人工智能生物学在军事医学研究中的应用

中共中央政治局2018年10月31日下午就人工智能(Artificial Intelligence, AI)发展现状和趋势举行第九次集体学习。中共中央总书记习近平在主持学习时强调，**人工智能是新一轮科技革命和产业变革的重要驱动力量，加快发展新一代人工智能是事关我国能否抓住新一轮科技革命和产业变革机遇的战略问题。**要深刻认识加快发展新一代人工智能的重大意义，加强领导，做好规划，明确任务，夯实基础，促进其同经济社会发展深度融合，推动我国新一代人工智能健康发展。习主席同时强调，人工智能是引领这一轮科技革命和产业变革的战略性技术，具有溢出带动性很强的“头雁”效应。在移动互联网、大数据、超级计算、传感网、脑科学等新理论新技术的驱动下，人工智能加速发展，呈现出深度学习、跨界融合、人机协同、群智开放、自主操控等新特征，正在对经济发展、社会进步、国际政治经济格局等方面产生重大而深远的影响。加快发展新一代人工智能是我们赢得全球科技竞争主动权的重要战略抓手，是推动我国科技跨越发展、产业优化升级、生产力整体跃升的重要战略资源。

而生命科学产业进入新世纪以来，以分子设计、基因组学新技术、合成生物技术、生物大数据、基因编辑（CRISPR/Cas9）技术为核心的技术突破，也推动了以生命科学为支撑的生物产业深刻改革。同时带来了一些与人类生命健康更有本质联系的，如基因测序、精准医学、合成生物学、再生医学等产业的发展。随着高通量组学技术的蓬勃发展，生物医学正加速进入大数据时代。其中以基因组为代表的组学数据的积累速度史无前例。借助生物医学多组学、多层次的海量数据，人们可以从前所未有的广度和深度来研究生物体运行机制，这对深入探索人类疾病的分子机理、推动个体化精准医疗有重大意义。**借助人工智能强大的计算分析能力，医学研究创新或将进入新一轮的蓬勃发展。因此，将人工智能技术和生命科学研究结合起来，筹建人工智能生物学(**Artificial Intelligence Biology, AIB**)，迫在眉睫。**

一．基本判断

**现在是人工智能与生物医学研究创新结合的绝佳时机，对军事医学研究和应用有很强的助力作用。**

**一是国家战略的大力支持。**2017年，科技部在京召开新一代人工智能发展规划暨重大科技项目启动会，在会上，科技部高新司公布了首批国家新一代人工智能开放创新平台名单，分别依托百度、阿里云、腾讯、科大讯飞，建设自动驾驶、城市大脑、医疗影像、智能语音国家新一代人工智能开放创新平台。目前在医疗领域，腾讯觅影已在中国和十余家三甲医院共建联合实验室，并与上百医院达成合作意向；阿里健康发布了医疗AI系统“Doctor You”发布，让人工智能承担专业医生助手的角色，降低医生工作量。2018年3月，天坛医院、清华大学等20多家医院和大学的学者参加了基于解放军总医院急救大数据进行的数据实践与竞赛活动（Data-thon），以急诊患者的连续生命体征数据、实验室检查数据、影像数据和电子病历四个方面为基本数据采集对象，通过三年的采集已经获得33万例急诊分诊数据和3万多例住院观察数据，以此数据为“粮食”，喂给人工智能平台，提高急性疾病的预测准确性，以便更早干预。

**军事医学的研究具有特殊研发优势，军事医学主要面向战场和部队的特殊需要**，其研发投入和研发群体相对较小，多为应急需求；从公共安全和保密角度，军事医学研究不宜全盘委托给企业和社会机构，毕竟信息时代数据为王，数据逐渐成为占领战略制高点的重要支撑；但同时也不能因此束手束脚，应及早组织高水平科技团队入手，**充分利用人工智能技术积累和利用大数据优势，推动和加速现有研究进展，解决现有军事医学药物研发及需求响应进程中高强度高负荷的复杂认知瓶颈问题**。

**二是国际环境相对缓和友好。**与其他被长期垄断的学科不同，人工智能作为一门新兴的发展迅猛的技术，在历史积累上华人处于世界领先水平。2016 年，美国白宫发布的《国家人工智能研究与发展策略规划》（The National Artificial Intelligence Research and Development Strategic Plan）报告中提到，从 2013 年到 2015 年，SCI 收录的人工智能方向的论文里，涉及“深度学习”的论文数量增长了约 6 倍。报告同时不无忧虑地提到，“ 如果按照文章数量统计，美国已经不再是世界第一了。”根据Google软件工程师王咏刚的调查和统计，中国人/华人在人工智能领域里的大牛比比皆是，吴恩达、孙剑、杨强、黄广斌、马毅、张大鹏 ……随口可以说出一大串，他自己在Google的研究团队、微软研究院等地亲眼看到，到处是中国人、华人的面孔。另外通过人工智能科研贡献数据统计，在2006到2016年的时间段里，近两万篇最顶级的人工智能文章中，由华人贡献的文章数和被引用数，分别占全部数字的29.2%和31.8%。近十年，华人用五分之一左右的作者人数，平均贡献了三成的顶级AI研究文章和被引用数。从统计角度看，这已经是超出平均水平的科研贡献了。从2014年、2015年开始，中国人/华人已经处于人工智能研究的领先地位，占据了人工智能科研世界的半壁江山！中国在这一轮科技变革的浪潮中已经抓住了浪潮之巅。目前由百度(BAIDU)，阿里(Alibaba)，腾讯(Tecent)三家公司为主导的BAT中国人工智能产业已经形成了开放的人工智能生态，正不断深入中国老百姓的日常生活之中。所以，**以AI技术为切入方向，是目前我国科技发展和军事医学研究领域弯道超车的绝佳机会。**

**三是符合事务发展规律。**世界发展史上有过三次工业革命。18世纪中叶以来，人类历史上先后发生了三次工业革命，发源于西方国家及衍生国家，并由他们所创新所主导。第一次工业革命开创了“蒸汽时代”（1760-1840年），第二次工业革命进入了“电气时代”（1840-1950年），两次世界大战之后开始的第三次工业革命，更是引领了“信息时代”（1950-），而如今人工智能技术或可成为第四次科技革命的中坚力量，它在21世纪发动和创新，恰逢中国通过改革开放后四十年的追赶，第一次与发达国家站在同一起跑线上。同样的，生命科学事关人的健康和发展，已成为许多国家科技创新的关键领域，全球生命科学同时面临前所未有的发展机遇。2018世界生命科学大会10月27日在京开幕，中共中央政治局常委、国务院总理李克强作出重要批示：“中国正在全面实施创新驱动发展战略和健康中国战略，不断加强生命科学研究，重视发展生命科学领域的新技术新产业，加快培育发展新动能，积极推进国际交流与合作。”**将AI技术和生命科学研究强强联合建立人工智能生物学(AIB)，是符合科技不断发展的历史趋势的。**

1. **对策思考**
2. **AIB数据意识建立。**

海量复杂数据带来的挑战非常巨大：一方面要处理的生物医学数据维数更高、要求AI模型有更强的学习适应能力；另一方面生物医学大数据更加分散破碎，数据结构复杂，常常要整合不同类型的信息，如基因组、蛋白质组和临床研究的数据等。这些问题都对数据特征提取提出了更高的要求。因此在生物医学大数据时代，统计分析方法及数据挖掘方法都面临巨大的挑战。

在传统生物医学数据分析中，一般将各领域专家手工提取的特征作为模型输入，大数据情境中这种过于依赖专家知识的研究思路存在两个问题：1）生物医学数据的特征集往往依赖数据集，例如基因表达数据的特征依赖具体的组织器官，因此实际应用中要对不同数据集分别提取特征。2）人工提取特征依赖先验知识，很难提取到潜在的复杂特征，而这类特征往往对分类至关重要。数据特征提取已逐渐成为生物医学数据分析的瓶颈。

人工智能技术的核心在于采用各种智能学习算法学习数据结构，在于各个维度数据的整合与应用。就军事医学研究而言，传统的研究方法采集的数据本身结构不清晰、标注不完整，只能部分满足人工智能技术应用。而早日**引入人工智能数据思维，整合清洗旧数据，建立新数据的采集标准，是目前需要迫切解决的问题，也是现有人工智能技术与军事医学研究切实结合的第一个要求。**

对于生命科学研究一线的同志而言，即使不了解AI算法细节，也需要大致了解其框架，因为AI分析的第一步就是输入数据。就目前AI在医疗领域最成熟的应用——医疗影像为例，常规的检查手段(CT、核磁共振、X光扫描)会产生大量图像数据，而人工智能目前有诸多算法(例如Google知名的深度学习算法：卷积神经网络算法)能够应用于临床图像分类、关键目标发现和图片自动分割等方面。国际上已有研究团队将深度学习模型用于图像中肿瘤细胞的自动发现，准确率相比传统方法有7%的提升，对癌症自动诊断有重要意义。这一领域的成功应用，主要归功于有经验的医生事先提供大量成熟的判断，帮助训练人工智能机器，通过大量的图片-标注关系“教”会了机器去自动识别健康的医学图像、癌变的医学图像。虽然经验丰富的医生也可以完成类似的工作，并且准确率可能比机器更高，但是识别效率和培养成本上而言，人工智能初筛能够大大帮助提升工作效率。

1. **优势领域建立，以点带面建设完整学科体系。**

医疗数据和生物数据分析中现有的人工智能应用可归纳为以下三个方面：**一是人工智能在临床诊断中的应用**。疾病诊断是深度学习在医学上的主要应用之一。它基于患者的疾病相关数据，通过深度学习模型预测异常病变或发病风险，进行疾病的辅助诊断。自动化的疾病辅助诊断能更快地处理数据，能为医师提供参考，且其判断不易受到主观因素的干扰，在减轻医师工作负担的同时提升效率和诊断准确率。自动疾病诊断包括疾病诊断、疾病分类和疾病分级等方面。**二是人工智能在医疗影像中的应用**。医疗机构的医学图像产出量巨大，图像数据往往包含大量潜在信息。目前主要依靠人工判读分析，效率较低且能挖掘到的信息有限，无法充分利用数据资源。深度学习在图像处理领域的优异表现为医学图像的自动化处理提供了新方法。目前深度学习在医学图像中主要应用于临床图像分类，关键目标发现和图片自动分割等方面。**三是人工智能在组学数据中的应用**。相比医学问题，生物组学中问题模型复杂，数据量更庞大。目前，深度学习技术主要被用在蛋白质结构预测、测序数据处理方面。2018年11月，google旗下的DeepMind团队再下一城，在蛋白质三维结构预测领域取得重大进展，获得两年一度Critical Assessment of Structure Prediction(CASP, 素有“蛋白质结构预测奥运会”之称)大赛的第一名，成功预测了43种蛋白质种的25种蛋白质结构，远远高于第二名的成功3种的成绩。深度学习在生物数据分析中的优秀表现有望为探究生物序列、分子结构、与疾病的关联提供新思路、新技术。

**（三） 人才队伍建设。**

AI作为一门具有普遍应用价值的技术，亟需有既熟练掌握AI技术应用，又充分理解生物大数据的人将这一强大的技术工具运用到恰当的地方。就好比AI是倚天剑屠龙刀，需要正确的人把它用在正义的战斗中。**因此，针对军事医学研究领域，必须建设起一支有梯度、有深度、有广度的人才队伍。**

**有梯度主要指建设教授牵头、博士后主抓、博士硕士研究生实施细节的研发团队，**在能力上、年龄上形成可持续发展的团队，这也是国际上著名的生物研究团队、计算机领域研发团队的成功模式。**有深度一是指理论深度**，AI本身是一个技术门槛很高的技术，不明就里拿来主义是非常难运用到位的，例如百度无人车团队聘请人工智能专家吴恩达做首席科学家，关键就是在理论指导上给予支持，因此军事医学研究也需要类似的人工智能“大拿”对技术运用进行掌舵，迅速培养起自己的理论团队；**有深度二是指应用深度**，AI再强大也只是技术，需要军事医学研究切实的需求作为支撑使技术落地具象化，这一点上需要AI技术团队和生物研究团队充分的讨论和交流，干实验(数据分析类)和湿实验(生物类实验)相互配合交替进行。**有广度主要指达成前两条标准之后**，不断拓展业务领域，占领国际战略高地，争取建立一个世界一流团队，实现第四次工业革命中的弯道超车。

1. **人工智能生物学(AIB)筹建建议**

基于以上的形势判断和对策思考，现建议筹建新学科——**人工智能生物学**(AIB)。主要期望解决人工智能与生物学交叉结合的系统方法学问题，针对军事医学研究需求建立相应的人工智能技术平台。关于 AIB的结构化定义，大体可分为：基础层、技术层、应用层。基础层涵盖各类数据的积累、融和、打通；技术层旨在利用语音、图像识别技术，对非结构化数据进行分析和总结，借助机器深度学习，达到监测、判断等能力；应用层即结合军事医学应用的各种场景，辅助或直接解决实际需求，例如药物研发、智能设备等。

1. **方法学建立**

人工智能算法基础在其原生的经典领域，例如图像识别、自然语言处理等方向发展迅猛，现主要分为有监督学习、无监督学习、强化学习三大主流，其中有监督学习中的深度学习算法应用最为广泛。但是，在生物研究领域，人工智能算法的广泛落地依然颇具难度，仍旧需要具体问题具体分析，很难实现多种问题的统一建模或者迁移学习。**从方法学角度看，主要待攻克的技术有以下两点**：

1. **AI与生物学研究联合建模**。

AI算法本身是一个笼统的说法，内在含有大量分支，例如传统的需要专家提取特征的机器学习算法、现流行的需要大量人工标注数据的深度学习算法、最新提出的生成对抗学习算法，每种算法在不同类型的问题应用中优劣势共存，没有一个彻底占优势的方法。例如前文提及的DeepMind团队预测蛋白质三维结构，采用了以往模拟退火的算法思路和蛋白质扭转角两大特征来训练自己的神经网络算法，同样的密歇根大学蛋白质预测团队采用类似的卷积神经网络人工智能算法，但是类似算法的最终预测结果却差距很大（总体DeepMind占上风），同时两种模型擅长预测的蛋白质类型也不一样。DeepMind在预测E. Coli中的RRSP蛋白时的表现，甚至差于平均水平，远逊于密歇根的预测结果。这个例子体现出，针对同一个问题，采用同样的AI算法，只是模型细节不同，都可以产生如此巨大的差距。由此可见，AI与广泛的生物学研究结合上依然面临严峻挑战，**AI算法中的数学本质如何在生物研究中被生物学本质具体量化，如何形成有效的可求解模型，如何评价模型求得的解的准确性**，都有亟需深入挖掘的方法学问题。

1. **AI本身方法学的发展进步。**

AI技术空前繁荣的背后，需要清醒认识到，当前应用广泛的AI算法仍旧处于“弱人工智能”时代。它能够部分摆脱过去显式编程中的机械笨拙，帮助人类从繁杂的数据中整理出相互联系的关系，并以概率值描绘联系的紧密性；也能够帮助人类从一些既需要知识储备、又重复性重体力的工作中摆脱出来，例如观察医疗影像、筛选SNP、预测蛋白结构。**但是，AI技术应用依然有一个大前提：它能解决的问题在小规模上人类自己也能解决**。因为它本质是模仿人类智能，它能够解决的问题一定是原本人类可以解决的，它能够大幅提高工作效率、大量减少人力成本，却无法代替人类完成原本无法完成的工作，例如需要大量因果逻辑推理的法律工作、需要灵感一现的基础科学创新研究工作，这些现阶段的人工智能很难做到。

针对“弱人工智能”向“强人工智能”时代跳跃的问题，国内外也启动了众多的研究项目，**寻求通过仿生学研究、脑科学和类脑科学研究来为“强人工智能”打开思路**。美国相继提出“神经科学研究蓝图”计划(2005年)、 “通过推动创新性神经技术开展大脑研究”国家专项计划(2013年投资30亿美元)、《国家人工智能研究与发展战略规划》(2016年)；欧盟将“人脑工程计划”列入未来新兴旗舰技术项目(2013年)，10年投资10亿欧元。日本、韩国、加拿大等先后发布大脑发展战略和共识，艾伦研究所、谷歌公司、微软公司、百度公司等研究机构和企业，纷纷加入。中国脑计划也在2016年底启动，在《国家新一代人工智能发展规化》中，脑科学和类脑科学研究被认为是生物科学研究反哺人工智能底层设计创新的重要研究方向。

通往“强人工智能”的**另外一个思路，是模仿人的因果推理能力教会机器学会推理因果**而非仅仅基于大数据拟合出概率上的关联性。贝叶斯网络理论之父、人工智能领域的奠基人之一Judea Pearl认为，**人工智能发展的新关键在于用因果推理(causal reasoning)取代关联推理(reasoning by association)**。例如机器不仅要能将发烧和疟疾联系起来，还需要具备推理出是疟疾导致发烧的能力。一旦具备这种因果推理的机制，机器就有可能提出虚拟的问题——追问在有干扰的情况下，因果关系将如何变化：例如一个日常不运动的人，在开始偶尔运动后、形成运动习惯后会产生怎样的变化？提出这样问题的能力，是下一步人工智能的发展方向，也是帮助解决现代生物学研究问题的重大助力。

1. **技术平台搭建**

军事医学研究主要面向战场和部队的特殊需要，可在合成生物学、生物识别、智能穿戴设备、基因组学、药物设计、生物防御快速响应等领域充分利用人工智能技术推动和加速现有研究进展。

**(1) 面向合成生物学的人工智能创新研究**。

合成生物学目前有两个主要的研究目标，一是科学家们努力建立新的生命形式，试图从零开始从最基本的化学物质组装出新的生命，这当前还是一个相对遥远的目标，最新的进展在于能够从头设计与化学合成酿酒酵母，中国在这个领域已处于世界领先地位(2017年天津大学、清华大学和华大基因共同完成了四条真核生物酿酒酵母染色体)，但是从染色体合成到组装新生命还有很长的路要走；二是利用生物技术生产其他手段难以获得的物质，特别是药品的合成和生物燃料的研制。核心是基因信息的选择性组装：通过引进外源基因的组合和定向进化来优化代谢工程而得到需要的产物，例如通过大肠杆菌和酵母生产青蒿素，通过修改细胞注入宿主监控糖尿病患者的血糖水平，通过微生物消耗糖来产生油等生物燃料。

合成生物学中涉及大量的化学反应，构成整个生化反应网络，涉及到其自身的途径合成和热动力学可行性，还有途径流量及其控制。其主要动力学机理尚处于未知状态，尤其是参与反应的速率和其他参数，传统的仿真模拟是个解决方案，但在先验信息大量缺失时无法解决这个问题。而人工智能算法，是帮助从以往数据中学习反应参数的新思路。结合的技术难度在于：经验数据的整合与建模和从小样本数据中提取特征学习参数的算法进化。

**(2) 人工智能在生物识别中的技术应用**。

生物特征识别技术是通过对人的生理特征（如DNA、指纹、面像、虹膜、掌纹等）和行为特征（步态、声音、笔迹等）等生物特征的采集、分析和比对辨识个人身份的一类技术。生物特征识别在定点打击、信息安全、情报侦察、反恐等军事领域具有广阔的应用，是目前世界军事强国竞相发展的高技术领域。

识别个体身份是生物识别其中的一环，除此以外，根据前期提取的特征和数据，对个体进行行为判别是另外一重目标。正如NBA通过统计球员的各项数据来判别球员价值，生物识别技术发展中获得的大量生物特征数据，可以帮助评判战斗人员的作战能力、评判训练中学员的表现、或者判断手术中主刀医生的水平。李飞飞带领的斯坦福人工智能实验室帮助临床治疗研究中心(Clinical Excellence Research Center，CERC)开发了一套新的软件程序，可以测量外科医生的技术水平。它的工作原理是“观看”一段手术视频，跟踪手术过程中使用的仪器的运动和时间，根据经验训练判断能力。类似的，通过记录军事训练中的个体生物表现、运动轨迹，实现个体运动状态能力识别将不是难题。该项工作的技术难点，在于生物数据集与对应行为标记集的建立、单兵智能穿戴设备的设计、动态时序特征提取模型、以及多种人工智能算法组合建立并求解模型。

**(3) 面向基因组学的人工智能创新研究。**

研究生物体内完整基因组，是人工智能近年来显著发展一大领域。由于生物体在一生中经历的疾病在很大程度上取决于他们的遗传基因，因此多年来人们对基因构成有着浓厚的兴趣。但是相关研究因需要评估数据的复杂性和巨大性而停滞不前。随着人工智能应用技术的进步，研究人员能够通过基因组测序和基因编辑更好地解释和处理基因组数据。加拿大公司Deep Genomics使用其AI平台解码基因组的含义，可以根据细胞中的DNA信息来确定个体的最佳药物治疗方式。该公司的学习软件可以分析突变情况，并基于数十万个突变实例的内容来预测突变的影响。谷歌子公司Deep Varient也开放自己的基因组分析算法代码以及云端用于存储高通量测序数据，他们的工作基于深度卷积神经网络，能够进行SNP和small indel检测，目前的性能已经优于生物信息学中普遍采用的GATK分析软件。总体而言，基因组学中的测序数据积累雄厚，发展迅猛的方向集中在SNP检测、启动子增强子预测、蛋白序列与结构预测等序列和性状关系对应比较明确的领域。(待扩充)

**(4) 面向生物防御快速响应的人工智能创新研究。**

生物防御对于国家安全的重要性越来越突出。生物战和生物恐怖袭击具有成本低、危害大、隐蔽性强等特点。在当前国际形势下，发生大规模生物战的可能性在减小，但随着相关生物技术的不断扩散，在非战争状态下使用生物恐怖剂的可能性越来越大。随着生物技术、特别是合成生物技术的飞速发展，生物威胁的种类和数量在迅速增长，未来可能遭受的生物打击很可能并不包括在传统的生物武器清单中。依靠现有方法和技术，研究人员需要数以年计的时间才能了解新的生物威胁的损伤机理，而发现有效的治疗手段则需要十年以上的时间。如何快速应对“未知”生物武器袭击，在短时间内拿出防治方案，成为当前生物防御技术迫切要解决的现实问题。

基于人工智能的生物大数据分析技术使得生物防御的快速响应成为可能，从而有希望通过高通量生物数据采集和智能化特征计算分析，加速整个生物防御体系的运行效率。针对生物威胁快速反应技术体系建设的需求，开展以仿脑概念学习和小样本机器学习为基础的生物威胁快速响应研究；通过引入先验知识控制模型复杂度，解决在有限的样本经验支持下学习训练方案中的可用模型问题，从复杂和稀缺的生物医疗样本数据中进行快速有效的认知学习。

**(5)** **面向军用特需药物研发的人工智能创新研究。**当前，传统药物研发模式已遇到瓶颈，已不能满足人们的医疗卫生需求，对于要求更高的军特药研究更是显得力不从心。药物重定位策略被认为是一种最快捷、最有效、风险/效益比最好的新药研发策略之一，目前除各大跨国公司纷纷采用，也被多国军方重视。基于大数据的药物重定位新技术是解决军队卫勤保障的特殊需求，快速研发应急品种、突破难研品种等军特药研发瓶颈问题的重要和崭新的手段。因此，突破传统药物研发模式，建立基于大数据的药物重定位新技术是我军跟上科学发展和技术进步的步伐，形成崭新的新药研发模式和技术体系的重要举措，在此基础上，开展军用急需特需药物重大品种研发，不但对解决我军卫勤保障所面临的现实迫切需求具有重要意义，对于满足长远需求、提升军用特需药物研发能力更是具有重大意义。

药物重定位策略的广泛商业应用需要考虑知识产权和专利保护方面的非技术问题，但这一限制并不适用于军事医学应用。药物重定位策略不仅能为复杂疾病的快速药物发现提供低成本应急药物筛选方案，还能够通过模拟创新手段开发新药奠定基础，有效降低了药物开发的周期、成本和风险,进一步结合当前部队卫勤保障的现实紧迫需求和未来战争的重大需求，面向化生威胁应急对抗、海上战斗力保障、新质战斗力保障等军事医学需求，建立崭新的军特药研发模式。

在生物大数据背景下，人工智能技术将超越专家知识，快速给出可验证的药物重定位计算预测方案。**其中关键问题是如何解决在有限的样本经验支持下让算法能够学习到可用的模型，从常见小规模而不是海量大数据中通过机器学习方法找到内在规律，构建可用模型**。

以上述各个方向作为军事医学研究创新切入口，不断扩展AI技术的生物研究应用范围，努力建设完备的研究方法和学科体系。

1. **组织实施**