Dezembro 2022/23

ISCTE IUL
Instituto Universitário de Lisboa

DESCONTINUIDADE ENTRE CIÊNCIA DE DADOS E ÉTICA

Escrita de Textos Técnicos e Científicos

Docente:
Professora Doutora Sílvia Rodrigues Cavalinhos

Licenciatura em Ciência de Dados
Ano curricular: 2º
Turma: CDA1

Marco Delgado Esperança nº 110451

Índice

Resumo

O objetivo central do presente texto é oferecer uma perspetiva sobre ética em ciência de dados.

O argumento central proposto é a necessidade de existência de normas regulatórias éticas mais elaboradas e exigentes.

A metodologia usada é mista, pois recorre-se à análise do número de artigos revistos por pares até 2014 no Google Scholar para mostrar a pouca preocupação com o tema, assim como uma análise qualitativa baseada nos problemas éticos existentes atualmente em ciência de dados.

Ao fim do texto, a partir da análise anteriormente mencionada, estabelece-se uma proposta de uma *framework* de ética.

Palavras-chave: dados, danos, ética, *framework*, privacidade

Introdução

De acordo com o relatório *The Future of Jobs 2020* (World Economic Forum, 2020), as profissões relativas a especialistas em *machine learning* e inteligência artificial irão ter cerca de mais 97 milhões de novos empregos até ao ano de 2025. Esta tendência é confirmada pelo Linkedin que refere que de 2020 para 2021, a procura de profissionais do setor cresceu 33%. Aliado a este aumento de procura de profissionais de análise de dados, surge também um crescimento da responsabilidade de cada um dos cientistas de dados, assim como das responsabilidades éticas das suas decisões, não enviesando-as ou tomando decisões com conclusões erradas. Contudo, por se tratar de uma área muito recente, as normas regulatórias relativamente à ética nesta área ainda se encontram muito verdes, visível pelo facto de até 2014, de acordo com o Google Scholar terem sido publicados apenas 17 artigos revistos por pares (Saltz, 2019).

Assim, as questões éticas em ciência de dados precisam de ser consolidadas de forma a dar maior confiabilidade e imparcialidade às decisões tomadas pelos cientistas de dados. O autor procurará enquadrar o trabalho na dimensão da proteção social do trabalho digno, na área de Ciência de Dados.

1. Descontinuidade entre ciência de dados e regulamentação ética

Um dos problemas resultantes da ética em ciência de dados é a grande descontinuidade entre os exercícios da ciência de dados e os instrumentos da regulamentação ética, não havendo um padrão ético tão rigoroso como a existente na investigação biomédica em humanos (Metcalf & Crawford, 2016).

A regulamentação da prática de investigação isenta os projetos que façam usos das bases de dados preexistentes disponíveis ao público, com base no argumento de que representam riscos mínimos para os sujeitos humanos. Contudo, este princípio revela-se errado, já que as bases de dados disponíveis ao público podem ser objeto de usos secundários e ainda combinados com outras bases de dados para diferentes fins, incorrendo em riscos indeterminados para os sujeitos em geral (Metcalf & Crawford, 2016).

2. Análise do número de *papers* publicados sobre ética em ciência de dados

De acordo com Saltaz (2019), a maioria dos *papers* identificados foram publicados muito recentemente, sendo que apenas oito dos 80 artigos identificados foram publicados antes de 2014. Este facto não é surpreendente pois coincide amplamente com o uso crescente da ciência de dados em uma variedade de contextos.

Como se pode observar na figura que se segue, a maior concentração de artigos foi publicada em periódicos / conferências de tecnologias de informação, onde foram publicados 17 artigos relevantes. Como o foco de domínio não era a literatura cinzenta, a análise baseou-se em artigos académicos revistos por pares.

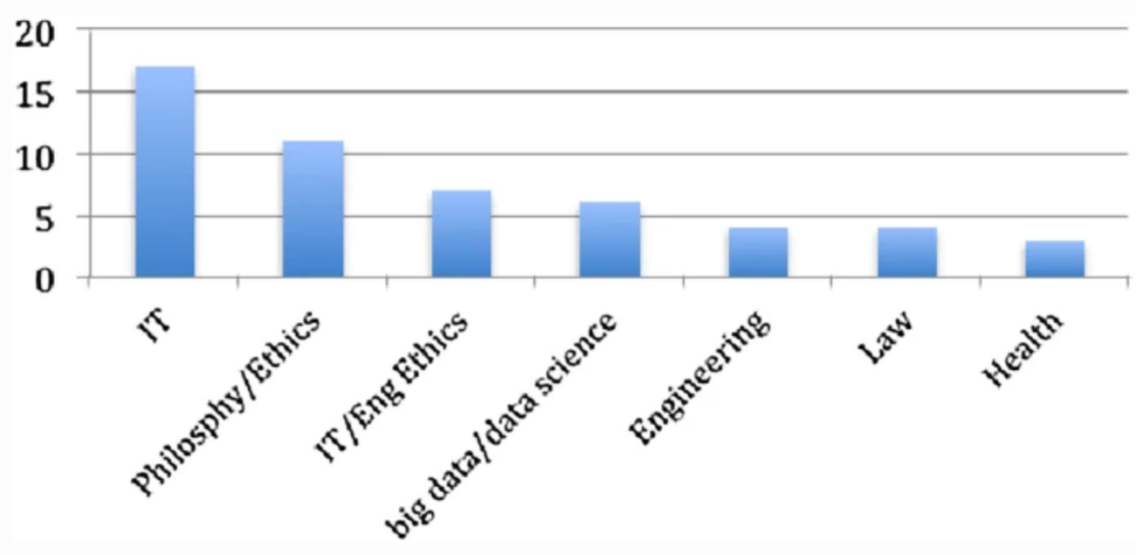

Figura 1 - Número de artigos por foco em journal.

3. Área muito recente

Um dos aspetos que a marca a ciência de dados deve-se ao facto de ser uma área muito recente. Especificamente, como o campo é novo, muitas normas e regulamentos éticos podem ainda não ter sido explorados ou definidos (Metcalf et al., 2016; Sweeney, 2013). Este ponto torna-se ainda mais complicado pelo facto de que a ética e a regulamentação tendem a atrasar as melhorias tecnológicas (Zwitter, 2014), sendo muitas vezes vistas como uma barreira ao desenvolvimento tecnológico e o facto de que a ciência de dados pode introduzir novas classes de risco para uma organização (Tiell & Metcalf, 2016). Em geral, pelo menos em parte devido à novidade desta área, acreditava-se que seria difícil prever todas as possíveis questões éticas relevantes (Tractenberg et al., 2015). Portanto, em um campo emergente como a ciência de dados, pode haver falta de clareza regulatória/legal para determinadas situações. Também pode haver implicações éticas que não foram previamente consideradas por outros ou mesmo destacadas como um potencial dilema ético.

4. A necessidade de uma framework de ética

A criação de uma *framework* de ética pode ajudar a clarificar o vocabulário necessário para discussão de problemas relacionados com ética em ciência de dados (Voronova & Kazantsev, 2015; Tractenberg et al., 2015). Além disso, pode permitir que as equipas de ciência de dados abordem o impacto ético e as implicações da ciência de dados e suas aplicações, usando uma abordagem consistente, holística e inclusiva (Tractenberg et al., 2015).

De forma a desenvolver o código de ética existente, muitos observaram que não havia nada disponível que abrangesse totalmente o que é necessário (Stoyanovich et al., 2017; Leonelli, 2016; Kazantsev, 2015), e também foi observado que o uso de um código de ética mais geral careceria da especificidade para ser útil (Stoyanovich et al., 2017). No entanto, outros defendem que uma estrutura geral encoraja o pensamento crítico e a reflexão ética, sendo que esta estrutura geral pode ajudar acerca das responsabilidades e obrigações das pessoas encarregues dos processos, estratégias e políticas de ciência de dados (Leonelli, 2016; Floridi & Taddeo, 2016).

5. Problemas éticos relacionado com a gestão de dados

A gestão de dados levanta grandes desafios éticos que podem surgir pela recolha e uso de dados. Se por um lado a análise de grandes volumes de dados permite prever eventos futuros com base em tendências passadas, por outro lado, os cientistas de dados integram várias fontes distintas para gerar novos *insights,* criando várias questões éticas que podem ser consideradas potenciais problemas na cadeia de fornecimento de dados (Martin, 2015).

5.1. Privacidade e anonimato

A privacidade e anonimato constitui um dos principais problemas do tratamento de dados. Os indivíduos têm o direito de escolher que atividades e factos que pretendem partilhar com outras pessoas. Na era digital, isso inclui o que o indivíduo escolhe publicar e a sua

capacidade de controlar com quem pretende compartilhar os seus dados. Assim, surge a questão da privacidade e do anonimato.

As questões de privacidade dizem respeito ao controlo do acesso aos dados, enquanto a propriedade diz respeito não apenas a quem possui os dados, mas também quais direitos podem ser transferidos e quais obrigações a recolha ou receção de tais dados implica (Mateosian, 2013; Wielki, 2015).

5.2. Uso indevido dos dados

Além disso, ao fazer a recolha e tratamento dos dados temos de garantir que eles não são indevidamente usados, sendo que em muitos sites não é permitido de todo usar dados de navegação. O armazenamento de dados em grande volume (*Big Data*) introduziu mudanças ao nível da recolha, uso, retenção e acesso aos dados pessoais. Infelizmente, esses dados muitas vezes são utilizados para fins diferentes do objetivo pretendido, sendo que muitos utilizadores consideraram tais práticas uma violação do seu direito à privacidade (Pascalev, 2017). Esta problemática está muitas vezes associada à criação de perfis a partir da conjugação de fontes diversas de dados sem autorização explícita dos proprietários dos dados.

A utilização de dados por um serviço requer o consentimento de uma política de privacidade, no entanto, muitos utilizadores não conseguem ler e entender essas políticas, o que leva a que não haja um consentimento real. Muitas vezes, a lista de termos e usos de privacidade é extensa, sendo que para muitas aplicações a rejeição desses termos não é possível para usufruir dela. No entanto, existem algumas sugestões de alto nível que as organizações podem seguir, como apropriar-se das suas fontes de dados, não entrar em acordos de confidencialidade que impeçam a explicação de quem são seus parceiros de dados e tornar a cadeia de fornecimento de dados visível para que uma organização tenha a capacidade de garantir que os dados não são indevidamente usados (Martin, 2015).

5.3. Precisão e validação dos dados

Outro aspeto fundamental que marca a ciência de dados é a precisão e validação dos dados. Um cientista de dados necessita de garantir *fitness of purpose*, isto é, que os dados são usados para os fins que tinham sido estabelecidos.

Um exemplo concreto da precisão e validação dos dados diz respeito à avaliação dos professores. Um número crescente de organizações usa pontuações de testes padronizados dos alunos de um professor para desenvolver pontuações de desempenho dos professores. Muitos questionam a precisão de uma pontuação de teste de um único aluno como entrada nesse modelo. Foi observado que "quando qualquer aluno faz um teste de matemática, em qualquer dia, há uma enorme incerteza em torno dessa pontuação. Pode ser que o estudante tenha tido sorte este ano e acertou duas ou três perguntas ou poderia na manhã do teste não estar a sentir-se bem. "Consequentemente, a pontuação em qualquer dia não é necessariamente um bom reflexo do nível de desempenho de uma criança" (Butrymowicz & Garland, 2012). Portanto, alguns argumentam que, embora a base de dados tenha as pontuações corretas armazenadas, os dados de um teste não são precisos e não devem ser usados como uma entrada chave para o modelo (Butrymowicz & Garland, 2012).

6. Framework proposta

Face aos problemas apresentados, o autor concorda com a proposta de Saltz, 2019 sobre as considerações éticas a ter em conta para cada fase de um projeto de ciência de dados, que se apresentam a seguir:

Project phase	Key ethical themes	Ethical considerations
Business understanding	Project initiation/management challenges	Personal and group harm
		Team accountability
Data understanding/data preparation	Data challenges	Data misuse
		Data privacy & anonymity
		Data accuracy
Modeling	Model challenges	Personal and group harm
Evaluation		Subjective model design
Deployment		Misuse/misinterpretation

Figura 2 - Considerações éticas a ter em conta em cada uma das fases de um projeto de ciência de dados.

Conclusão

Em síntese, ainda existem muitos passos a serem dados para a consolidação das questões éticas em ciência de dados.

O facto de ser uma área muito recente, assim como os problemas gerados pela existência de um volume de dados cada vez maior numa era digital, exigem que sejam aperfeiçoadas as normas éticas existentes, como também a elaboração de uma *framework* de ética, bem definida e estruturada que auxilie a identificar quais as considerações éticas a ter em conta em cada fase dos projetos de ciência de dados.

Desta forma, poderemos contribuir para decisões mais independentes de influências externas, assim como aumentar a confiabilidade das decisões tomadas pelos cientistas de dados.

Referências Bibliográficas

Butrymowicz, S., & Garland, S. (2012). *How New York city's value-added model compares to what other districts, states are doing, hechingerreport*. http://hechingerreport.org/content/how-new-york-citys-value-added-model-compares-to-what-other-districts-states-are-doing_7757/

Crawford, K., & Metcalf , J. (2016). Where are human subjects in Big Data research? The emerging ethics divide. *SAGE Journals*. https://journals.sagepub.com/doi/full/10.1177/2053951716650211

Drosou, M., Jagadish, H. V., Pitoura, E., & Stoyanovich, J. (2017). Diversity in big data: A review. *Big data*, 5(2), 73–84.

Farfield, Joshua, & Shtein (2014). Big Data, Big Problems: Emerging Issues in the Ethics of Data Science and Journalism. *Mass Media Ethics*, 29. https://www.tandfonline.com/doi/citedby/10.1080/08900523.2014.863126?scroll=top&needAccess=true&role=tab

Floridi, L., & Taddeo, M. (2016). What is data ethics?. *Philosophical Transactions Series A,* 374, 2083.

Forum, W. E. (Ed.). (s.d.). *The Future of Jobs Report*. (October 2020). Genebra, Suíça. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf

Leonelli, S. (2016). Locating ethics in data science: Responsibility and accountability in global and distributed knowledge production systems. *Philosophical Transactions of the Royal Society* A, 374(2083), 20160122.

Kosinski, M., Stillwell, D. and Graepel, T. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15): 5802–5805.

Martin, K. E. (2015). Ethical issues in the big data industry. *MIS Quarterly Executive*, 14, 2.

Mateosian, R. (2013). Ethics of big data. *IEEE Micro*, 33(2), 60–61.

Pascalev, M. (2017). Privacy exchanges: Restoring consent in privacy self-management. *Ethics and Information Technology*, 19(1), 39–48. https://doi.org/10.1007/s10676-016-9410-4.

Reid, E. 1996. Informed consent in the study of on-line communities: A reflection on the effects of computer-mediated social research. *The Informational Society Journal*, 12: 169

Saltz, J.S., Dewar, N (2019). Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics Inf Technol* 21, 197–208. https://doi.org/10.1007/s10676-019-09502-5

Stoyanovich, J., Howe, B., Abiteboul, S., Miklau, G., Sahuguet, A., & Weikum, G. (2017). Fides: Towards a platform for responsible data science. *SSDBM'17-29th International Conference on Scientific and Statistical Database Management*.

Sweeney, L. (2013). Discrimination in Online Ad Delivery. ACM Queue 11(3). *Association of Computing Machinery*.

Tiell, S., & Metcalf, J. (2016). The Universal Principles of Data Science Ethics. Accenture Labs. https://www.accenture.com/t20160629T012639__w__/us-en/_acnmedia/PDF-24/Accenture-Universal-Principles-Data-Ethics.pdf

Tractenberg, R. E., Russell, A. J., Morgan, G. J., FitzGerald, K. T., Collmann, J., Vinsel, L., … Dolling, L. M. (2015). Using ethical reasoning to amplify the reach and resonance of professional codes of conduct in training big data scientists. *Science and Engineering Ethics*, 21(6), 1485–1507.

Voronova, L., & Kazantsev, N. (2015). The ethics of big data: Analytical survey. *Business informatics (CBI), 2015 IEEE 17th conference on* (Vol. 2, pp. 57–63). IEEE.

Wielki, J. (2015). The social and ethical challenges connected with the big data phenomenon. *Polish Journal of Management Studies*, 11(2), 192–202.

Zwitter, A. (2014). Big data ethics. *Big Data & Society*, 1(2), 2053951714559253.