General Data Protection Regulation* (GDPR), a regulamentação europeia de Maio de 2018 sobre direitos dos donos de dados (Data Owners), em particular o *direito à explicação* (Right to Explanation); • Diferentes perspectivas sobre o que é uma "explicação"; explicabilidade e interpretabilidade de modelos de aprendizado automático; • O problema da explicação automática nas décadas de 1980 e 1990; o que mudou por volta de 2015; • Valores e vieses embutidos em algoritmos (*algorithmic accountability*); • Itens para uma agenda de pesquisa sobre responsabilidade social no desenvolvimento de modelos e aplicações de aprendizado de máquina.
PROGRAMA	**Aula 1:** Introdução ao assunto; Panorâmica do Estado da Arte; a *GDPR*; **Aula 2:** Aspectos do direito a explicação; **Aula 3:** Exercitando interpretações do 'direito a explicação' em cenários reais e plausíveis;

	Aula 4: Seminário sobre o que constitui um *explicação*; **Aula 5:** Seminário sobre o que constitui um *explicação* (continuação); **Aula 6:** *Explicações* em sistemas especialistas (IA simbólica); **Aula 7:** Desafios de sistemas que utilizam redes neurais ou aprendizado profundo; **Aula 8:** Desafios de sistemas que utilizam redes neurais ou aprendizado profundo (continuação); **Aula 9:** Seminário sobre *Explainable AI*; **Aula 10:** Seminário sobre *Explainable AI* (continuação); **Aula 11:** Valores e vieses embutidos em algoritmos (*algorithmic accountability*); **Aula 12:** Valores e vieses embutidos em algoritmos (continuação); **Aula 13:** Itens para uma agenda de pesquisa sobre responsabilidade social no desenvolvimento de modelos e aplicações de aprendizado de máquina; **Aula 14:** Itens para uma agenda de pesquisa sobre responsabilidade social no desenvolvimento de modelos e aplicações de aprendizado de máquina (continuação); **Aula 15:** Resumo e encerramento da disciplina.
AVALIAÇÃO	Os alunos serão avaliados pela participação e apresentação em seminários da disciplina (aulas 4, 5, 9, 10), além de um relatório final (de até 6 páginas de texto, imagens, tabelas e gráficos, excetuadas as referências bibliográficas), contendo sua proposta de itens para uma agenda de pesquisa, não necessariamente pessoal, sobre *Explainable AI*. A disciplina inclui 1h por semana de estudos individuais sem horário fixo para os alunos se dedicarem ao estudo da bibliografia indicada.
BIBLIOGRAFIA PRINCIPAL	1) Alan Dix, **Neural Networks and Pattern Recognition in Human-computer Interaction**, pp. 429-451, Upper Saddle River, NJ, USA. Ellis Horwood. 1992. 2) Victoria Bellotti and Keith Edwards, **Intelligibility and Accountability: Human Considerations in Context-Aware Systems,** *Human-Computer Interaction*, vol. 16, no. 2-4, pp. 193-212, Taylor & Francis. 2001. 3) Federica Russo and Jon Williamson, **Interpreting Causality in the Health Sciences,** *International Studies in the Philosophy of Science*, vol. 21, no. 2, pp. 157-170, 2007. 4) Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan and Jonathan Herlocker, **Interacting meaningfully with machine learning systems: Three experiments,** *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 639-662, 2009. 5) Jonathan Grudin, **AI and HCI: Two fields divided by a common focus,** *AI Magazine*, vol. 30, no. 4, pp. 48, 2009.

6) Jon Kolko, **Abductive Thinking and Sensemaking: The Drivers of Design Synthesis,** *Design Issues*, vol. 26, no. 1, pp. 15-28, 2010.

7) Lorenzo Casini, Phyllis McKay Illari, Federica Russo and Jon Williamson, **Models for Prediction, Explanation and Control: Recursive Bayesian Networks,** *Theoria. Revista de Teoría, Historia y Fundamentos de la Ciencia*, vol. 26, no. 1, pp. 5-33, January. 2011.

8) Illari, Phyllis and Russo, Federica. *Causality: Philosophical Theory meets Scientific Practice*. Oxford University Press, Oxfprd. Oxford University Press. 2014

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

10) Zachary C. Lipton, **The mythos of model interpretability,** *arXiv:1606.03490*, 2016.

11) Lucian Leahu, **Ontological Surprises: A Relational Perspective on Machine Learning,** in *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*, pp. 182-186, New York, NY, USA. ACM. 2016.

12) C. Hill, R. Bellamy, T. Erickson and M. Burnett, **Trials and tribulations of developers of intelligent systems: A field study,** in *2016 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pp. 162-170, September. 2016.

13) Sandra Wachter, Brent Mittelstadt and Luciano Floridi, **Why a right to explanation of automated decision-making does not exist in the general data protection regulation,** *International Data Privacy Law*, vol. 7, no. 2, pp. 76-99, 2017.

14) Andrew D. Selbst and Julia Powles, **Meaningful information and the right to explanation,** *International Data Privacy Law*, vol. 7, no. 4, pp. 233-242, 2017.

15) Tim Miller, **Explanation in Artificial Intelligence: Insights from the Social Sciences,** *arXiv preprint arXiv:1706.07269*, 2017.

16) Lilian Edwards and Michael Veale, **Slave to the Algorithm: Why a Right to an Explanation Is Probably Not the Remedy You Are Looking for,** *Duke Law and Technology Review*, vol. 16, pp. 1-65, 2017.

17) Maja Brkan, **AI-supported Decision-making Under the General Data Protection Regulation,** in *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law*, pp. 3-8, New York, NY, USA. ACM. 2017.

18) Reuben Binns, **Algorithmic Accountability and Public Reason,** *Philosophy & Technology*, May. 2017.

19) Sean Penney, Jonathan Dodge, Claudia Hilderbrand, Andrew Anderson, Logan Simpson and Margaret Burnett, **Toward Foraging for Understanding of StarCraft Agents: An Empirical Study,** in *23rd International Conference on Intelligent User Interfaces*, pp. 225-237, New York, NY, USA. ACM. 2018.

20) Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect*. Basic Books., New York, NY. Basic Books.. 2018.

	21) James Larus, Chris Hankin, Siri Granum Carson, Markus Christen, Silvia Crafa, Oliver Grau, Claude Kirchner, Bran Knowles, Andrew McGettrick, Damian Andrew Tamburri and Hannes Werthner, **When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making**, New York, NY, USA. ACM. 2018. Technical Report: Informatics Europe & EUACM 2018 (20 p.). 22) Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim and Mohan Kankanhalli, **Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda,** in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 582:1-582:18, New York, NY, USA. ACM. 2018.
BIBLIOGRAFIA COMPLEMEN-TAR	1) Jaime Nubiola, **Abduction or the Logic of Surprise**, *Semiotica*, vol. 2005, no. 153-1/4, pp. 117-130, 2005. 2) Lorenzo Magnani, **An Abductive Theory of Scientific Reasoning**, *Semiotica*, vol. 153, no. 1/4, pp. 261-286, 2005. 3) Maria Eunice Quilici Gonzalez and Willem Ferdinand Gerardus Haselager, **Creativity: Surprise and Abductive Reasoning**, *Semiotica*, vol. 153, no. 1/4, pp. 325-342, January. 2005 4) T. Kulesza, S. Stumpf, M. Burnett, W. K. Wong, Y. Riche, T. Moore, I. Oberst, A. Shinsel and K. McIntosh, **Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs**, in 2010 IEEE Symposium on Visual Languages and Human-Centric Computing, pp. 41-48, September. 2010 5) P. J. G. Lisboa, **Interpretability in Machine Learning – Principles and Practice**, in Fuzzy Logic and Applications: 10th International Workshop, WILF 2013, Genoa, Italy, November 19-22, 2013. Proceedings, pp. 15-21, Cham. Springer International Publishing. 2013. 6) Todd Kulesza, Margaret Burnett, Weng-Keen Wong and Simone Stumpf, **Principles of Explanatory Debugging to Personalize Interactive Machine Learning**, in Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 126-137, New York, NY, USA. ACM. 2015. 7) Robert Folger and Christopher Stein, **Abduction 101: Reasoning processes to aid discovery**, *Human Resource Management Review*, vol. 27, no. 2, pp. 306-315, 2017. 8) Simone Stumpf, Simonas Skrebe, Graeme Aymer and Julie Hobson, **Explaining smart heating systems to discourage fiddling with optimized behavior**, in IUI 2018 Workshop on Explainable Smart Systems (ExSS) - March 7-11, 2018, pp. 1-4, Tokyo, Japan. Online Publication at: http://explainablesystems.comp.nus.edu.sg/. 2018 9) Jonathan Dodge, Sean Penney, Andrew Anderson and Margaret Burnett, **What should be in an XAI explanation? What IFT reveals**, in IUI 2018 Workshop on Explainable Smart Systems (ExSS) - March 7-11, 2018, pp. 1-4, Tokyo, Japan. Online Publication at: http://explainablesystems.comp.nus.edu.sg/. 2018.

	10) Jonathan Dodge, Sean Penney, Claudia Hilderbrand, Andrew Anderson and Margaret Burnett, **How the Experts Do It: Assessing and Explaining Agent Behaviors in Real-Time Strategy Games**, in Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 562:1-562:12, New York, NY, USA. ACM. 2018.