arXiv:2304.03674v2 [cs.LG] 2 Feb 2024

机器学习与需求：一份宣言

Eleonora Giunchiglia [a,*]，Fergus Imrie [b]，Mihaela van der Schaar [c,e] 和 Thomas Lukasiewicz [a]，[d]a奥地利维也纳工业大学逻辑与计算研究所 电子邮箱：*eleonora.giunchiglia@tuwien.ac.at,thomas.lukasiewicz@tuwien.ac.at*[b]美国加利福尼亚大学洛杉矶分校电气与计算机工程系 电子邮箱：*imrie@ucla.edu*[c]英国剑桥大学 *DAMTP* 电子邮箱：*mv472@cam.ac.uk*[d]英国牛津大学计算机科学系[e]英国艾伦·图灵研究所

摘要。 近年来，机器学习在不同应用领域取得了巨大进展，并成为许多突破的根源。然而，如何使其适用于高风险或安全性关键的应用领域仍然是一个悬而未决的问题，因为它们往往可能脆弱且不可靠。在本文中，我们论证了需求定义和满足能够大大提高机器学习模型对现实世界的适用性，尤其是在关键领域。为此，我们提出了两个问题，其中（*i*）需求自然产生，（*ii*）机器学习模型已经或能够得到富有成效的部署，并且（*iii*）忽视这些需求可能导致灾难性的后果。我们展示了如何将需求规格说明富有成效地整合到标准机器学习开发流程中，并提出一种新颖的金字塔式开发过程，在该过程中，需求定义可能影响流程中的所有后续阶段，反之亦然。

关键词：机器学习、需求、机器学习运维、软件工程

1. 引言

近年来，机器学习取得了巨大进展，成为不同应用领域众多突破的根源。例如，AlphaFold [1] 是一种深度学习模型，解决了“蛋白质折叠问题”——这是生物学领域半个多世纪以来的一项重大挑战；而 halicin [2] 则是首个借助机器学习发现的抗生素，有望帮助应对细菌耐药性。尽管上述成果具有开创性，却掩盖了在关键应用中不慎使用机器学习模型所带来的危险。事实上，即使此类模型在准确率（或另行选定的指标）方面表现惊人，也无法保证模型在实际使用时不会出现任何意外行为。正如 Rudin 在 [3], 中指出的，已经出现过一些案例：有人被错误地拒绝假释 [4],；用于监测美国空气质量的深度学习模型声称污染严重的空气可以安全呼吸 [5] ；以及在医疗、刑事司法、金融和其他领域普遍存在对有限资源的不当使用 [6]。模型行为中的这类问题往往根源于质量】【。

*通讯作者。电子邮件：eleonora.giunchiglia@tuwien.ac.at。

用于训练模型的数据集。因此，例如，在 COVID-19 疫情早期就出现过这种情况：正如 [7], 所报道的，来自 [8] 的一个常见胸部扫描数据集由儿童扫描图像构成，被用作与 COVID-19 阳性扫描图像进行对比的对照组；结果，构建出来的不是新冠肺炎检测器，而是成人/儿童分类器。无论原因如何，这类意外结果（*i*）在安全关键应用中尤其危险，因为哪怕一个未预见到的错误也可能导致严重后果；（*ii*）还会削弱人们对模型本身的信任，从而减缓其应用。

本文认为，需求规格说明和验证能够大幅提升机器学习模型对现实世界的适用性，降低出现潜在危险的非预期行为的风险。我们的出发点是一个简单的观察：非预期行为意味着模型违反了某些需求，而这些需求可能在数据收集和模型开发开始之前就已明确。为支持这一观点，我们考察了两个例子，一个来自医疗保健领域，另一个来自自动驾驶领域。在这两个例子中，(*i*) 某些需求是预先已知的；(*ii*) 机器学习模型在准确率（或其他选定指标）方面表现非常出色；(*iii*) 尽管表现出色，按照常规方式开发的机器学习模型，其结果往往仍会违反明示的需求，而考虑到应用领域的关键性，这可能造成严重后果。尽管我们只考察了两个例子，但我们认为，鉴于以下几点，类似的考量适用于大多数应用领域：(*i*) 多年来各应用领域积累的知识可以转化为相应的需求；(*ii*) 机器学习模型取得了令人瞩目的性能成果，尽管可能出现非预期行为，这些成果仍将不断推动其应用；(*iii*) 如果在数据收集和模型开发过程中不考虑需求，那么最终模型违反需求并不令人意外。进一步说，如果在部署之前，没有以某种方式依据刻画预期行为的需求对数据和模型本身进行验证，那么机器学习模型出现非预期行为也不足为奇。因此，我们主张，需求定义应先于整个机器学习模型开发周期，并贯穿其中，包括数据集构建。具体而言，我们提出一种新颖的金字塔式机器学习开发过程，它(*i*) 强调需求定义如何帮助改进标准机器学习模型开发过程中的每一个步骤；(*ii*) 同时考虑到，机器学习模型开发过程中出现的困难和／或新发现可能使某些需求有必要进行调整，尤其是鉴于机器学习模型的数据驱动特性。上述提议可以视为呼吁调整并采用软件工程中用于通用系统开发的一般方法论。在这一领域，众所周知：(*i*) 越晚考虑需求，修复系统的成本就越高；(*ii*) 需求定义是一个迭代过程，需求会影响系统开发，反之亦然，因为在系统开发过程中，某些需求可能会被新发现、调整，甚至取消，参见例如 [9–11]。我们将在具体的机器学习模型开发流程中阐明上述事实，并突出在不同阶段采用需求的优势。

我们的提案代表了对传统“性能驱动”机器学习开发流程的一次重大转变；传统流程只专注于如何获得更好的性能，以准确率或其他选定指标来衡量。事实上，我们提出的是一种“需求驱动”机器学习开发流程，在该流程中，性能只是模型必须满足的众多需求之一。因此，我们的提案与近期大多数相关工作是一致的：这些工作不仅关注性能，还会检查模型是否满足公平性、鲁棒性、可解释性、可持续性和安全性等其他属性。显然，这可能会使开发过程更加复杂，因为不同需求之间可能相互矛盾：这是软件工程中一个广为人知的问题 [9]。例如，在较高的抽象层次上，复杂模型很可能具有更好的性能，但代价可能是可解释性和/或可验证性和/或可持续性较差。类似地，在某些情况下，相较于更高的性能，满足特定的公平性属性（例如，均等化赔率或人口统计平等）可能更为重要。无论我们希望模型具备哪些属性，有一点是明确的：（*i*）越早明确需求越好；（*ii*）在数据集定义和模型构建中考虑这些需求，很可能会得到更易于根据既定需求进行验证的模型。为此，神经符号领域的近期研究提供了一个很好的例子：研究人员已经开始开发一些模型，这些模型在设计上就能保证符合一组以逻辑约束表达的给定需求（参见例如 [12–14]）：这些研究表明，如何将需求纳入模型，使模型能够自动满足这些需求，并且还能对性能产生积极影响。



图 1. 标准机器学习流程的可视化 .

我们的提案与 [15–17] 倡导采用更结构化的方法进行机器学习模型开发的作品秉持着相同的精神。具体而言，Gebru 等人在 [15] 中提出了针对新数据集的文档编写指南；Mitchell 等人在 [16] 中倡导对模型进行标准化报告，其中包括训练数据、性能指标和局限性；而 Seedat等人则在 [17] 中提出了一种可操作的清单式框架，用于在开发流程的不同阶段引出以数据为中心的考量。

本文其余部分的结构如下。首先，在第 2 节中，我们考察医疗保健和自动驾驶应用领域的示例，以说明在这些领域中，*i*需求会自然产生，*ii*机器学习模型已得到或能够得到有效部署，且 *iii*忽视需求可能造成严重后果。其次，在第 3 节中，我们展示如何将需求定义有效地融入标准的机器学习开发流程，并影响流程中的所有阶段。最后，第 4 节总结全文，并探讨未来可能的研究方向。

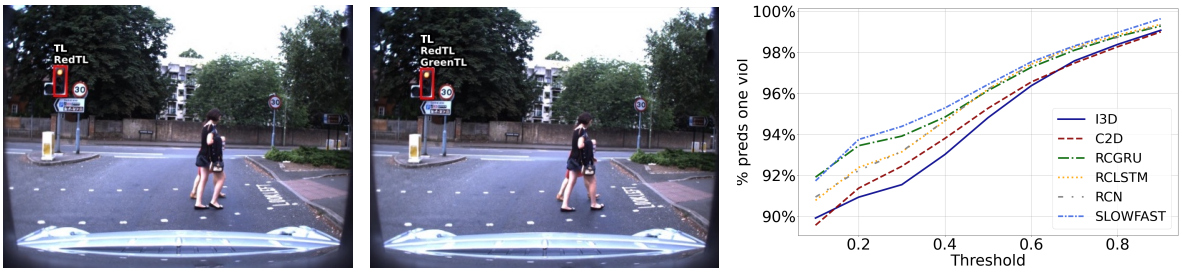
2. 需求的必要性

标准的机器学习流程本质上大多是顺序进行的，其主要步骤包括：

1. **数据整理**，其中创建用于训练模型的数据集，涵盖以下阶段：(*i*) 数据收集、(*ii*) 数据预处理、(*iii*) 数据增强，以及(*iv*) 数据质量评估；2. **模型创建**，其中设计并构建模型；3.**模型训练**，其中对所创建的模型进行训练，并选择合适的超参数（通常通过在验证数据集上进行测试）；4.**模型测试**，其中评估模型的性能（理想情况下使用不同的指标和测试集）。5. **模型部署**，其中模型实际部署到现实世界中。

标准顺序流水线的可视化表示如图1所示。尽管图中没有展示，但在实践中，可能需要回溯，以便修改数据集，或者更常见的是修改模型。不过，这些可能的变更大多发生在模型创建步骤与模型训练步骤之间，因为它们通常是由提升性能的愿望所驱动的，而数据集往往被视为流水线中的静态元素。鉴于人们常常认为机器学习模型唯一需要满足的需求，就是根据某些选定的指标取得良好性能，这种情况并不令人意外。然而，正如引言中所强调的，取得高分往往不足以确保模型在实践中使用时表现出恰当的行为。为了使模型能够适用于特定应用领域，甚至获得在该领域应用的认证，可能还需要满足其他需求。

现在，我们聚焦于两个应用领域——医疗保健和自动驾驶，并展示需求如何自然地出现在这两个应用领域中，应用机器学习技术如何已经为这些领域带来并且还能带来巨大优势，以及在这些领域部署机器学习模型时若不明确考虑需求，如何可能导致意料之外的行为，进而违反需求。尽管我们只考察了两个应用领域，但我们相信，鉴于（*i*）这些年来在各应用领域积累的知识体系可转化为相应的需求，（*ii*）机器学习模型取得的卓越性能成果将继续推动其应用，尽管它们可能产生非预期行为，以及（*iii*）若不明确阐明相应的需求，并针对这些需求对模型进行测试和验证，就不可能证明不存在不良行为或排除其发生，因此，类似的考量和结论几乎适用于任何应用领域。



(a) (b) (c)

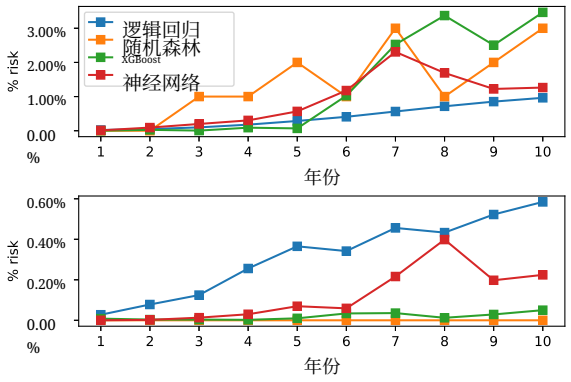
图2。图2a和图2b展示了 I3D模型（采用 $\theta = 0.5$）针对同一交通信号灯、仅相隔一帧所作的预测。图2c展示了改变 θ 时违反至少一项需求的预测数量。图2c取自 [14]。

2.1. 自动驾驶

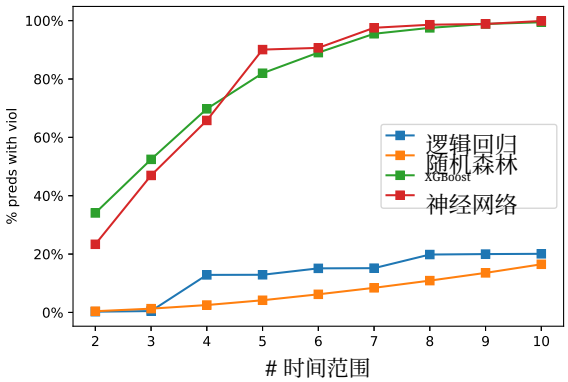
近年来，机器学习领域，尤其是计算机视觉领域的发展，燃起了人们的希望，让自动驾驶车辆最终成为现实似乎指日可待。然而，这样的希望时常被车祸击碎；有些事故甚至导致人员受伤或死亡。例如，2018年3月，优步先进技术集团开发并在公共道路上测试的一辆自动驾驶车辆，撞死了一名推着自行车在非指定人行横道处过马路的行人 [18]。这起事故最令人震惊之处在于，汽车完全没有尝试刹车和／或避让行人 [19]。遗憾的是，这并非孤立事件：在短短十个月内，美国报告了近400起涉及自动驾驶车辆的车祸 2022 [20]。正如 [20],所报告的，这些车辆在做出决策时会依赖计算机视觉模型等技术；如果这些模型仅以最大化性能为训练目标，并遵循标准的机器学习开发流程，就可能连最基本的需求都无法满足。

为了说明上述问题，我们考虑 ROAD-R 数据集 [14], ，该数据集由以下内容组成：(*i*) 22 段相对较长的视频（每段 ∼8 分钟），这些视频标注了道路事件，即一系列按时间关联的逐帧边界框，每个边界框均标注了其中的主体，以及其动作和位置；(*ii*) 243 条以命题逻辑表达的需求，用于说明哪些道路事件是允许的。[1] 与每个边界框关联的可能标签有 41 个，这些需求规定了模型可以输出哪些标签组合。例如，一条需求规定交通信号灯不能同时为红色和绿色，另一条则规定交通信号灯不能移动。基于 ROAD-R，研究人员训练了六种最先进的时序特征学习架构（I3D [21], C2D [22], RCGRU [23], RCLSTM [23],RCN [24] 和 SlowFast [25]），将其作为用于事件检测的 3D-RetinaNet模型[26]（采用由 Resnet50 [27]构成的 2D-ConvNet骨干网络）的组成部分。这些模型以一组连续帧作为输入，并对每一帧输出 (*i*) 一组边界框，以及 (*ii*) 每个边界框对应的一组标签。标签按照标准方式确定：对于 41 个标签中的每一个，模型都会输出一个数值 $o \in [0, 1]$；如果 $o > \theta$，则返回该标签，否则不返回。 θ 是用户定义的阈值，通常取 0.5。图 2a 给出了一个预测示例，其中所示边界框的预测结果为 {交通信号灯，红色交通信号灯}。该研究 [14] 的惊人发现是，如图 2c 所示，无论选择何种阈值 θ，至少 89% 的预测都会违反至少一条需求。更进一步，有些预测可能违反人类拥有的常识所对应的需求，使得这些预测难以解释和处理。例如，考虑图 2b 中的预测：该预测由与图 2a 相同的模型做出，针对的是仅晚一帧的同一盏交通信号灯。由此可见，预测不仅有误，还违反了交通信号灯不能同时为红色和绿色这一常识（在本例中，这一常识也对应一条正式表述的需求）。如果控制车辆的系统不对这种预测作进一步处理，它确实可能造成严重后果。因此，

[1] Dat数据集地址：https://github.com/gurkirt/road-dataset。需求地址：https://github.com/EGiunchiglia/ROAD- R.



（a）不同时间范围内的预测示例。



（b）存在 $\geqslant 1$ 违规的患者比例。

图 3。图 3a 展示了两个示例，说明不同分类模型在不同时间范围内对同一患者所作的预测。图 3b 展示了预测违反“时间范围越长，风险越高”这一需求的样本比例。

显然，任何此类预测都必须由控制车辆的系统专门处理，而该系统必定会纳入并满足被违反的需求。然而，如果整个系统都纳入了这些需求，并且确实对其进行了验证（在这种情况下本就应当如此），那么为何在构建系统的机器学习组件时通常会忽略这些需求，就很难说得通了。事实上，所评估的六个系统都无法处理关于其输入输出的需求。更重要的是，正如 [14], 所报告的，用于训练模型的数据集违反了一些约束——例如，车辆不可能同时处于驶入和驶出状态——而这些约束必然需要纳入模型的任何实际应用中。[2]

2.2. 医疗保健

作为第二个示例，我们讨论医疗保健。人们对机器学习的发展将彻底改变医学并转变临床实践寄予厚望 [28]。医学中的应用范围十分广泛，从分析放射学医学影像的计算机视觉系统、住院期间对患者轨迹的长期监测，到未来疾病风险的基因组筛查，等等。然而，与许多其他高风险和安全关键的机器学习应用一样，医疗保健领域的机器学习系统也有许多需求。这些需求包括伦理方面的考量，例如公平性和偏见；实践方面的考量，例如控制假阳性率以防止警报疲劳 [29] 并确保合理分配资源；以及逻辑方面的考量，例如下文讨论的示例，等等。

作为一个具体示例，我们考虑临床风险评分。临床风险评分用于估计某一特定结局在一段时间后发生的可能性，例如患者在未来十年内罹患某种疾病或病症，或在医疗操作后出现不良事件。根据定义，某一结局在特定时间范围 t 内发生的概率必须随时间增加，因此风险评分必须是关于时间的单调递增函数，也就是说，对于所有 $n > m$，预测的 n 年风险都高于其 m 年风险。然而，许多研究使用分类模型来预测特定时间范围后的风险，例如 5 年风险。显然，我们可以将这种方法推广到多个时间范围，而非只预测一个时间点的风险。从临床角度看，这通常非常有用，因为患者的病程可能存在很大差异，长期风险相同的患者，其短期风险可能截然不同。一种直接的做法是训练多个分类模型，分别预测每个感兴趣时间点的风险。然而，无法保证预测风险会随着时间范围延长而增加。

[2]虽然有人可能会认为，数据集因噪声而存在错误是正常的，但先用已知错误的数据训练模型，再在模型输出符合这些已知错误数据时予以纠正，这似乎有些奇怪。

为说明未将上述简单需求纳入机器学习系统会造成什么后果，我们考虑预测糖尿病患病风险这一问题。我们从英国生物样本库 [30], 中构建了一个患者队列。该大型观察性研究于 2006 年至 2010 年间招募了约 50 万名参与者，他们来自英格兰、威尔士和苏格兰的 22 个评估中心。我们提取了一个参与者队列，其中所有人在入组时均年满 40 岁，且基线时没有糖尿病诊断或病史。我们采用了 QDiabetes [31] 使用的 18 项特征。我们使用 HyperImpute[32] 进行数据插补，并使用 AutoPrognosis [33]训练分类算法，以预测n 年糖尿病患病风险，其中$n \in \{1, \ldots, 10\}$。我们对五次随机初始化的结果取平均值。图 3a 展示了两个示例。对于第一位患者（图 3a，上图），只有逻辑回归模型满足糖尿病患病风险应随时间范围延长而单调递增这一要求；随机森林、XGBoost 和神经网络模型的预测中，至少有一个时间范围的风险低于前一个时间范围。第二位患者（图 3a，下图）也有类似情况，只有随机森林模型满足这一要求。考察整个数据集和全部十个时间范围后发现，逻辑回归和随机森林模型中，约有 20% 的患者至少一次出现风险预测未能单调递增的情况；而对于 XGBoost 和神经网络模型，超过 99% 的患者都出现了违反这一要求的预测。需要注意的是，模型违反要求的次数与其性能不一定相关：以 5 年时间范围的受试者工作特征曲线下面积衡量，性能最低的模型（随机森林）反而违规次数最少。这类预测即便最好的情况也只是准确性不足，最坏的情况下则可能削弱人们对机器学习系统的信任，并根据后续采取的行动造成更严重的后果。满足这一特定要求有多种解决方案，例如采用专门设计的函数形式 [34, 35], 、使用专门损失函数的多任务学习 [36],，或定制架构。尽管此类预测在这一特定示例中对现实世界的直接影响可能有限，但问题本身以及对需求的必要性显而易见。

3. 金字塔模型

在上一节中，我们展示了标准的性能驱动机器学习开发流程，其中需求通常会被忽视；同时，我们通过两个例子强调了这种疏忽可能带来的严重后果。因此，开发过程必然需要纳入需求，而需求正是软件工程领域提出的任何开发流程的核心。由此，我们建议调整软件工程中用于通用系统开发的一般方法论；众所周知，需求被纳入考虑得越晚，修复系统的成本就越高。为了调整这类方法论，我们需要考虑到，与标准软件不同，机器学习模型是从数据中学习得到的，因此我们对模型行为的控制程度较低。这不仅意味着在设计阶段可能无法预测模型未来的行为，而且在测试阶段，如果发现了缺陷，那么修复该缺陷所需的步骤也完全不同。因此，机器学习模型的开发流程需要考虑以下几点：

1. 数据是这一过程的核心，因为模型的质量在很大程度上取决于所收集数据的质量；2. 模型是从数据中学习得到的，因此无法事先完全了解其行为；3. 如果模型的行为不符合预期，建议不仅检查模型，也检查数据。

此外，不能不切实际地期望所有需求都能事先明确，并且在整个开发过程中始终保持不变。这一点在通用软件工程领域早已为人所知，而且也已在我们这一领域得到专门验证。事实上，正如 [37],开展的调查所指出的，机器学习从业者每天面临的一些最大难题包括数据的可获取性——这在需求定义阶段可能尚不明确——以及系统利益相关者不切实际的期望——这些期望可能会在模型的测试阶段显现出来。因此，我们还需要考虑到，需求可能会在开发过程的每个阶段发生变化。

鉴于上述内容，我们提出图4中的金字塔机器学习开发模型。在该模型中，需求定义阶段如预期般被置于开发过程的起始位置，而金字塔形状则更清晰地凸显了需求定义阶段与

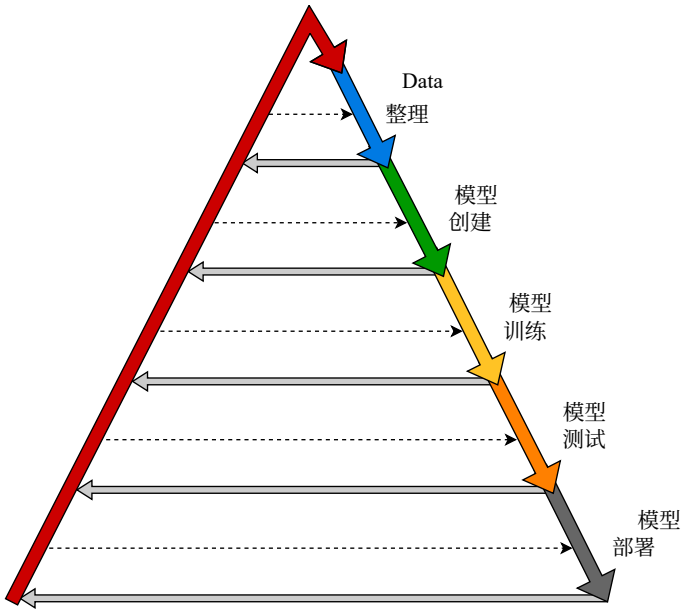


图 4. 金字塔模型的可视化。实线箭头表示标准流程，虚线箭头则表示需求会影响流程的每个阶段。

流程中所有其他阶段之间必须存在的紧密关系。请看灰色的反向箭头。这些箭头表示，如果流程中的任何环节发生意外情况（例如，模型在测试阶段确实出现了某种非预期行为），就有必要回头检查已定义的需求是否合适或需要更新，然后检查所收集的数据是否能代表我们想要捕捉的现象，依此类推。显然，我们越晚在流程中发现缺陷，修复成本就越高；图4通过流程推进时需要走过的路径长度体现了这一点。再看虚线箭头，它们说明需求有助于塑造每个开发步骤，尤其是：

1. **数据整理**：有关数据的需求显然定义了数据收集阶段，因为这些需求说明了数据集应具备哪些属性。这可以对该阶段产生积极影响，因为它会促使系统的利益相关者仔细思考数据集应涵盖总体中的哪些方面，以及哪些方面应予以忽略。此外，这些需求还有助于减少标注阶段的错误——任何不符合需求的标注都必须予以更正——甚至还能加快标注过程：可以只向标注人员展示符合需求的选项，也可以实时自动更正标注。

2. **模型创建**：有关模型的需求定义了模型应如何运行，因此在创建模型时应将其纳入考虑。遵循数据整理阶段所采用的相同原则，明确需求可以对这一阶段产生积极影响，因为这会促使系统的利益相关者思考模型的预期行为，以及必须始终避免哪些结果。随后，这些需求可以映射到模型本身。例如，给定一组形式化需求，神经符号领域中的一些研究 [12, 13, 38] 能够将需求直接映射到神经网络的拓扑结构中，并保证始终满足约束，同时利用需求所表达的背景知识，进一步提升性能。

3. **模型训练**：对模型的需求同样可以用来定义训练目标。例如，公平性领域的许多研究便是如此：模型使用两个目标进行训练，一个用于最大化性能，另一个用于最大化所需的公平性定义（参见例如 [39]）。此外，如果我们将需求映射到损失函数中，甚至可以在半监督设置下减少对标注数据的需求（参见例如 [40–42]）。实际上，对于给定的未标注数据点，我们可以教模型只需返回符合需求的输出。

4. **模型测试**：正如标准软件工程中所做的那样，需求应指导测试阶段。每项需求都应检查其是否成立（最终可使用形式化验证技术，参见例如， [43,44]），如果不成立，那么要么需要进一步分析该需求并最终修改，要么必须制定某种程序，以发出信号表明该需求必须在机器学习模型之外得到妥善处理。

5. **模型部署**：需求也可以影响部署阶段，尤其是当利益相关者希望模型即使在存在数据变化和/或漂移时仍保留某些属性。

4. 结论与未来展望

本文表明，尽管将机器学习应用于特定问题能够带来巨大益处，但如果不明确说明并考虑问题的需求，就可能导致意料之外的行为，甚至造成严重后果。因此，我们提出了一种新的机器学习开发流程，在流程之初明确纳入需求定义阶段，从而使需求定义与开发过程中的所有其他阶段紧密相连；与软件工程中的常规做法一致，前者可能影响后者，反之亦然。从这一角度来看，本文可以视为呼吁将软件工程中用于通用系统开发的方法推广至机器学习这一特定领域，并强调不这样做所带来的风险。我们认为，在许多情况下，明确界定需求并在其他阶段加以考虑所带来的益处，超过了相应成本。当机器学习模型作为独立系统用于必须满足严格要求的领域（例如安全攸关领域），或作为具有一系列严格要求（例如内存使用量，见 [45]）的大型系统的一部分时，需求的定义与落实显然不可或缺。机器学习领域的最新趋势也支持这一观点：性能只是需要满足的诸多需求之一。具体而言，可解释性（见 [46, 47]）、公平性（见 [48, 49]）和鲁棒性（见 [50, 51]）领域的研究都优先考虑模型的某些特性，而非性能；这些研究的动机都源于必须考虑各自应用领域所产生的需求。然而，如何在开发过程中定义需求，并纳入和充分利用需求所表达的领域知识，仍是一个悬而未决的问题，因为还有许多技术挑战有待克服。为此，神经符号领域可以视为标准机器学习与软件工程之间的桥梁，因为该领域的研究人员已经开始开发相关技术，以构建不仅由数据驱动、还明确将其他类型知识（如形式化表述的需求）作为输入的模型，并利用这些知识来设计模型（见 [12, 13]）或将需求纳入损失函数（见 [41, 42]）。

尽管大量工作都可以归入“需求驱动的”机器学习这一范畴，但这是首篇明确倡导一种通用模型开发方法论的论文。在这种方法论中，需求定义与流水线中的其他阶段一样，都是核心环节。我们认为，这体现了一种更具结构化的机器学习模型开发方法，而这种方法已在 [15–17]中得到倡导。

参考文献

[1] A.W. Senior 等，《利用深度学习的势能改进蛋白质结构预测》，*Nature* **577**(7792)（2020年）。ISBN 1476-4687。[2] J.M. Stokes 等，《用于抗生素发现的深度学习方法》，*Cell* **180**(4)（2020年）。[3] C. Rudin，《停止为高风险决策解释黑箱机器学习模型，改用可解释模型》，*Nature Machine Intelligence***1**(5)（2019年）。[4] R. Wexler，《当计算机程序让你身陷囹圄：计算机如何危害刑事司法》，*The New York Times* （2017年）。https://www.nytimes.com/2017/06/13/opinion/how-computers-are-harming-\criminal-justice.html。[5] M. McGough，《萨克拉门托的空气究竟有多糟？有人称，Google结果似乎与现实不符》，*Sacramento Bee* （2018年）。https: // www.sacbee.com/news/california/fires/article216227775.html。[6] K.R. Varshney 和 H. Alemzadeh，《论机器学习的安全性：网络物理系统、决策科学与数据产品》，*Big data***5**(3)（2016年）。

[7] M. Roberts、D. Driggs、M. Thorpe、J. Gilbey、M. Yeung、S. Ursprung、A.I. Aviles-Rivero、C. Etmann、C. McCague、L. Beer、J.R. Weir- McCall、Z. Teng、E. Gkrania-Klotsas、A. Ruggiero、A. Korhonen、E. Jefferson、E. Ako、G. Langs、G. Gozaliasl、G. Yang、H. Prosch、J. Preller、J. Stanczuk、J. Tang、J. Hofmanninger、J. Babar、L.E. Sánchez、M. Thillai、P.M. Gonzalez、P. Teare、X. Zhu、M. Patel、C. Cafolla、H. Azadbakht、J. Jacob、J. Lowe、K. Zhang、K. Bradley、M. Wassin、M. Holzer、K. Ji、M.D. Ortet、T. Ai、N. Walton、P. Lio、S. Stranks、T. Shadbahr、W. Lin、Y. Zha、Z. Niu、J.H.F. Rudd、E. Sala、C.-B. Schönlieb 和 AIX-COVNET，《使用机器学习通过胸部X线片和CT扫描检测COVID-19并进行预后判断时的常见陷阱与建议》，Nature Machine Intelligence 3(3)（2021年），199–217。[8] D. Kermany、M. Goldbaum、W. Cai、C. Valentim、H.-Y. Liang、S. Baxter、A. McKeown、G. Yang、X. Wu、F. Yan、J. Dong、M. Prasadha、J. Pei、M. Ting、J. Zhu、C. Li、S. Hewett、J. Dong、I. Ziyar 和 K. Zhang，《通过基于图像的深度学习识别医学诊断和可治疗疾病》，Cell 172（2018年），1122–1131.e9。[9] I. Sommerville 和 P. Sawyer，《需求工程：良好实践指南》，Wiley，1997年。[10] I. Sommerville，《软件工程》，第10版，国际计算机科学丛书，Addison-Wesley，2015年。[11] B. Nuseibeh 和 S.M. Easterbrook，《需求工程：路线图》，载于：ICSE论文集，A. Finkelstein 编，ACM，2000年，第35–46页。[12] E. Giunchiglia 和 T. Lukasiewicz，《带硬逻辑约束的多标签分类神经网络》，JAIR 72（2021年）。[13] N. Hoernle、R. Karampatsis、V. Belle 和 K. Gal，《MultiplexNet：迈向神经网络中完全满足的逻辑约束》，载于：AAAI论文集，2022年。[14] E. Giunchiglia、M.C. Stoian、S. Khan、F. Cuzzolin 和 T. Lukasiewicz，《ROAD-R：用于基于需求学习的自动驾驶数据集》，Machine Learning Journal（2023年）。[15] T. Gebru、J. Morgenstern、B. Vecchione、J.W. Vaughan、H.M. Wallach、H.D. III 和 K. Crawford，《数据集数据表》，CoRR abs/1803.09010（2018年）。http://arxiv.org/abs/1803.09010。[16] M. Mitchell、S. Wu、A. Zaldivar、P. Barnes、L. Vasserman、B. Hutchinson、E. Spitzer、I.D. Raji 和 T. Gebru，《模型报告模型卡片》，载于：FAT*论文集，danah boyd 和 J.H. Morgenstern 编，ACM，2019年，第220–229页。[17] N. Seedat、F. Imrie 和 M. van der Schaar，《DC-Check：指导可靠机器学习系统开发的以数据为中心的人工智能检查清单》，CoRR abs/2211.05764（2022年）。doi:10.48550/arXiv.2211.05764。[18] NTSB，《初步报告：HWY18MH010号公路》，2018年，华盛顿特区：美国国家运输安全委员会。[19] C. Macrae，《从自主与智能系统的失败中学习：事故、安全性与社会技术风险来源》，Risk analysis：Society for Risk Analysis 的官方出版物 42（2022年）。[20] N.E. Boudette、C. Metz 和 J. Ewing，《特斯拉自动驾驶辅助系统及其他驾驶辅助系统与数百起碰撞事故有关》，《纽约时报》（2022年）。https://www.nytimes.com/2022/06/15/business/self-driving-car-nhtsa-crash-data.html。[21] J. Carreira 和 A. Zisserman，《何去何从，动作识别？一种新模型与Kinetics数据集》，载于：CVPR论文集，2017年。[22] X. Wang、R. Girshick、A. Gupta 和 K. He，《非局部神经网络》，载于：CVPR论文集，2018年。[23] Y. Hua、Z. Zhao、Z. Liu、X. Chen、R. Li 和 H. Zhang，《基于长短期记忆深度学习中随机连接的交通预测》，载于：VTC-Fall论文集，2018年。[24] G. Singh 和 F. Cuzzolin，《因果三维卷积神经网络的循环卷积》，载于：ICCV研讨会论文集，2019年。[25] C. Feichtenhofer、H. Fan、J. Malik 和 K. He，《用于视频识别的SlowFast网络》，载于：ICCV论文集，2019年。[26] G. Singh、S. Akrigg、M.D. Maio、V. Fontana、R.J. Alitappeh、S. Saha、K.J. Saravi、F. Yousefi、J. Culley、T. Nicholson、J. Omokeowa、S. Khan、S. Grazioso、A. Bradley、G.D. Gironimo 和 F. Cuzzolin，《ROAD：面向自动驾驶的道路事件感知数据集》，IEEE TPAMI（2022年）。[27] K. He、X. Zhang、S. Ren 和 J. Sun，《用于图像识别的深度残差学习》，载于：CVPR论文集，2016年。[28] E.J. Topol，《高性能医学：人类智能与人工智能的融合》，Nature Medicine 25(1)（2019年），44–56。doi:10.1038/s41591-018-0300-7。[29] S. Sendelbach 和 M. Funk，《警报疲劳：患者安全问题》，AACN Advanced Critical Care 24(4)（2013年），378–386。[30] C. Sudlow、J. Gallacher、N. Allen、V. Beral、P. Burton、J. Danesh、P. Downey、P. Elliott、J. Green、M. Landray、B. Liu、P. Matthews、G. Ong、J. Pell、A. Silman、A. Young、T. Sprosen、T. Peakman 和 R. Collins，《英国生物样本库：用于识别中老年期多种复杂疾病病因的开放获取资源》，PLOS Medicine 12(3)（2015年），1–10。doi:10.1371/journal.pmed.1001779。[31] J. Hippisley-Cox 和 C. Coupland，《QDiabetes-2018风险预测算法的开发与验证：估计未来2型糖尿病风险的队列研究》，BMJ 359（2017年）。doi:10.1136/bmj.j5019。[32] D. Jarrett、B.C. Cebere、T. Liu、A. Curth 和 M. van der Schaar，《HyperImpute：结合自动模型选择的通用迭代插补》，载于《机器学习研究论文集》第162卷，PMLR，2022年，第9916–9937页。[33] F. Imrie、B. Cebere、E.F. McKinney 和 M. van der Schaar，《AutoPrognosis 2.0：通过自动化机器学习推动医疗保健中的诊断与预后建模民主化》，arXiv预印本 arXiv:2210.12090（2022年）。[34] D.R. Cox，《回归模型与生命表》，英国皇家统计学会会刊：B辑（方法论）34(2)（1972年），187–202。doi:https://doi.org/10.1111/j.2517-6161.1972.tb00899.x。[35] C. Lee、W. Zame、J. Yoon 和 M. Van Der Schaar，《DeepHit：一种处理竞争风险的深度学习生存分析方法》，载于：AAAI论文集，第32卷，2018年。[36] Y. Li、J. Wang、J. Ye 和 C.K. Reddy，《生存分析的多任务学习表述》，载于：ACM SIGKDD论文集，2016年，第1715–1724页。[37] S. Mäkinen、H. Skogström、E. Laaksonen 和 T. Mikkonen，《谁需要MLOps：数据科学家希望完成什么，以及MLOps如何提供帮助？》，载于：第1届IEEE/ACM人工智能工程研讨会——人工智能软件工程，IEEE，2021年，第109–112页。[38] K. Ahmed、S. Teso、K. Chang、G.V. den Broeck 和 A. Vergari，《用于神经符号学习的语义概率层》，载于：NeurIPS论文集，2022年。

[39] A.Agarwal、A. Beygelzimer、M. Dudík、J.L. 0001和 H.M. Wallach，《一种公平分类的约简方法》，载于：ICML 论文集，PMLR，2018年，第 60–69 页。[40] J. Xu、Z. Zhang、T. Friedman、Y.Liang和 G.Van denBroeck，《用于结合符号知识进行深度学习的语义损失函数》，载于：ICML 论文集，2018年。[41] M. Diligenti、M. Gori、M. Maggini 和L. Rigutini，《连接逻辑与核机器》，《机器学习》86（2012年）。[42] S. Badreddine、A. d'Avila Garcez、L. Serafini 和 M.Spranger，《逻辑张量网络》，《人工智能》303（2022年）。[43] L. Pulina 和 A.Tacchella，《用于人工神经网络验证的抽象细化方法》，载于：CAV论文集，《计算机科学讲义》，第6174 卷，施普林格，2010年，第 243–257 页。[44] A. Lomuscio 和 L. Maganti，《前馈ReLU 神经网络的可达性分析方法》，CoRR abs/1706.07351（2017年）。http://arxiv.org/abs/1706.07351。[45] M. Rhu、N. Gimelshein、J. Clemons、A.Zulfiqar和S.W. Keckler，《虚拟化深度神经网络：可扩展且节省内存的神经网络设计》，载于：IEEE/ACM 微体系结构国际研讨会论文集，IEEE计算机学会，2016年，第18:1–18:13 页。[46] C. Chen、O. Li、D. Tao、A. Barnett、C.Rudin 和J. Su，《看起来像那个：用于可解释图像识别的深度学习》，载于：NeurIPS论文集，H.M.Wallach、H. Larochelle、A. Beygelzimer、F. d'Alché-Buc、E.B.Fox和 R.Garnett 编，2019年，第 8928–8939页。[47] F. Doshi-Velez和B. Kim，《走向一门严谨的可解释机器学习科学》，2017年。[48] N.Mehrabi、F. Morstatter、N. Saxena、K. Lerman和 A. Galstyan，《机器学习中的偏见与公平性综述》，《ACM 计算调查》54(6)（2022年），115:1–115:35。doi:10.1145/3457607。[49] M. Wick、S.Panda和 J.-B. Tristan，《释放公平性：重新审视权衡》，载于：NeurIPS 论文集，第 32卷，Curran Associates, Inc.，2019年。[50] T. Weng、H. Zhang、P.Chen、J. Yi、D. Su、Y. Gao、C. Hsieh 和 L.Daniel，《评估神经网络的鲁棒性：一种极值理论方法》，载于：ICLR论文集，OpenReview.net，2018年。[51] G. Katz、C.W. Barrett、D.L. Dill、K. Julian和M.J. Kochenderfer，《迈向证明深度神经网络的对抗鲁棒性》，载于：首届自动驾驶车辆形式化验证研讨会论文集，FVAV@iFM 2017，意大利都灵，2017年9月19日，L.Bulwahn、M. Kamali 和 S. Linker 编，电子理论计算机科学论文集，第 257 卷，2017年，第19–26页。doi:10.4204/EPTCS.257.3。[52] P. Dragone、S.Teso 和 A.Passerini，《用于结构化预测的神经符号约束规划》，载于：IJCLR-NeSy，2021年。[53] X. Huang、D. Kroening、W.Ruan、J.Sharp、Y.Sun、E.Thamo、M.Wu和X.Yi，《深度神经网络的安全性与可信性综述：验证、测试、对抗攻击与防御及可解释性》，《计算机科学评论》37（2020年），100270。doi:10.1016/j.cosrev.2020.100270。[54] C. Cheng、C. Huang和H.Yasuoka，《用于测试机器学习赋能的自主系统的定量投影覆盖率》，载于：自动化技术验证与分析论文集，《计算机科学讲义》，第11138卷，施普林格，2018年，第 126–142 页。[55] A. Odena、C. Olsson、D.G. Andersen 和 I.J.Goodfellow，《TensorFuzz：使用覆盖率引导模糊测试调试神经网络》，载于：ICML论文集，《机器学习研究论文集》，第97卷，PMLR，2019年，第 4901–4911页。[56] W.W. Royce，《大型软件系统的开发管理：概念与技术》，载于：ICSE论文集，ACM 出版社，1987年，第328–339 页。[57] B.W.Boehm，《软件开发与增强的螺旋模型》，《计算机》21(5)（1988年），61–72。doi:10.1109/2.59。[58] G. Symeonidis、E. Nerantzis、A. Kazakis和G.A.Papakostas，《MLOps——定义、工具与挑战》，载于：IEEE 计算与通信研讨会论文集，IEEE，2022年，第 453–460 页。[59] M.M. John、H.H. Olsson和J. Bosch，《迈向MLOps：框架与成熟度模型》，载于：软件工程与先进应用论文集，IEEE，2021年，第 1–8页。[60] N. Eén和 N. Sörensson，《可扩展SAT求解器》，载于：可满足性问题会议论文集，《计算机科学讲义》，第 2919 卷，施普林格，2003年，第 502–518 页。