像血液,像食物,像罗盘,像空气

数据在现代社会中流动,像血液在人体内循环,为各种系统提供动力。

数据还是人工智能和机器学习的“食物”。需要大量的数据来“消化”,机器才能更好地服务于人类。

在商业、医疗、教育等领域,数据又像罗盘,帮助我们确定方向。

总之,数据已成为现代社会不可或缺的一部分,就像空气一样,至关重要。

所以,数据污染就如同空气污染、食物污染一样,是一个不容忽视的、事关安全的重要问题。设想一下:一辆无人驾驶汽车行驶到一个红绿灯路口,正常情况下应该“红灯停、绿灯行”,但如果车辆学习到的相关数据被改变,逻辑设定为“红灯行、绿灯停”,就会造成交通混乱,甚至车祸。

还有公司的账目、工厂的生产信息、网上银行的数据等,都有可能成为黑客的目标被污染,进而造成恶劣的后果。

随着生成式人工智能技术的快速发展,由AI生成的图片、新闻、视频等内容在网络世界泛滥。这些由AI生成的内容不少缺乏真实性,可能引发谣言的广泛传播,对社会造成不良影响。

本期《读+》记者采访的数据治理专家石秀峰告诉我们:数据已经成为核心资产,数据治理是企业数字化转型的必经之路,数据治理,是一个难题也是一个迫切需要解决的问题。 王永芳

访谈

数据治理专家石秀峰:

数据治理是数字化转型必经之路

□记者马梦娅

随着互联网、物联网、人工智能等技术的不断发展,数据量呈现爆发式增长的态势。海量的数据为人们提供了更多的信息和机会,但也给数据治理带来了巨大的压力。无论是传统制造业的智能化转型,还是新兴数字经济产业的发展,都离不开准确、可靠的数据支持。在庞大的数据量中,如何确保数据的真实性、完整性和准确性,是一个亟待解决的问题。

近日,用友集团数据咨询专家,全数联(DGBOK)数据治理专家石秀峰及其团队出版新作《一本书讲透数据治理:战略、方法、工具与实践》,书中强调“用数据思考,用数据说话,用数据决策”。长江日报《读+》记者专访石秀峰,他指出,在当下的时代背景下,数据治理的迫切性日益突出。只有加强数据治理,才能确保数据的质量和安全,充分发挥数据的价值,推动经济社会的健康发展。

石秀峰

数据发展影响经济与生活

从大学毕业开始,石秀峰就在做与数据相关的工作。他最早做数据仓库项目,后来做数据治理,包括做数据系统,再往后是围绕数据中台开展工作……他深切地感受到数据对我们的生活产生的影响越来越重要。

2015年,全国第一家数据流通交易场所——贵阳大数据交易所成立。在石秀峰的印象里,当时这是一个比较新鲜的概念。截至2022年7月30日,由地方政府发起、主导或批复的数据交易所已有39家,具体的数量动态还在不断变化中。几乎每个省都有数据交易中心,有些市级城市也有数据交易中心或是数据资产登记所。从这些发展变化可以看出来,整个数据发展市场是非常迅猛的。

石秀峰表示,企业做一项业务的时候,之前历史数据的一些分析挖掘是可以对当前的业务决策起到一定指导作用,甚至是控制作用的——举例来说,采购业务,选择一个供应商,突然发现之前供应商的一些信息,包括一些付款质量情况不是特别好,平台系统就会做出及时的响应或者预警,数据的管理规避了采购的风险。

目前,很多企业对数据的有效治理,是可以直接带来经济效益的。

我们在网上看到的和“数据”相关的所有新闻,都属于数据治理的范畴,数据治理只是把你日常做的事进行了理论升华,便于把一个行业的数据管理方法,推广到所有行业,形成标准。

在石秀峰和他的团队所著的《一本书讲透数据治理:战略、方法、工具与实践》书中,清晰地解释了“数据治理、数据战略、数据中台、元数据、主数据”这些术语。石秀峰说,如果你管理着公司里的几个数据表或者数据库,并定期对这些数据表进行整理、规范、关联、扩展等,那你日常做的事,就和数据治理相关性很大。

数据清污如“下水道工程”

生成式人工智能的普及,使数据规模以摩尔定律的速度快速增加,互联网上每日新增的图片、语料等内容中AI生成内容的占比已经远远超过了真实的、人类生产的内容。随之而来的是数据质量和数据污染问题,诸如深度伪造、偏见和有害内容生成、有效数据“稀释”等。大量低质量或不客观的数据充斥其中,不仅对模型的训练产生负面影响,甚至可能导致错误的决策和偏差。这种数据质量问题,引发了人们对“数据污染”的担忧。

石秀峰笑称,将污染的数据弄干净再清洗好,相当于一个“下水道工程”,清洗数据,就必然要对那些脏数据、噪声数据进行处理。

面对互联网上新增的海量原始数据,我们亟须构建完善的数据内容维护、数据质量管理和数据安全监管体系。

联合国大学关于AIGC数据利用与风险的一份调研报告中给出了一些建议:全面推进AIGC相关立法,从源头防范数据污染问题;构建全面的AIGC数据质量与安全评价体系,促进合成数据的合理利用;将AIGC数据管理纳入全球AI治理合作,促成全球范围内全面解决合成数据带来的风险;加快明确生成式人工智能伦理与隐私准则,增强生成式人工智能技术的透明性,防范有违全人类共同价值的AIGC内容扩散,并解决用户隐私和数据滥用等问题。

数据治理与我们的生活息息相关

读+:您能阐述一下什么是数据污染,在当今数字化时代,您认为数据污染主要表现在哪些方面?能否举一些具体的例子?

石秀峰:数据污染是一个比较形象的概念,简单来说,就是数据里混进了“脏东西”,这些数据不完整、不准确、不可靠,对业务来讲甚至可能是有害的。我们网上搜索信息,经常看到一些结果都是过时、重复的,甚至是矛盾的内容,这就是低质量数据在捣乱。

数据污染发生在数据的源头,也可能发生在数据传输的过程中。最常见的是数据录入过程中被“污染”,工作人员粗心大意或缺乏专业培训,就可能导致数据录入错误。例如,在医疗系统中,如果医生或护士在记录患者信息时出错,那么患者的医疗记录就会受到污染。这种污染是无意识的,还有一种源头数据污染的情况是“数据造假”,在某些情况下,为了达成某种目的(如提高业绩、骗取补贴等),相关人员可能会故意篡改或伪造数据。

另外,在数据传输过程中,由于网络不稳定或设备故障等原因,也可能导致数据丢失、乱码或重复,从而造成“数据污染”。还拿远程医疗系统举例,试想一下如果网络不稳定,那么患者的医疗图像数据就很可能在传输过程中丢失或失真,这对于现代化的智能诊疗影响是非常大的。随着我们网络技术的不断发展这个问题目前已经基本得到了解决,但在一些偏远的地区,网络不够发达,由于网络问题导致的数据丢失、数据失真、数据延迟等数据污染问题依然存在。

读+:数据污染和我们的生活有怎样的关系?数据污染对个人、企业和社会带来哪些具体的危害?

石秀峰:数据污染与我们的生活息息相关,它像一股无形的“毒流”,悄然影响着我们的日常生活和工作。数据污染不仅危害个人,还有企业的经营,乃至对整个社会都会带来一系列危害。

首先,数据污染严重影响我们的生活。相信大家都知道网络诈骗,其实这个严格来讲跟数据污染有很大的关系。例如:你在某个网站注册了账号,并填写了个人信息。然而,由于该网站的数据保护措施不到位,你的个人信息被泄露给了不法分子。他们可能会利用这些信息对你进行诈骗或骚扰,这就是数据污染带来的隐私泄露问题。再举个例子,现在大家都经常在网上购物,大部分人在网上浏览商品的时候都喜欢看看这个商品的评价如何,然后根据大家的评价决策是否购买。假设你正在考虑从网上购买一件商品,你查看了某电商平台的用户评价,但其中混杂了大量虚假好评或恶意差评,这就会导致你无法准确判断这件商品的真实质量。因此,我们在购物的时候就会买到一些“以次充好、以假充真”的情况。

其次,数据污染影响业务运营和管理决策。数据驱动业务,数据赋能管理,数据要素是当前企业数字化转型的核心。然而,企业决策时,需要的是准确、及时、完整、真实的数据。例如,如果市场数据被污染(如被夸大或缩小),那么企业的市场策略可能会偏离实际,导致决策失误和经济损失。如果企业的产品数据(如质量、性能等)被污染,那么客户在使用产品时可能会遇到各种问题。这些问题会降低客户对企业的信任度,进而影响企业的品牌形象和市场声誉。

数据污染对社会危害也很大。例如:在金融行业,如果金融机构的数据被污染(如被人为篡改或伪造),那么可能会导致金融市场的波动和风险。在公共卫生领域,如果疫情数据被污染(如被隐瞒或夸大),那么可能会导致政府和社会对疫情的误判和误操作……这种由于数据污染带来的灾难有时候是难以想象的。

读+:与过去相比,现在的数据污染有哪些新的特点和趋势?

石秀峰:过去的数据污染是由人为(如粗心、不专业或恶意)、设备故障问题以及数据传输中的网络问题等造成的。

随着生成式人工智能(AIGC)技术的快速发展,由AI生成的图片、新闻、问答和视频等内容在网络世界中泛滥。这些由AI生成的内容虽然视觉效果惊艳,但往往缺乏真实性和可靠性,导致真实与虚假的界限变得模糊不清。一旦这些错误信息被大众误认为是真实的,就可能引发谣言的广泛传播,对社会造成不良影响。

根据有关调查报告,生成虚假文章的网站数量自2023年5月以来激增了1000%以上,涉及的领域包括政治、军事、科技、社会、艺术等多个方面。这些虚假新闻和信息不仅误导公众,还可能引发社会问题和信任危机。例如:在司法领域,伪造的图像、音频或视频可能被用于误导法庭或干扰司法程序。

读+:为什么数据治理在当下变得如此重要?它对于个人、企业和国家分别有着怎样的意义?

石秀峰:有数据污染,自然就需要数据治理,污染问题越大,影响越大,数据治理的重要性和必要性就越强。

对个人来讲,数据治理涉及个人信息的合法使用和保护,它的重要性不言而喻。随着数字安全技术的发展,可以通过加强数据加密、身份验证和授权机制等数据治理技术措施,来保障个人数据免受未经授权的访问、篡改或泄露,保障个人信息安全。当然,我们的数据安全意识也需要不断提升,在不确定的环境下,尽量不要将个人信息泄露给别人或各种网络APP。

对企业而言,数据治理是企业数字化转型的必经之路。高质量的数据是企业进行业务分析、决策制定和运营优化的基础。被污染的数据不仅不能帮助企业产生有价值的数据洞察,还可能对企业决策造成误导。因此,数据治理有助于企业更好地理解和利用数据资产,以及确保使用数据过程的合规性,从而帮助决策者做出更加科学的决策,预防企业经营风险的发生。

对国家而言,数据安全直接关系到国家的经济安全、政治安全和社会稳定。通过数据治理,可以优化数据资源配置,提高数据利用效率,促进数据产业的创新发展,为数字经济的健康发展提供有力支撑。

利用人工智能技术进行“数据清污”

读+:您能否用一些实际案例来说明数据治理的成功带来的积极影响?

石秀峰:由于我一直从事企业级的数据治理,给大家讲一个企业数据治理的案例吧。

我曾经做过一家煤炭企业矿井侧的数据治理。这家煤矿集团企业,在煤炭板块建立了集团数据标准化体系,但聚焦到煤矿板块尚无矿井侧数据采、治、管等能力。这使得企业无法获取到准确、实时的矿井生产数据,各个生产系统之间的核心数据也不一致,导致企业无法准确评估矿井的生产能力和资源状况。

我们首先为这家企业建立统一的数据采集标准和规范,通过数据治理平台,对采集到的矿井数据进行清洗、转换、整合和存储,解决了各个生产系统之间核心数据不一致的问题,增强了数据的一致性和可比性。同时部署先进的传感器和物联网技术,实时监测矿井生产数据,如产量、设备状态、人员位置。

通过数据治理,成功整合了矿井侧的煤炭产业、安全管理、生产辅助等多方面的数据资源,构建了矿井侧数据采集、治理、共享和应用等能力,利用先进的大数据、物联网、人工智能等技术实现了IT数据和OT数据的融合,并围绕企业的生产优化、资源调配和安全管理等核心业务,构建了数据应用场景,帮助企业实现了数字化转型。

读+:在数据治理过程中,您认为最大的挑战是什么?是技术问题、管理问题还是观念问题?

石秀峰:在数据治理过程中,面临的挑战是多元化的,但要说最大的挑战,往往因组织而异。在观念上,许多企业内部缺乏对数据作为重要资产的认识,导致数据质量低下、数据使用不当等问题。另外一个很大的观念问题是:数据治理需要整个企业各部门的支持和参与,但由于有的业务部门可能对数据治理的重要性认识不足,或者担心数据治理会限制其灵活性,往往成为数据治理的阻力。

在管理上,有的企业由于缺乏具备数据治理专业的知识和技能的人才,或者现有人员缺乏必要的培训和支持,导致数据管理的成熟度不够。还有就是数据治理需要跨部门的协作,但往往很多企业缺乏明确的责任分工和协作机制,导致治理工作难以推进。

我想最后说说技术的挑战,之所以将技术放到最后说,并不是说技术不重要,而是在我看来,技术并不是企业数据治理最大的挑战。随着新技术的不断发展和提升,由于技术问题而带来的挑战会迎刃而解。

读+:未来数据治理的发展方向是什么?会有哪些新的技术和方法出现?

石秀峰:我相信,未来越来越多的企业会将数据治理视为战略层面的工作。随着数据重要性的提升,数据治理也将不再局限于IT部门,而是需要各个业务部门以及法务部门、合规部门等多个部门的紧密协作。这是必然的趋势。

在技术方面,AI与传统数据治理技术的结合将越来越紧密。随着数据治理技术和人工智能的不断融合,数据治理将变得更加主动和智能。例如在数据采集方面,利用图像识别、语音识别、自然语言处理等AI技术自动化采集各种半结构化和非结构化的数据。

在数据建模方面,通过知识图谱、机器学习、图数据库等新技术,帮助企业对结构化数据、半结构化数据、非结构化数据进行文本识别、语言识别、全面梳理,自动转为结构化数据,自动化捕获元数据进行数据建模,通过数据解析、结构化建立数据标准和数据关系,让企业暗数据无处藏身。

在元数据管理方面,人工智能技术可以帮助更好地管理和整合元数据,将机器学习和NLP植入元数据管理工具,对以往难以索引的丰富数据类型自动创建高质量的元数据。

在主数据管理方面,利用人工智能技术对数据集进行监控,可以帮助自动鉴别和筛选出主数据,通过监控主数据的数据质量,维护和确保主数据的“黄金记录”。

在数据标准方面,通过机器学习算法可以自动识别出数据标准的使用频度和热度,找出那些没有使用或使用过程中存在问题的数据标准,以便企业对数据标准进行评估和优化。

在数据安全方面,利用人工智能、机器学习技术,可以帮助企业识别、清洗、转换、处理数据集中的敏感数据,例如通过分类、聚类神经网络等算法模型及自然语言处理、智能搜索等技术实现对敏感数据的实时、动态识别,自动化生成标注,自动化分类分级,加强敏感数据的安全防护。

在数据分析方面,将机器学习技术应用到数据建模、数据处理、数据质量等环节,实现自动化数据的清洗与处理,减少人为干预;利用机器学习、人工智能技术将传统的分析模型(如杜邦分析)在大数据环境下进行“锤炼”,形成适合企业且更加智能、可靠的数据模型。

读+:对于普通民众来说,我们在数据治理中可以发挥怎样的作用?应该如何增强公众的数据治理意识?

石秀峰:我认为对于普通民众来说,增强公众的数据治理意识是至关重要的。我们都应了解自己的个人信息如何被收集、使用和共享,并谨慎处理个人信息,避免泄露给不安全的平台或机构。以及学习数据分析和解读的基本技能,以便更好地理解和使用数据,避免被误导或误导他人。

我们可以通过普及数据治理知识、开展教育活动,提升民众数据安全意识,从而有助于构建一个更加安全、透明和高效的数据环境。

本社社址:湖北省武汉市江岸区金桥大道113号新长江传媒大厦 | 邮政编码:430013 | 中闻集团武汉印务有限公司承印 | 印刷质量监督电话:027-85888888