ICS 35.030
CCS L80

NATIONAL STANDARD OF THE PEOPLE'S REPUBLIC OF CHINA

GB/T 42888-2023

Information security technology - Assessment specification for security of machine learning algorithms

信息安全技术 机器学习算法安全评估规范

Issued on: August 6, 2023 **Implemented on: March 1, 2024**

Issued by: State Administration for Market Regulation;
Standardization Administration of the PRC

Table of Contents

Foreword

This document complies with the provisions of GB/T 1.1-2020 "Standardization Work Guidelines Part

1.Structure and Drafting Rules of Standardization Documents" Drafting. Please note that some content in this document may be subject to patents. The publisher of this document assumes no responsibility for identifying patents. This document is proposed and coordinated by the National Information Security Standardization Technical Committee (SAC/TC260). This document was drafted by: Beijing Saixi Technology Development Co., Ltd., Institute of Computing Technology, Chinese Academy of Sciences, Tsinghua University, National Planning Computer Network Emergency Technology Coordination Center, Shanghai SenseTime Intelligent Technology Co., Ltd., Beijing Ruilai Intelligent Technology Co., Ltd., Alibaba (China) Co., Ltd., Institute of Information Engineering,

Chinese Academy of Sciences, China Academy of Information and Communications Technology, China Electronics Technology Group Corporation Fifteenth Research Institute Institute, National Information Technology Security Research Center, Guangzhou University, Peking University, East China Normal University, Beihang University, Huawei Technologies Co., Ltd. Co., Ltd., Beijing Megvii Technology Co., Ltd., Beijing Baidu Network Technology Co., Ltd., Shenzhen Tencent Computer Systems Co., Ltd., Zhejiang University School, Beijing Qihu Technology Co., Ltd., Beijing Xiaoju Technology Co., Ltd., Anhui Engineering University, Beijing Zhizhi Tianxia Technology Co., Ltd., Beijing Jiaotong University, Zhejiang University of Technology, Shanghai Industrial Control Safety Innovation Technology Co., Ltd., People's Public Security University of China, Shenzhen Big Data Research Academy, Beijing Institute of Computer Technology and Applications, Institute of Automation, Chinese Academy of Sciences, Shanghai Suiyuan Technology Co., Ltd., Fengtai Technology (Beijing) Co., Ltd., China Electronics Technology Standardization Institute. The main drafters of this document. Shangguan Xiaoli, Hao Chunliang, Xu Xiaogeng, Hu Ying, Chen Zhong, Shen Huawei, Jiang Hui, Mei Jingqing, Zhang Yuguang, Peng Juntao, Guo Yan, Li Pengxiao, Ai Zhengyang, Zhao Yunwei, Han Han, Liu Ming, Yin Zhiyi, Pang Liang, Wang Xiaoshi, Liu Zongzhen, Zhou Xi, Meng Guozhu, Jing Huiyun, Zhang Linlin, Zhu Chunchao, Huo Shanshan, Liu Jian, Liu He, Su Hang, Jin Tao, Liu Jiqiang, Ren Kui, Zhang Xudong, Cheng Jin, Zhu Hongru, Yang Tao, Li Qin, Liu Xianglong, Wang Yifei, Wu Geng, He Ran, Gu Zhaoquan, Li Shi, Cao Xiaoqi, Yan Minrui, Fu Yingbo, Guo Ying, Sun Air Force, Tang Jiayu, Liu Xize, Wang Zhelin, Ren Lu, Xu Yongtai, Zhang Yi, Qin Zhan, An Zeliang, Xu Yuqing, Li Xue, Li Dahai, Xu Guangxia, Bao Shenfu, Guo Jianling, Xuan Qi, Zhang Shitian, Zhao Yongxin, Wang Jiao, Wang Bingzheng, Lu Tianliang, Wu Baoyuan, Han Lei, Zhang Yutong, Peng Quan. Information security technology Machine learning algorithm security assessment specifications

1 Scope

GB/T 42888-2023 specifies how to assess the security of a machine learning algorithm, which is a different question from whether the system around it is secure. A model has attack surfaces of its own: adversarial examples that cause misclassification from perturbations a person cannot see, poisoning of the training data so the model learns a backdoor, extraction attacks that reconstruct a proprietary model by querying it, and inference attacks that recover whether a particular record was in the training set. None of these are software vulnerabilities in the usual sense - the code is correct and the model behaves as trained - so conventional security testing does not find them. This document specifies the security requirements and assessment methods for machine learning algorithm technologies and services, and the process by which such an assessment is carried out. It is intended to guide the development, deployment and evaluation of machine learning systems where security matters. Under ICS 35.030 and CCS L80, it is written for AI developers and the organisations deploying their models, for security assessment

bodies, and for regulators of algorithmic systems.

This document specifies the security requirements and assessment methods for machine learning algorithm technologies and services, as well as the machine learning algorithm security assessment process. This document is suitable for guiding machine learning algorithm providers to ensure the security of the machine learning algorithm life cycle and carry out machine learning algorithm security. A comprehensive assessment can also provide reference for regulatory assessment.

2 Normative reference documents

This document has no normative references.

3 Terms and definitions

The following terms and definitions apply to this document.

3.1 Algorithms that functional units improve their performance by learning new knowledge and skills or sorting out existing knowledge and skills.

3.2 Organizations that utilize machine learning algorithms to perform specific functions.

Note. This document is referred to as algorithm provider, including algorithm technology providers and algorithm service providers. Algorithm technology provider refers to the development and Provider, algorithm service provider refers to a service provider that uses applied algorithm technology.

3.3 Services that apply algorithm recommendation technology to provide information. Note

1.Application algorithm recommendation technology refers to the use of machine learning algorithms to generate synthetic categories, personalized push categories, sorting and selection categories, retrieval and filtering categories, and scheduling decisions. Categorization and other algorithmic technologies are used to provide information to users. Note

2.This document refers to algorithms such as generation and synthesis, personalized push, sorting and selection, retrieval and filtering, and scheduling and decision-making as five categories of algorithms.

3.4 algorithm lifecyclealgorithmlifecycle The evolution of machine learning algorithms from design to retirement. Note

1.The algorithm life cycle includes design and development, verification and confirmation, deployment and operation, maintenance and upgrade, and decommissioning. Note

2.General algorithm services are in the deployment and running stage.