数据生产理论

——数据资源权利配置的基础理论

高富平

【摘　　要】人类正在进入数据驱动的时代，数据成为社会的基础资源，但数据一直被认为处于公共领域并妨碍着数据权利化，困扰着数据资源利用秩序的建立。描述特定对象的数据并不是天然存在的，而是被生产出来的，并将数据价值(预测分析、发现新知)的实现过程界分为原始数据生产(采集)、数据集生产(汇集性处理)和数据分析(分析性处理)三种行为，并将前两个行为称为数据生产，提出数据生产理论。数据生产理论首先应区分数据生产和数据分析，原始数据的生产是建立在分析原材料提供者基础之上，应承认其价值并配置适当权利，以满足各种分析目的的数据集的生产。同时，数据生产还应区别数据来源，来源于个人的数据并不一定是个人生产的，只有个人在提供或创制了数据时才是数据的生产者。因此，数据生产理论是在将数据视为一种资源的情形下为数据上权利配置提供理论支撑，通过配置相应权利，构筑从原始数据生产者到数据集生产者，再到数据分析者的数据利用秩序。

【关 键 词】数据生产；个人数据；数据财产权；数据权利配置

【作者简介】高富平，华东政法大学法律学院教授、法学博士。

【原文出处】《交大法学》(沪)，2019.4.5～19

【基金项目】本文为国家社会科学基金重大项目“大数据时代个人数据保护与数据权利体系研究”(项目编号：18ZDA145)的研究成果。

一、引言

信息和通信技术发展到今天，尤其是随着网络技术的普遍和深度应用，人、物(自然界、机器等)、组织的活动或运行数据被各种网络系统、传感器和智能设备记录下来，形成可以数字化再现世界状态和运行的数据世界。今天，我们不仅可以全息地数字化记录(即数据化)人类所处的物理和社会环境、物或人的轨迹或行为，而且具有处理和分析海量数据所需要的运算能力，这便是大数据技术。[①]大数据分析可以克服传统基于统计学数据分析的局限性，实现对海量、动态和多样化的数据分析，由此人类被认为进入到大数据时代(本文称为数据时代)，开启数据驱动的经济(datadriven economy)，[②]或称为数据文明时代。[③]数据时代(数据文明、数据经济)的标志是数据成为社会基础资源、经济活动的要素，成为比土地、资本、劳动力等更为核心的要素，它被比作“石油”。事实上，现在个人和组织均已经开始重视并利用数据资源，尽可能多地获取和控制数据，并利用各种数据处理工具分析数据(包括人工智能)，应用于科学研究、社会治理、商业活动等领域。可以说，数据之所以被视为资源就在于其具有分析价值，单个数据可以直接描述对象的某个或某类特征，但海量数据相互联系，就可能够抽象出数据对象背后的普遍特征，通过其透析客观世界或分析对象的规律、特征，预测未来的价值。

每个社会主体所掌握的数据是有限的，而要形成足够大、满足各种使用目的的数据集，[④]就必须利用他人掌握的数据(掌握数据的主体，称为数据控制

者),[5]同时也要让他人利用自己的资源,即实现数据的社会化利用,而不只是自我利用。显然,在数据资源化、资产化的背景下,已被公开的公共数据(特指公开可自由利用的数据)的利用价值是有限的。因此,必须给控制者一定的激励,才有可能实现数据的社会化利用。这便是困扰数据经济发展的数据赋权问题,即通过赋予产权来实现数据的商业化(市场化)利用。

但是,在人类文明的长河中,信息一旦被公开即被认为处于公共领域(public domain),是任何人可以自由利用的公共资源,任何利用者也不能排他支配或独享。[6]而且数据一直被认为是非竞争性的,也不适合私人独享。这是因为人类社会是在不断认识世界的过程中进步发展的,而人类对客观事实的认识需要借助符号、文字等工具,使用这些工具(即数据)对世界的客观描述(即信息)不能为任何个人所垄断,否则会妨碍人类共同生存和进步。[7]按照人类知识或智慧的DIKW(Data-Information-Knowledge-Wisdom)经典表述,"智慧源于知识,知识源于信息,信息源于数据"。[8]人类借助数据表述各种含义(信息),而对信息的应用组成知识,人类学习知识之后形成了智慧,于是人类文明呈"数据→信息→知识→智慧"递进式结构。为了激励人们的知识创造,设计出了知识产权制度来对创新成果给予有限度的保护,以激励创新,但这种保护给予的是对创新成果的商业性使用的专有权,而不是对知识内容(信息)或构成要素(数据)的专有权。[9]另外,信息自由也关系到言论自由、政治民主等内容的实现。因此,对信息的任何私权利均与人类社会的基本价值观相悖。因此,在DIKW体系下,法律一直拒绝赋予或承认私人对信息的排他性支配权,[10]更不用说作为信息载体的数据。

在当今数据时代,人类获取知识的能力和方式发生了巨大变化。数据时代的数据已经不再是传统的人类文明公共元素意义上的数据,也不是人类观察、测量、计算形成的对自然和社会现象的客观描述或记录,而是利用信息和通信技术(网络设备、传感器、智能设备等)生产出来的描述特定对象和客观现象的数字化记录。这些大规模和多样化的数据为人类自身识读分析的可能性极低,但通过汇集一定量的数据进行关联分析,可以分析挖掘特定对象的特性、规律或趋势。这种数据分析处理技术被称为大数据分析或数据挖掘。借助这种新型数据处理技术(包括人工智能),从各种网络设施和各种传感器(sensors)形成的大量和复杂的数据,成为大数据分析(人工智能分析)的原材料。这些原材料不再需要加工成为信息、形成知识再由人来理解、形成对各种事务的判断或分析,而是通过各种算法直接得出结论、预测或分析,以支撑各种决定——也就意味着,数据直接成为智慧的来源。本文分析数据时代人类知识或智慧生产新方式中数据本身生产利用的新特征,认为此种意义上的数据不是天然处于公共领域的,而是被生产出来的,由此提出数据生产理论,为数据时代的数据权利配置提供理论支撑。

二、数据的产生:原始数据的生产

数据时代是全息数据化的时代,无时无处不在网络上形成多样性海量数据,成为大数据分析(智能分析)的原材料。只有数据生产者愿意提供数据给他人,才能有充分的数据供应,满足大数据分析之需要。这需要对数据生产者角色的肯定,并确认和保护其数据权利。

(一)数据化:数据的生产

从2009年开始"大数据"(big data)成为互联网技术行业中的热门词汇。"大数据"是需要一种全新处理模式,才能成为具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。从数据的类别上看,"大数据"是指无法使用传统流程、工具处理或分析的信息,用来定义那些超出正常处理范围、大小和迫使用户采用非传统处理方法的数据集。实际上,人们早已发现,互联网本质上是一台超级数据生产机器,它将所有用户输入、传输、存储、交互的信息(内容)及这些行为的过程(元数据)记录下来。首先是每个用户利用网络生成、制作和交流的文字、图片、照片、音频、视频等内容都被记录下来并通过网络传播,形成网络数据的重要来源,

称为用户生成的内容(user-generated content)。其次是用户利用网络发布这些内容的行为本身也被记录下来,比如人们利用网络发送电子邮件、即时通讯,从事网络社交、购物、接受服务等行为均被网络自动地记录下来,被称为元数据,成为网络中最为重要、最有价值的数据。这两方面的数据使网络成为数据的生产机器。互联网产业之所以兴起且多以无偿模式运营,就是因为用户是其最重要的资源,而用户资源本质上在于有关个人的数据。

2012年是标志大数据时代到来的重要年份,这是因为此时各种传感器、智能设备和终端的大量出现,尤其是移动互联网、物联网等技术出现,使得网络不仅可以收集人的数据,[11]而且还可以收集机器运行、自然界变化、组织运行等数据,人类开始进入了万物互联(Internet of everything,简称IoE)[12]的时代。今天,我们的电子终端都是智能化的,智能手机、智能手表、谷歌眼镜等可以随时感应和记录我们的时间、位置、活动、行走轨迹等信息。从地铁到智能马桶,几乎所有的设备机器均成为数据源。即使用户没有联网,因到处布设、互联的传感器也会留下用户的数字轨迹。除此之外,智能终端又进一步推动了巨量级的网络社区、电子购物、物流网等出现,线上业务与线下业务相融合,产品服务智能化不断升级,经济或商业模式转型升级,数据收集系统不断普及,网络数据开始出现海量集聚,真正的大数据时代由此而生。正如Microshare公司的两位作者指出:信息科技在不断生成和处理数据。直到最近,这类数据管理的增长还是线性的、可管理的和可预测的,但是,世界已经达到数据奇点——数据的体积、速度和种类呈现非线性激增。我们生活在一个产生、传输和存储额外一个字节数据的边际成本几乎为零的世界,并且正因为边际成本几乎为零,所以每一个能够生成数据的事件所做的就是:生成数据。[13]"大数据不仅指数据的绝对规模大,还指就某主题的综合性数据集而言,相对规模也大。"[14]大数据技术首先意味着以复杂多样的数据形式并以惊人的速度产生新数据,我们可以用"数据化"来描述这一伟大的社会变革。

上述所讲的数据化有两个要素:一是数据有关或描述的对象(主题),即数据源(data source);二是对该对象的数字化记录、描述和呈现。数据是对现实世界(构成要素)的记录或描述,而不是现实世界本身。将位于某处的树木拍摄下来,形成有关于树木的数据,成为分析树木本身或某位置有什么的数据;将机器运行过程或轨迹记录下来,成为有关机器及其性能的数据;对用户网络浏览、订购商品或服务内容及其时间、地点等信息可以描述勾勒一个用户的特性、偏好、轨迹等。这些人、物等本身是一种现实存在,而数据则是它们的数字化呈现(图片)或数字化描述。我们将数据与所描述对象的分离过程(即数据化过程)称为数据生产。数据生产即数据采集——通过技术手段将特定对象本身及其行为或过程以数字形式记录下来,形成用"0"和"1"记录的数据。数据生产旨在说明数据并不是天然存在的,而是通过各种网络设施和设备记录生成的。在英文中,一般用"generate"来描述数据生产(该词的本义是"产生"或"源自"),以区别于物之生产(加工制造)。

在大数据语境下的数据包括人"生产"的信息(比如,人写的文章或作品、用户创制内容等),但是更多的是机器生产的数据。这里的机器包括网络设施、传感器、智能设备等能够自动数字化记录特定对象活动、周边环境等的任何机器设备(或系统)。在笔者看来,大数据分析时常会用到两类数据,即人创作或制作行为生成的数据(信息)以及机器生成的数据。但是,作为一种数据类型,大数据是指其大小超出了典型数据库软件的采集、储存、管理和分析等能力的数据集,而如此巨大的数据量显然也非人力可采集。大数据是指机器自动生产的数据。机器生产的数据是人类文化史上不曾有的独特数据,是当今数据化客观世界的主要力量。我们仍然可以采取大量田野调查、实验记录等人工观察和记录的方式来描述世界(生成数据),但是借助计算机系统、网络系统、传感设备等来记录、感知、抓取活动对象或描述对象的行为或变化,形成可用的数据,已成为数据化

客观世界的主流。这种将特定对象转化为数字化的信息(数据)的过程,就是本文所称的数据生产。因此,数据生产就是产生独立于描述对象的数据,使我们通过数据的处理和分析即可以了解该对象。为此,我们首先需要区分数据来源与数据生产。

(二)数据来源与数据生产的区分

数据生产与数据来源有密切的联系。研究者多从来源角度研究数据的产生。例如,联合国经济和社会事务署统计局曾提出,将数据分为源于人的数据、源于组织运营的数据和源于机器的数据。[15]显然,这样的分类是依据数据源头或描述对象的一种分类,而不是从数据生成或生产的角度进行分类。为了更加清晰地揭示数据的生成,我们需要从理论上区分数据来源与数据生产。

在数据时代,任何数据皆有源,甚至数据必须有源,没有源就丧失了作为大数据分析的价值,因而数据的来源问题是数据最为重要的"标签"。正因为数据总是对特定对象进行描述,所以才成为现实世界的再现工具。作为一种社会资源,数据一定是对特定对象的描述或者是关于特定对象的,这便是数据来源。数据有源,可以用来分析其描述对象,揭示其规律,预测其行为,因此就具有价值。数据来源实际上只是数据"关于"的对象、可识别(认知或描述)的对象,而不涉及数据如何生成。数据描述的对象包括个人(自然人),也包括组织、物、机器、天体等非个人。这些对象是将数据与现实世界关联起来的"媒介",所谓的数据分析,即通过数据本身逻辑演算来认识或识别、预测数据描述的对象(来源)。因此,标记数据的来源和确认其可描述或分析的对象是非常重要的。

当数据不能指向或联系特定对象时,该数据就是抽象的存在,不具有分析价值,数据也会因为时间推移而丧失对特定对象的分析价值,尤其对于时间敏感的数据。[16]当数据不具有分析特定对象的价值时,就会进入公共领域,成为任何主体都可以自由利用的数据。在网络世界存在着大量来源不明或者难以归属某个对象的数据,这样的数据也可能具有潜在的价值。但是,它需要新的技术或劳动来挖掘其价值。就相当于在以语言文字、符号、图形等为载体的人类文明中存在公共领域,在数据世界中我们也承认公共领域的存在,以给人们再次开发与利用数据提供公共空间。

一旦我们承认资源性数据是对特定对象的数字化记录,而这些数字化记录又不是天然存在的,不是从数据描述对象自然"流出"的,那么我们就必须承认数据生产。数据生产意味着数据是外在力量作用的结果,而不是数据源于自然的产物。在具有分析价值的数据应当被作为一种资源背景下,数据生产就成为构筑整个数据资源利用秩序的基础,以此可以构筑后续数据加工处理、流通交换和分析利用秩序。如果不承认数据生产,数据加工处理、数据的流通利用秩序就没有起点,整个数据社会化利用的秩序大厦就无从构建。

既然存在数据生产问题,那么我们就应当区分数据来源者和数据生产者。数据描述的对象只是数据的源头,而并不一定生成、产生数据。为了更准确地表述数据,我们将数据描述的对象(主题)称为数据源或数据来源者,而将设计数据采集工具系统或设备、从事数据采集的活动称为数据生产,而将对数据生产做出实质贡献的主体称为数据生产者。数据的生产活动实现数据与描述对象之间的分离,形成与描述对象的独立存在,形成供数据分析的原始数据(raw data)。

数据描述的对象为数据采集的对象,也是数据的来源"主体"。实际上,描述的对象即数据主题(subject)。根据数据描述的对象(主题),数据大致分为关于人的数据(描述人身份、属性、行为等的数据,即个人数据)、关于组织的数据(描述组织基本情况、运营情况等,即组织数据)和关于物的数据(描述自然界和物的属性及其变化或运行轨迹等)。这些描述人、组织和物的数据均是由特定的主体生产出来的。这里的生产既包括人(自然人)的录制,也包括人或组织通过网络设施、智能设备和传感器等(统称为机器)记录或生产。此处先分析后者——机器自动

产生的数据，前者将归入个人数据部分一并进行分析。

在数据描述自然界、机器设备的情形下，该数据是对自然现象、机器运行的记录。该记录存在两种情形：一是对设备之外的物体结构、运行等的记录，此时记录对象和记录设备不同；另一种情形是，数据来源于机器记录本身，是对机器自身运行的记录，此时记录的对象和记录设备本身是一体的。前者，如传感器对气温、空气质量、天气变化的测量和记录；后者，如飞机对自身飞行状况的记录、汽车对自身运行状态的记录等。因此，机器设备既可以感知和记录外部，也可以记录机器本身的运行(智能设备具有该功能)，形成了源于机器的数据。在这两种情形下，数据全部来源于或产生于机器。在这里，数据经历了从无到有的过程，也可以说是机器生产出了数据。只是我们在用“数据生产”一词的时候，主要目的在于确定是由谁“生产”或“制造”出了数据，而不是机器本身。由于自然界和物本身不是主体，说数据来源于机器，不如说是设置和运营机器的主体(组织或个人)生产出数据。对于来源于系统、设备等的关于物的数据，系统、设备的所有者、运营者就是数据生产者。来源于物的数据的初始权利配置相对比较简单，我们可以单纯基于数据生产将数据控制权配置于数据生产者。

当数据描述对象是组织和人时，因为组织和人的主体属性，这些主体可能参与到数据的生成与生产过程中，因而存在这些主体是否是数据生产者或享有何种权益的问题。

当数据关于组织本身时，组织是该数据主体，但同时，组织也是该组织数据的生产者。[17]在这里，来源者与生产者也高度合一。一旦我们认可组织数据也具有价值时，那么作为组织数据的生产者，亦应当保护其利益。在传统法律框架下，除非关于组织的数据落入商业秘密范畴，对于组织产生的数据一般不予保护或者法律上没有明确如何保护。在数据时代，在各个企业建设自己的信息系统或者上云接受云服务的情形下，企业运行所形成的大量关于自己的数据以及该数据对外提供服务和利用问题也是今后值得关注的。因此，我们有必要确认组织作为自身数据生产者的地位，以便解决今后组织数据的赋权问题。需要指出的是，当组织生产关于自身的数据时，数据来源者与生产者是合一的，数据来源者(组织)的合法权益可以归并到数据生产者权益中加以保护；而当组织生产关于其他组织的数据，如果其他组织对该数据享有权益，那么还存在数据来源者的保护问题。原则上，作为数据的生产者，只要不侵害数据上组织主体或个人主体的权益，那么数据生产者即享有关组织或关于个人的数据生产者权益。至于组织数据的生产者是否享有权利、享有什么样的权利，不是本文的主题。

当描述的对象是个人时，因人是具有能动性的主体，需要给予特别考量。

(三)个人数据的生产：机器(系统)与人的作用

人是有思维、能创造的主体，人有获取数据和分析数据的能力，从而创造了人类灿烂的文明。进入数据时代，由计算机和网络生成的各种数据，成为海量数据的来源。从数据生产的角度，关于特定个人的数据大致分为两类：

一类来源于人的创制，即个人可以通过录制、拍摄、汇编、制作等创制、创作形成的各种数据，不论这些是否构成作品，均成为大数据分析的原材料。同时，人类在利用网络进行各种形式的“创作”过程反映了创作者的思想，其创作的成果因满足作品构成要件而受著作权保护。但是，受著作权保护(保护思想的表达)并不妨碍创作内容(信息)作为数据分析的对象。也就是说，来源于用户创制的内容即使受著作权保护仍然具有分析价值，构成重要的(大)数据资源。数据分析是挖掘信息本身所蕴含的来源主体的个性或行为特性的过程，而非信息所表达的思想内容。对于数据分析来讲，能够联系到特定创作主体的信息(即使构成作品)，既可以分析作品内容本身，还可以分析信息生成的时间、地点、方式等创作行为本身，以形成创作主体或分析相关事物的特性。例如，各类网络交易或网络服务平台上的用户点评信

息，不管是否被认定为作品，都具有分析该用户特性和被点评商家或商品(服务)特性的价值。这便是用户创制内容的价值所在。除了此类信息来源于用户外，用户在从事各种网络交易、接受各种服务过程中还会主动提供有关个人的一些数据(通常是个人身份信息、联系方式、账户等)，被称为用户提供的数据。总之，用户提供的和用户创制的信息构成用户个人为数据分析生产的原料。

另一类数据是关于特定个人因人使用计算机和网络的行为过程被网络服务器记录下来而形成的行为轨迹或过程数据。此类数据并不是来源于人的创制(生产)，而是来源于人的数据。它属于关于人的数据(个人数据范畴)，但人只是被动地参与到数据生产中，而没有积极地提供和生成数据。在这种情形下，人是被记录的对象，而人的行为、事实或事件的数据化是由系统设备完成的，是机器生产了数据。个人行为或运行过程转化为数据，并不一定是主体本身实现的，而是设备拥有者架设系统环境，通过技术手段形成的。这一过程实质上就是对数据的采集，即本文所称的“生产”。我们将架设基础设施、形成描述来源对象的数据的主体，称为数据的生产者。数据生产者就是将来源者(人—对象)本身状态和行为以数字方式记录下来，形成可供进一步分析该对象的数据主体。因为数据描述的对象是人，是主体，但是此时的主体并没有发挥其主观能动性来创设数据，只是数据的源头。也就是说，数据主体只是来源者(描述对象)而不是数据生产者。此时我们需要区分数据来源者和数据生产者。数据来源者是数据描述的对象，而数据生产者完成了来源者的数据化。

在海量的关于个人的数据资源中，大多数数据属于后一类型，即人在利用各种网络设备从事各种活动过程中，由系统、物、设备记录所形成的轨迹数据或行为数据。虽然个人只是数据来源者，没有产生数据，但是没有人的参与，光有网络设备也是不能生产出数据的。因此，数据生产者如何考虑数据来源者利益，是个人数据保护中需要考虑的因素。对此，本文亦不作深入讨论。

三、数据的加工处理：数据价值的“生产”

生产出来的原始数据一般还不能直接成为数据分析的生产资料。生产出来的数据需要经过清洗、整理、汇集后才能进行逻辑推演、运算分析，洞察分析对象。这些数据处理行为都会改变数据的价值或形成新的价值，实践中统称为数据处理。笔者将这一过程区分为相互联系的两个过程或行为：一种是将原始的数据加工处理成为数据分析的材料，这便是汇集性数据处理(本文也称为数据集的生产)；另一种是分析性处理，经过数据演算分析，为人们提供新知识、新判断，支撑人们的决定。分析性处理即通常所说的数据挖掘。正是这两种行为最终实现了数据资源的分析价值。

(一)数据汇集处理：数据集的生产

数据从描述对象采集或分离后，还不具有直接的使用价值或者仅有有限的使用价值，需要打破不同类型数据的孤立性，实现数据的互通、再提炼和形成更有价值的数据——这便是数据集的生产。数据集的生产是对已形成数据的加工处理活动，它是按照特定目的，收集汇聚、清洗整理、分类归集，形成可用的数据资源，使原生态的数据加工成为具有使用价值的产品性数据，被称为数据集(datasets)。从价值实现的角度，数据从“原材料”的变现到“粗加工”后的变现有两种方式，即API接口的调用和数据文件集的生产。[18]显然，以API接口的方式将原始数据变现，是那些拥有巨大用户群体和良好数据采集和归集架构的网络服务商或平台实现数据价值的方式。这些网络平台大多在采集过程中即对用户数据进行了初步整理，因而可以实现原始数据的直接变现；而数据集或数据文件集的生产则是原始数据变现等更普遍的方式。

数据量大并不一定满足对特定对象数据分析的需要。数据集的生产核心目的是从他人处获取关于相同对象或相同主题且足够多的数据。由于每个主体的数据是有限的，每个数据主体都需要从其他数据控制者那里获取更多维度或数量的数据，才能生产出满足不同应用或分析需要的数据产品(数据集)，

而且随着数据数量越大、维度越广,分析的结论就越全面和精准,各数据生产者则成为数据集生产者的原料供应者。显然,这需要在承认数据生产者对数据的使用控制权前提下,通过各种共享、交换、许可使用等方式来获取数据。

从价值产生的角度,数据的收集、汇集、整理等加工处理活动也属于数据的生产活动,它改变的是数据存在形态,由原始形态的数据变为具有特定使用价值的数据集合,供各种数据分析使用。数据集的生产是大数据分析的基础,数据集就是这种生产活动的产品,只是这种产品相对于大数据分析(或人工智能)仍然是原材料。因此,数据集的生产者是数据分析者的原材料供应者。这种原材料生产者也需要投入大量物力和财力,其劳动成果也需要得到保护。而这种保护只能通过赋予数据集的生产者对数据集一定的控制能力来实现。

之所以要保护数据初始加工处理者的利益,是因为他们的劳动创造了价值。数据的采集完成了数据与数据主体的分离,使数据成为人类处理分析的对象,所形成的数据具有潜在的价值,即原生固有价值(intrinsic value)。而之后的数据处理加工,不改变数据,但改变了数据质量、维度、数量等,相当于在数据原生固有价值基础上添附了新价值,形成了添附价值(added value)。正因此,我们也将后续的加工处理称为生产性活动。数据的采集完成了数据与数据主体的分离,而数据汇集整理则使数据具有使用价值,可供人们分析使用。相应地,我们分别称为数据的生产(者)和数据集的生产(者)。

数据集是经过初步加工处理后,区别于原始数据的形态、含义和价值的数据,属于加工处理的数据(processed data),具有产品属性,因而数据集亦可以被称为数据产品。在数据驱动的时代,数据集的生产占有非常重要的位置,是支撑整个大数据产业、人工智能的基础。当然,我们也不排除对原始数据的加工处理后,可以形成可供人类查询使用的数据库,以直接获取知识的司能。

需要指出的是,数据一直处于被加工处理的过程中,加工处理后的数据可能再次地作为原始数据被进一步加工处理。因此,原始数据与加工数据并没有严格的界限,而主要取决于数据的目的或用途及其人工干预的程度。如同数据和信息可以转化一样,原始数据和产品数据也是相对的,产品数据可以因时间、用途或环境的改变而成为其他数据产品加工的原材料(成为原始数据)。如此,原始数据和产品数据的区分也仅仅具有规范价值。

(二)数据分析处理:数据挖掘

数据分析最终的价值是为人类各种决策提供知识或决策支持服务,而这一过程需要对数据进行深度的加工处理,发现分析对象的规律或预测未来趋势,从数据中得出新知识、新发现,以做出预测性判断或解决方案。实际上,大数据与大分析(big analytics)捆绑在一起,数据的价值正是源于大分析。这是因为数据分析(data analytics)可以"获得洞见(知识创造)"和"自动决策(决策自动化)",[19]是数据价值创造和实现的机制。

在DIKW结构下,数据分析的结论表现为信息,或者说是不同于分析所使用数据(原料)的新数据,是从原始性数据(数据集)分析、解读出来的数据,也被称为推断数据(inferred data)或衍生数据。[20]相对于原始数据,分析结论不再是识别特定对象的数据,而是从数据中发现新知识、新规律甚至作出决策。基于原始数据的推断、演算分析出来的数据类似于传统的将玉石加工成玉器、木材加工成家具,是一种新物或产品。[21]对原材料性的数据进行处理、分析形成的新数据,与原始数据(数据集中的数据)存在推断、推演或衍生关系,但已经是两种性质的东西,不宜继续在相同的层次、运用相同的原理讨论其归属。数据发现需要保护的不再是数据本身,而是信息的应用价值。至于如何保护则取决于数据挖掘所形成的成果形式和价值。有些发现的价值可直接用于提升决策精准度和效率(辅助决策),就不需要提供额外的保护;有些可能需要借助著作权来保护AI创制内容,[22]甚至可能会借助对专利保护具有独创性的AI系统或算法本身来对其应用价值进行保护。

新数据的产生离不开不断进步的计算机科学。数据分析处理伴随机器学习、深度学习等技术应用而不断发展，缺失大运算能力的算法系统是不可能完成的。大数据为人工智能提供了原材料基础，而人工智能也为数据分析提供算法支持，为大数据应用插上翅膀，使得传统的数据分析技术有了新的想象空间。通过不同形式的人工智能分析大数据，使人类具有从数据中获取新知识或者新洞察的能力。有了针对不同需求或目的的数据分析，原材料性质的数据才最终发挥出应用价值。可以说，数据分析处理是数据经济价值最终实现的前提。

显然，数据的分析处理是一个技术能力问题，有数据分析技术和能力的主体并不一定拥有数据，因而数据分析就逐渐分化为一种专业服务。于是，在数据经济时代，逐渐形成数据生产者、数据集的生产者(数据汇集处理)和数据分析者的社会分工，而促成这种分工的关键就是数据的流通(为数据集生产提供原料)和数据集流通(为数据分析提供原料)。没有数据和数据集流通，只有数据生产者利用自己的数据从事数据分析，是低效率的，甚至也不可能有大数据分析。为了形成适应数据经济发展的社会分工，需要界定数据经济价值链条中每个主体的角色，肯定或承认他们各自的贡献，并从最终的数据价值中获取相应的回报或利益。在这个过程中，数据生产者是数据矿石的生产者，数据集的生产者(汇集处理者)是数据矿石的分拣者，而数据分析者是数据矿石的“冶炼”者，三者共同构筑数据经济产业链(如图1所示)，而社会中的每一个主体都可能是其中的某个角色承担者。不过，数据产业中这样的分工并不妨碍这些角色的合一。因为在数据即采即用的情形下，原始数据的生产者、汇集处理者和分析处理者是由同一个主体完成的。因此，这里的角色旨在为数据产业分工提供分析工具，通过对每个角色配置相应的权利来构筑产业或经济秩序。为此，基于前面的论述，笔者提出数据生产理论。

四、数据生产理论

大数据分析处理通常被概括为收集、存储、处理、应用四个环节[23]，也有分析者将该过程描述为数据从“原材料”到“粗加工”，再到“精加工”过程。[24]本文将该过程区分为数据生产和数据分析，并将数据生产细分为原始数据生产(采集)和数据集的生产(汇集性处理)。这样的区分并不否认数据分析在整个过程中被应用的价值(实际上整个过程是由分析驱动的)，但有利于区分需要保护的客体，因而正确配置数据权利。笔者称之为数据生产理论。

(一)数据生产理论的内涵

基于上面的论述，笔者提出以下三个方面的区分理论，作为笔者的数据生产理论：

其一，数据生产和数据分析区分理论。数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息并形成结论，从而对数据加以详细研究和概括总结的过程。数据分析是“指各种揭示洞见的技术手段以及促成理解、影响或控制该等洞见之数据客体(例如自然现象、社会制度、个

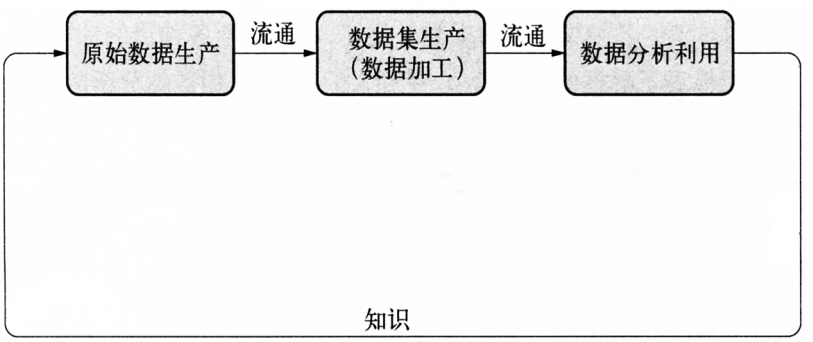

图1 数据产业链

体)的各种工具。”[3]数据分析和建模工作高度依赖计算机,多数工作使用算法/程序完成。随着大数据时代的来临,传统的数据分析逐渐为大数据分析所替代,大数据分析演绎为人工智能。受益于计算机技术在数据采集、存储、计算等环节的突破,人工智能已从简单的“算法+数据”发展演化到了“机器学习+深度学习”阶段。这种数据规模和分析技术的改变并没有改变数据分析的本质,即提出/发现问题,分析数据得出结论。如果说数据分析(含大数据分析,下同)是通过数据的运算分析得出有关数据描述对象的一些结论或新发现,辅助人类决策的话,那么人工智能则还具有创作或创新能力,能创作出独立于原数据的作品。因此,数据分析是行业知识和计算技术(算法)的应用,本质上属于创造性活动或智力劳动范畴。因此,数据分析不是在生产数据,而是对数据进行运算分析,对数据描述对象作精准的判断或预测,如果将二者区分开来就是要将数据分析或AI的结果保护与分析所依赖的数据保护区分开来。

数据生产为数据分析提供原料,它包括生成描述对象的原始数据和汇集更多有关该对象的数据。为此笔者将其划分为原始数据的生产和数据集的生产,前者是描述数据与描述对象的分离过程;而后者则主要是:描述相同对象的各种维度的数据被汇集在一起,用于各种目的的数据分析,提供各种解决方案。虽然数据分析体现了数据的最终价值、受到人们的重视,但数据分析的精准性取决于所依据的数据(集)的数量和质量,因而只有解决了数据分析原材料供给,才能够有真正的大数据应用,才能够有数据经济。因此,数据生产理论旨在解决数据分析的原材料(数据)供给问题。

其二,原始数据生产和数据集生产区分理论。数据生产实现数据从无到有,数据集生产则实现关于某个对象的数据汇聚和优化。大数据应用最为关键的是能够形成用于分析特定对象的数据集,因而数据集的形成得到业界的广泛关注。在某种意义上,收集足够量关于某个分析对象的数据是数据集生产的关键,至于数据的清洗、分拣或筛选、分类整理等则基本上是通过技术和人力能够解决的事情。这样的汇集处理活动能够改进数据质量,满足特定目的的数据分析需求,创造一定的价值,因此应当予以保护。对此没有人质疑,而问题在于如何获取足够多的数据。数据生产理论将数据生产区分为数据集的生产和原始数据的生产,不仅在于肯定收集并加工处理数据形成有用的数据集的劳动价值,而且在于将数据与描述对象的分离过程视为一种生产劳动,在明确数据产业分工或价值链的同时,为原始数据生产者寻求法律保护提供理论支持。一旦将原始数据的形成也视为一种生产劳动,生产者对于数据的控制权亦应当予以尊重和保护。在肯定其对生产数据具有财产利益前提下,有利于数据生产者将其生产的数据提供或流通给他人,以满足数据集生产(数据收集)的需要。由此使数据持有者愿意将数据提供给他人使用,需求者可以获取足够多的数据,最终促成数据的社会化利用,充分发挥数据的价值。

其三,数据来源和数据生产区分理论。在大数据概念下,数据皆有源,而“源”就是数据描述的对象或者数据主题。数据生产理论便是建立在数据皆是特定对象描述或者关于某对象数字化记录的理念上,由此将数据描述的对象称为数据源,而将描述或数字化记录本身的过程(与描述对象的分离)称为数据生产。从源头上,数据有关于个人(个人数据),也有关于组织、物(机器)、自然界等对象,这些源构成数据主题。从数据的形成角度而言,后一类数据往往是被记录的结果,而前者则相对复杂。因为人是主体,个人可能参与到数据生成甚至创设数据或提供关于自己的数据。因此,当我们在实践中宽泛地将“关于人的数据”描述为“来源于人的数据”时,就会掩盖个人数据形成的复杂性。而根据数据生产理论,来源于人的数据分两类:一类是由人创制和提供的,另一类是被记录生成的。个人在网络上的言论、发布和提供的信息等不仅来源于个人,也是个人“生产”的;但是网络系统和传感设备也会实时地记录个人所有的网络行为和非网络行为,无论个人是否能够感知记录过程和内容,无时无处不在网络和传感

设备实时记录人的行为。事实上,这部分才是如今大数据(分析)概念中的主体部分,而这部分数据是由网络设施和传感设备拥有或运营者生产的,而不是由个人"生产"的。

数据来源和数据生产的区分是对个人数据的重要分析工具。个人数据是关于个人的数据,个人是源头,但源头并不表明所有个人数据均产生于个人。关于个人数据的一个基本结论:个人数据来源于个人,但个人并不都是个人数据的初始的生产者。在网络化时代,个人数据保护必须面对的现实是,在能够联系到个人的数据中,只有少数是由个人提供和创立的,在大多数情形下,个人不能准确知道自己的哪些数据被他人采集,一开始也不拥有或控制数据。这与20世纪七八十年代以个人提供的个人基本信息为主体的个人数据保护所面临的场景完全不同。因此,数据生产理论除了清晰地描述数据的产生、价值形成和实现过程外,还有助于揭示当今个人数据的产生和控制事实,因而可以成为数据权利配置的理论基础。

(二)数据生产理论的价值:数据权利配置的理论基础

在人类进入到数据驱动的时代后,数据成为社会的基础资源,成为具有经济价值的资产。这需要我们将数据视为一种资源,研究和构筑数据资源的社会化利用的秩序来支撑未来的数据经济。由于产权是所有有价值资源得以有效配置和利用的工具,因而有关数据产权的讨论成为近期及未来的学术热点。尽管赋予某种财产权已经成为主流,但也不乏反对者。[29]本文无意详细评论各家学说并探讨可行的数据权利配置,而是研究数据产生、价值形成与实现的过程,并将之概括为数据生产理论,以期为数据权利配置提供理论支撑。

在数据权利配置研究中,似乎存在一个假设,即数据是天然存在的,或者被当作一个"无主"资源。然而,数据生产理论则提出数据并不是天然存在的,而是被有意识地记录、生产出来的。有了这样的认知,就使我们可以解决数据赋权的起点问题。至于赋予什么样的权利,则是需要进一步讨论的事情,这里暂且将数据上需要配置的权利理解为数据的"使用控制权"。如果承认数据是被生产出来的,那么数据权利配置的核心问题就是——谁是数据的生产者,数据生产者应当被给予怎样的数据使用控制权,权利内容、效力或限制如何等等。

数据生产理论首先承认数据与描述对象的分离和记录过程(数据采集)属于数据生产,它为数据分析提供原始素材,这样的数据生产者地位应当被给予确认和配置相应使用控制权,使其成为数据资源配置的初始发动者。在弄清源头之后,数据权利配置应当沿着数据产业链形成价值或创造展开,价值创造者应当给予适当使用控制权,以保护其投资或劳动。数据生产理论将数据与数据源分离之后的数据处理过程区分为两类行为:一种是生产出为特定目的分析所需要的一定数据和质量的数据(表现形式为数据集);另一种是对数据进行演算分析。前者改变的是数据的质量和数量,仍然属于数据形式的改变、属于数据生产范畴;而后者则是发现数据背后含义,对特定对象做出新的认知或预测,形成的数据实质上已经属于信息或知识范畴,不再属于数据生产的领域。尽管数据分析的结果形式上也最终表现为数据,但该数据不再是描述特定对象的数据资源(准确地讲是信息),一个完整的数据分析过程就此结束(当然,如果分析结论可以作为新一轮数据分析的原材料,那么分析结论数据的控制者可以行使使用控制权,进入下一数据分析过程)。

在某种程度上,数据分析技术引领着数据生产,有怎样的数据分析方法和技术就有怎样的数据生产。数据分析的精准性既取决于数据集的质量,也取决于数据分析方法和技术。在算力确定的情形下,数据分析的精准性主要取决于质量,即数据规模、数据维度、关联度、正确性、时效性等。区分数据(集)生产者与数据分析者的主要目的是明确数据拥有者与数据分析者之间的分工。至于数据拥有者和数据处理者(算力提供者)如何分享数据分析的结果,则需要二者通过合同约定来确定数据分析结果和收

益的分配。也就是说,即便数据分析结果产生了新数据、属于信息或知识范畴,其保护问题也是一个知识产权法命题,但不作为本文探讨的主题。本文的数据权利的配置仅限于原材料性质的数据、数据集的权利配置问题,是人类进入数据时代开启的全新课题。

因此,数据权利配置主要讨论数据生产者和数据集生产者的赋权问题,而在涉及个人数据时,还涉及数据来源者享有什么权利问题。本文以个人数据为例说明数据生产理论对权利配置的适用。

在个人数据权利配置方面,世界各国均将个人数据保护理论建立在数据与特定个人有联系或能够识别特定个人的基础上。因为人是主体,所以对个人数据的处理应当尊重个人意志,给予个人必要的控制权利。个人控制论成为世界各国对个人数据保护的普遍认知。这样的理论形成于20世纪七八十年代。这个时代的个人数据限于个人属性、联系方式、账户等基本身份信息,且以个人提供给特定主体使用为主。在计算机开始应用之时,人们最为担心的问题是,当这些个人信息被提供给政府或企业后,这些机构可以长期保存这些信息且重复使用,可能会危害个人自由和尊严。如今,网络化生存无疑给个人基本信息的滥用带来更大可能性。问题还不仅限于此,无时无处不在的网络导致个人行为被全息记录下来,形成对个人特性的深度观察,跟踪个人行为、透析出许多个人隐秘信息,甚至被用于危害人身和财产安全的犯罪。与此同时,网络化、数据化和智能化又使人类进入数据驱动经济时代,凡是与特定个人有联系的数据都具有资源价值,于是个人数据成为潜在的经济资源。因此,个人控制是否当然地演变为个人数据使用控制、让个人成为数据经济资源的初始决定者,就成为这个时代需要解决的首要问题。

数据生产理论旨在揭示数据时代作为资源意义的数据是被生产出来的,并经过汇集处理后才能实现其分析价值,并把汇集处理视为数据分析原料的生产。而在这一过程中,个人只是数据描述的对象和源头。个人有可能参与到原始数据的生产过程,但数据生产的主体是数据采集设施设备运营者。这样的理论真正将数据视为一种经济资源,来构思数据权利的配置。作为个人数据来源者在数据上享有主体利益,但这种主体利益仅仅建立在个人是数据的主题(数据与个人有关)基础上,而非个人数据归属于个人、是个人的财产。也正因此,普遍认为去除关联性(即去识别处理或匿名处理)的个人数据就不再受个人控制。[27]个人数据保护旨在保护数据上主体权利,目的在于防范个人数据使用行为对主体利益的侵害(即个人数据的滥用行为),而不是赋予个人对数据的支配(个人对个人数据使用不享有控制权)。[28]换言之,赋予个人对个人数据处理一定的控制权,主旨在于防范个人数据的滥用(保护个人尊严和自由),而不是赋予个人数据以使用控制权。通常,个人数据权利需要通过转化数据控制者的维护个人主体权益的系列义务来实现,而不是通过个人对个人数据的支配来实现。至于作为源头的个人如何分享个人数据后续分析利益的经济利益,大致属于商业问题,可能需要借助市场力量来调节,而不是通过法律上的权利配置。笔者相信数据生产理论可以为数据资源的权利配置提供新思路和理论支撑,由此设计出契合数据经济需要的数据权利,来构筑数据利用基本秩序。对此的系统论述超出本文范围,留待后论。

注释:

①1998年,美国计算机科学家Mashey在一次演讲中首先使用"big data"概念。《自然》杂志在2008年9月发表了有关大数据的封面文章《大数据:从数据中提取内涵》,"大数据"由此成为IT行业中的热门词汇。现在人们多用大数据来描述当今信息技术及其应用。IBM将大数据概括为4个"V",即数据量(volume)大、数据形态多样(variety)、数据处理速度(velocity)快、数据价值(value)密度低。

②2013年,软件和信息产业协会发布了《数据驱动创新》白皮书,提出并定义了数据驱动创新;OECD于2015年发布了研究报告《数据驱动创新:为增长和福祉的大数据》;2014

年,欧盟议会、欧盟理事会、欧洲经济和社会理事会和地区理事会联合发布一份通讯《走向繁荣的数据驱动经济》,明确提出“数据是未来知识经济和社会的核心”;2017年,欧洲政治战略中心发布了《进入数据经济》。伴随这些文件的使用,数据经济逐渐成为描述大数据应用下社会经济形态的主流词汇。

③涂子沛认为,“数据正在改变所有那些组成文明的要素”“就如同农耕之于古代文明,工业革命之于现代文明,数据将催生一种全新的文明形态”,他称之为“数文明”。参见涂子沛:《数文明——大数据如何重塑人类文明、商业形态和个人世界》,中信出版集团股份有限公司2018年版,前言。

④数据集(big datasets)是设备、传感器、互联网交易、电子邮件现在和未来产生的大规模、多样、复杂的、纵向和(或)分布式数据集(datasets)。See National Science Foundation,"Solicitation 12-499: Core Techniques and Technologies for Advancing Big Data Science & Engineering(BIGDATA)"(http://www.nsf.gov/pubs/2012/nsf12499/nsf12499.pdf, last access 2019-03-10).

⑤数据控制者是国际社会在个人信息保护立法中提出的一个概念,它指实际控制使用数据的主体;而与此相对的概念是数据主体,特指个人数据描述的对象,即自然人(因为人是主体)。数据主体对个人数据也有控制,只是这种控制是基于维护数据上的主体利益。

⑥许多知识产权学者对知识产权法并不保护信息作过论述,例如美国知名知识产权法学者Pamela Samuelson在1989年撰文指出,版权法视信息为可以被无限制使用的公共领域的素材(public domain material),专利法规定在授予专利后将涉及专利发明的信息置于公共领域,信息自由传播而不是通过财产权给予限制是美国联邦知识产权法体制目标。但是她也指出美国在特定条件下将信息视为财产的判例。在文中她检讨了美国法院存在将信息视为财产的两个判例:如在Ruckelshaus v. Monsanto Co., 467 U.S. 986(1984)中,法院认为提交给联邦机构的研究数据仍然是企业的财产;在Carpenter v. United States, 108 S. Ct. 316(1987)中,法院认为报社对出版时间表和栏目内容视为报社的财产,她认为是否将信息视为财产没有充分论证。因此,她认为在人们对于何时是财产而何时不是财产缺乏共识的前提下,定义信息是财产有些为时过早。参见Pamela Samuelson,"Information as Property: Do Ruckelshaus and Carpenter Signal a Changing Direction in Intellectual Property Law?" 38 Cath. U.L. Rev.365(1989).

⑦联合国教科文组织信息与信息学部主任菲利普·奎奥在2000年5月10日的一次专家研讨会上指出,“知识的获取,是开放社会的重要原则之一。垄断信息,大大限制了对知识的获取。解决这个问题,需要通过国家和国际立法对信息网络实施调控”。参见[美]格里钦·西德胡:《信息时代,人人有权充分获得信息信息垄断,只会妨碍经济社会发展》,载《科技潮》2000年第10期,第95页;有学者指出,各国著作权法都认为,对于人类进步,保护著作权与保证公众获取信息同等重要。技术锁闭、网上合同以及欧盟对于数据库的保护指令对人类进步的这一基础提出了挑战。参见[美]安娜—路易丝·玛丹:《保护知识产权防止信息垄断——IT的公平问题》,载《中国青年科技》2001年第3期,第42~43页。

⑧Jennifer Rowley,"The wisdom hierarchy: representations of the DIKW hierarchy", 33(2) Journal of Information Science 163-180(2007). 也有学者从信息科学的角度出发,认为DIKW结构并不合理,方法上有缺陷,不应当作为信息科学和管理的标准。数据是以适当语义和程序方式记录的所有东西,信息是一种弱知识(知识也是弱知识),智慧是那些适当行为人对宽泛的实践知识的掌握和使用。参见Martin Frické,"The knowledge pyramid: a critique of the DIKW hierarchy", 35(2) Journal of Information Science 131-142(2008).

⑨技术发明取得专利权的前提是向专利管理机构(进而向社会)披露其设计或发明的内容,从而在一定期限内排他实施其技术方案,以激励技术创新(获取商业利益);著作权法通过思想和表达二分法,既保护作者有独创性的表达,同时置思想(作品内容)于人们可以公开获取、学习和再创作的状态,再加上合理使用制度赋予个人合理利用作品的权利,因而著作权法保持了作品内容足够的开放性和公众可获知性。因此,知识产权体系中保护人们思想创造的这两项制度并没有给权利人以排他控制其信息或知识的权利。

⑩在法律制度史上亦有例外,欧盟1996年制定的《数据库法律保护指令》对于数据库在给予版权保护同时,给予数据库制作者以“数据库权”(database rights),这种权利核心是对抗他人非法提取和再利用数据库中信息内容的行为,被认为是针对信息内容的赋权。但是,该指令在欧盟的适用颇有争议:2004年欧洲法院在BHB v. William Hill等案的判决中明确地宣布了除非是从外部收集、核实、汇编形成的数据库,否则不应纳入数据库权范围进行保护,也就是说为了自己业务需要而自然形成的信息库(副产品)不受保护。这大大限缩了其适用范围,有学者认为这等于宣布“数据库权之死”。See Sarah Wright & Priya Vatvani,"Death of the database right", Copyrights

World(2005). 从国际上来说，1996年2月，欧盟在日内瓦会议上向WTPO提交了一份关于数据库法律保护协议的提案，在1996年12月外交会议上，WCT、WPPT两个条约的磋商几乎占据整个会议日程，因而对该数据库条约草案并没有进行实质性地讨论。之后因为美国的反对，该条约再没有进入WIPO知识产权条约的视野。这充分说明，到目前为止，国际社会对于在信息上设置排他性支配权的谨慎态度。

⑪例如通过身份证管理系统、新生儿管理系统，以及学生证、士兵证、社保、驾驶执照等各种系统，就可以获得几乎覆盖全部居民的基本数据，这在过去几乎不可能做到。

⑫万物互联(IoE)被定义为将人、流程、数据和事物结合在一起，使得网络连接变得更加相关、更有价值。在万物互联时代，梅特卡夫定律(即网络的价值与联网用户数的平方呈正比)更加受到关注。

⑬Ron Rock & Michael Moran,"A 21st Century Framework for Data Ownership"(https://microshare.io/wp- content/uploads/2018/06/21stCenturyFramework- June2018- ML- R3.pdf, last access 2018-02-25).

⑭Viktor Mayer- Sch{D412AE103.JPG}nberger & Kenneth Cukier, Big Data: A Revolution That Will Transform How We Live, Work, and Think(London: John Murray, 2013), p.29. 转引自：[美]莫里斯·E. 斯图克(Maurice Stucke)、艾伦·P. 格鲁内斯(Allen Grunes)：《大数据与竞争政策》，兰磊译，法律出版社2019年版，第21页。

⑮联合国经济和社会事务署统计局(Department of Economic and Social Affairs Statistics Division)设立的大数据工作组(Task Team on Big Data)，于2013年提出了大数据分类，将数据分为源于人的数据、源于组织运营的数据和源于机器的数据。参见Meeting of the Expert Group on International Statistical Classifications(New York: 19- 22 May 2015) (http://unstats.un.org/unsd/class/intercop/expertgroup/2015/AC289- 26.PDF, last access 2018-03-20)。

⑯实时数据(real-time data)又称“快数据”(fast data)，时间敏感性数据或数据的价值在于立即应用数据于决策。

⑰企业、社会组织等任何机构日常运营和活动中均会形成许多记录。企业的业务流程和运行中会形成注册客户、制造产品、接受订单等记录。业务流程数据通常是高度结构化的，既包括事务、参考表格和关系(transactions, reference tables and relationships)，也包括支撑其环境的元数据(metadata)。

⑱参见图图：《大数据时代下的商业变现(二)：从“原材料”到“粗加工”》(https://www.weiyangx.com/276979.html，最后访问时间2019-01-07)。

⑲OECD Data- Driven Innovation Big Data for Growth and Welt- Being,"Interim Synthesis Report" 4(October 2014) (http://www.oecd.org/sti/inno/data-driven-innovation-interim-synthesis.pdf, last access 2019-03-04).

⑳“衍生数据指原生数据被记录存储后经过算法加工、计算、聚合成系统可读取的数据，例如偏好数据、信用数据。”“原生数据就像原油一样，原油并不能直接被使用，它需要经过加工、提炼成汽油才能被使用。数据记录方对于合法收集的数据需要经过加工计算，产出为数据处理系统可读取的数据方可实现数据价值，而其间的数据采集、存储、计算、加工、管理等需要投入巨大的成本，因此数据记录者的合法权利应当得到立法认可。”参见陈小江：《数据权利初探》，载《法制日报》2015年07月11日第006版。

㉑根据《数据产品的前世今生》作者的分类，数据产品从最初的报表型(如静态报表、Dashboard、即席查询)，到多维分析型(OLAP等工具型数据产品)，到定制服务型数据产品，再到智能型数据产品、使能型数据产品等多种形态。参见老读悟：《数据产品的前世今生》(http://www.woshipm.com/pmd/76203.html，最后访问时间2019-03-12)。

㉒许多AI创制内容满足作品要件，可受著作权法保护，当无太多疑问，现在主要争论的问题是该作品的归属。在不承认AI系统为独立主体的情形下，其创制的内容只能归属于AI系统研发或运用者，而问题又在于研发或运用者是否有足够的智力投入。

㉓参见曹思龙：《冰与火之歌：数据分析的前世今生(二)》(http://www.woshipm.com/data- analysis/686309.html，最后访问时间2019-03-12)。

㉔参见图图：《大数据时代下的商业变现(一)：大数据浪潮》(https://www.weiyangx.com/276972.html，最后访问时间2019-01-10)；《大数据时代下的商业变现(二)：从“原材料”到“粗加工”》(https://www.weiyangx.com/276979.html，最后访问时间2019-01-10)；《大数据时代下的商业变现(三)：数据的“精加工”》(http://www.sohu.com/a/228934245_117965，最后访问时间2019-01-10)。

㉕见前注⑲。

㉖参见齐爱民：《数字文化商品确权与交易规则的构建》，载《中国法学》2012年第5期(信息权是一种相对独立的新型财产权)；龙卫球：《数据新型财产权构建及其体系研究》，载《政

法论坛》2017年第4期；程啸：《论大数据时代的个人数据权利》，载《中国社会科学》2018年第3期(数据企业对合法收集的包括个人数据在内的全部数据享有支配性财产权)；纪海龙：《数据的私法定位与保护》，载《法学研究》2018年第6期(提出数据文件所有权说)。

㉗《网络安全法》第42条规定："网络运营者不得泄露、篡改、毁损其收集的个人信息；未经被收集者同意，不得向他人提供个人信息。但是，经过处理无法识别特定个人且不能复原的除外。"依据该规定，向他人提供个人信息之所以要征得个人的同意在于该信息是可以用于识别该个人，一旦"无法识别"，那么就不再受该个人控制。欧盟《统一数据保护条例》(GDPR)的序第26段(Recital 26)也明确指出条例不适用于匿名化数据(anonymous data)。

㉘法律对个人数据的保护也被称为个人数据保护权(个人信息保护权)，这种保护权不同于支配权，主要是维护数据上主体利益，而不是财产权意义上的支配权。参见高富平：《个人信息保护：从个人控制到社会控制》，载《法学研究》2018年第3期，第86～103页。

The Theory of Data Production: A Fundamental Theory of the Right Allocation of Data Resource

Gao Fuping

Abstract: With the coming of a data-driven society, data has become the fundamental resource of society. However, data has always been considered to be in the public domain and hinder the data rights, which has puzzled the establishment of the order of data resource utilization. In this paper, it is argued that data, a digital record describing an object, is produced by a human being or by a machine other than is a natural existence. Data production(data collection), dataset creation(data assembling process) and data analysis(including AI) constitute the three processes for the realization of the data´s value. The first two behaviors are called data production, and the theory of data production is put forward. The data production theory differentiates data production from data analysis. It holds that the subject of raw data production is the raw material´s analyst. Therefore, the value should to be recognized and appropriate rights should be allocated to them to meet the production of datasets for various analytical purposes. At the same time, data production is also different from the data´s origin, which implies that when the data´s origin relates to a person, it does not mean the person produces that data. A person is a data producer only when he or she provides or creates data. This theory aims to provide a legal foundation for the right allocation for data usage as a resource. The legal order of data usage is constructed through the right allocation between the raw data producer, the dataset creator and the data analyst.

Key words: Data Production; Personal Data; Property Right Related to Data; Right Allocation of Data Resource