USA und KI: Neues Artificial Intelligence Risk Management Framework des NIST

Dr. Axel Spies ist Rechtsanwalt in der Kanzlei Morgan Lewis & Bockius in Washington DC und Mitherausgeber der MMR.

Das National Institute for Standards and Technology (NIST) in den USA hat sein Artificial Intelligence Risk Management Framework (AI RMF) veröffentlicht (48 Seiten). Dessen Vorgaben sollen den verschiedenen Akteuren im Bereich der künstlichen Intelligenz, wie Organisationen und Einzelpersonen helfen, die Risiken von KI zu bewältigen. Das AI RMF ist in zwei Teile gegliedert.

1. Erster Teil der AI RMF

Teil 1 enthält Richtlinien, wie KI-Akteure die potenziellen Risiken von KI erfassen können. Er konzentriert sich auf vier Maßnahmen: Steuern, Abbilden, Messen und Verwalten. Es behandelt in dem Zusammenhang drei Kategorien von potenziellen Schäden, die KI-Akteure beim Einsatz von KI-Systemen berücksichtigen müssen:

> Schäden für den Menschen: Verletzung individueller Freiheiten oder Bedrohung der physischen, psychischen oder wirtschaftlichen Sicherheit und Chancen; Diskriminierung von Personengruppen und Schäden für den gesellschaftlichen Zugang zu Demokratie und Bildung.
> Schäden für eine Organisation: Unterbrechung des Geschäftsbetriebs, mögliche Sicherheitsverletzungen und Rufschädigung.
> Systemschäden („Ecosystem"): Unterbrechung der globalen Finanz- oder Lieferkettensysteme und Schädigung der Umwelt und der natürlichen Ressourcen.

Im AI RMF wird genauer dargestellt, welche Eigenschaften ein vertrauenswürdiges KI-System erfüllen muss. Dazu gehört nicht nur der Schutz der Privatsphäre, sondern auch dass das System transparent und nachvollziehbar arbeitet, sowie diskriminierungsfrei und gegen schädliche Verzerrungen gewappnet ist.

NIST betont, dass KI-Akteure die Effektivität des KI-Risikomanagements regelmäßig bewerten sollten, um ihre Fähigkeiten zum Umgang mit KI-Risiken zu verbessern.

2. Zweiter Teil des AI RMF

Teil 2 des AI RMF beschreibt die vier Maßnahmen Steuern, Abbilden, Messen und Verwalten als Kern des AI RMF. Jede Kernfunktion ist in zahlreiche Kategorien und Unterkategorien unterteilt, wobei jede dieser Kategorien bestimmte Maßnahmen enthält, die während des gesamten Lebenszyklus eines KI-Systems kontinuierlich durchgeführt werden sollten. Das NIST hat zu diesem Zweck ein begleitendes AI RMF „Playbook" als Handbuch veröffentlicht, das Unternehmen dabei helfen soll, das RMF anzuwenden und ihre Ziele durch vorgeschlagene Maßnahmen zu erreichen, die Unternehmen an ihren eigenen Kontext anpassen können.

Das AI RMF ist skeptisch, was den Einsatz von Privacy Enhanced Technologies (PETs) betrifft: „PETs für KI sowie Methoden zur Datenminimierung, wie z. B. De-Identifizierung und Aggregation für bestimmte Modell-Outputs, können die Entwicklung von KI-Systemen mit erhöhtem Datenschutz unterstützen. Unter bestimmten Bedingungen, z. B. bei spärlichen Daten, können Techniken zur Verbesserung der Privatsphäre zu einem Verlust an Genauigkeit führen, was sich auf Entscheidungen über Fairness und andere Werte in bestimmten Bereichen auswirkt." (S. 17).

3. Fazit

Das AI RMF und das dazugehörige Handbuch bieten Leitlinien, Maßnahmen und Ergebnisse, die Einzelpersonen und Organisationen bei der Entwicklung von und dem Umgang mit KI-Systemen umsetzen können. Deren Einhaltung ist zwar zurzeit noch freiwillig, aber Organisationen, die das AI RMF anwenden, sind dann sicherlich besser auf die einzigartigen und oft unvorhergesehenen Risiken vorbereitet, die KI-Systeme darstellen.