人工智能的监管机制构想

[澳]罗杰·克拉克（Roger Clarke）*

马玉**译

摘要：人工智能正掀起又一波周期性流行浪潮，若能实现当前的承诺，人工智能可以带来巨大的收益。然而，无论是否如此，人工智能都隐含着相当大的威胁。本系列的第一篇文章探讨了这些威胁，第二篇则提出了一套原则和业务流程，以便各组织能负责审慎地应用人工智能。考虑到人工智能的重大影响，并且想让所有组织都负责行动几乎是不可能的，因此有必要构建一个适用于人工智能的监管机制。

本文回顾了重要的监管概念，并考虑了监管机制能采用的各种形式。鉴于这些威胁所具有的技术和政治上的复杂性和强度，合作监管形式应当最为适宜。这包括建立具有一些重要特征的立法框架，议会需要公布相关的要求、执行过程和制裁措施，并向适当的监管机构分配权责；另外，它需授权一个独立机构以开发并维护履行具体的义务，该机构由所有的利益相关群体的代表组成，包括不同类别受到影响的公众。

关键词：自然控制，预防原则，基础设施监管，合作监管，数据分析，人工智能/机器学习，机器人技术，半机械人化

一、 引言

人工智能当前的表现再次引起了相当大的关注，并据称为很多领域提供了广阔的前景，例如节约劳动力，更快、更可靠或更高质量的决策和行动，以及发现新的有价值的隐藏信息等。另一方面，人工智能包含着许多严重的威胁。本系列的第一篇文章中阐释了公众对人工智能的担忧的本质："人工智能导致了推理、决策和行动的失误，这是由于智能产品几乎是独立运作的，无法做出理性的解释，不具备调查、纠正和赔偿的能力"。

这些担忧的根本原因被认为有：智能产品的自主性、数据假设和推理过程假设、推理过程的不透明以及未能将责任归咎于法律实体。我们有必要对社会能够管控风险的同时仍能利用人工智能技术获利的恰当的方式进行理性分析，文中探讨了这些方式的可能性。

---

* 罗杰·克拉克（Roger Clarke），澳大利亚堪培拉 Xamax 咨询公司负责人，澳大利亚国立大学客座教授，新南威尔士大学客座教授。电子邮件：Roger.Clarke@xamax.com.au. 本文载于《计算机法律与安全评论》杂志 2019 年 8 月第 4 期第 398-409 页，https://doi.org/10.1016/j.clsr.2019.04.008. 预印本载于 http://www.rogerclarke.com/EC/AIR.html. 本文为作者"负责任的人工智能"系列文章的第 3 篇。

** 马玉，上海外国语大学 2019 级法律硕士研究生。

本文首先回顾了重要的监管概念，包括监管机制的设计和评估标准。此外讨论了自然控制以对监管干预的阈值测试进行限定；明确了现行法律的失效。接着概述了监管制度可采取的各种形式及其与人工智能特征的相关性，最后提议建立一个合作监管框架以实现对人工智能的公共风险管理。

二、 监管概念

这部分简述了监管理论的重要概念，从而为分析人工智能监管的可选择形式构建基础。对“监管”这一概念有许多种定义，如参见 Black(2002)以及 Brownsword and Goodwin(2012)。本文则采纳了工具性使用的定义：监管是对各种实体的行为进行控制。

这一定义方式表明它不仅是实现政策目标的工具，同时也是意外灾害和事故的管控机制；还避免了有关监管手段和监管目的的表述。监管制度的目的常常引起争议，而且它们在随着时间改变。另外，实现目的的手段有效性并不是一个定义上的特征，而应是监管机制的特质。

本系列先前的文章探讨了人工智能计划的利益相关者范围，认为人工智能的产品和系统太具影响力以至于组织需要考虑所有利益相关者的利益。监管要在个人的智能产品、系统以及技术之上运作，一般是通过开发和应用它们的实体实现管控。

与其说利益相关者的概念是组织中有用的分析工具，不如说是用不同的方法将相关的实体分类更符合当前的目标。上文提到的有效的监管定义针对的是行为受到控制的实体，总体上分为三个主要类型，即：

（一）受监管者，是指受监管制度约束的实体，一般包括公司、非公司型企业、政府机构、合作社、法人或非法人组织以及私人。

（二）监管者，是指为了管控受监管者的行为而行使权力的实体，一般是受严格控制的政府机构或一个相对独立的委员会，但也可能是诸如证券交易所的法人团体，甚至可以是管理行业行为准则的协会。

（三）受益者，是指有意、无意或偶然地因监管安排获益的实体，一般包括“受监管者”列下的任何形式的实体。不过，这一概念可以延伸到社会价值，例如对社会和经济体系的信任、环境质量等；而那些利益遭到人工智能威胁和因监管安排缓和了劣势的实体属于重要的子类实体。但是，监管安排不仅仅要涵盖预期受益者，因为多数方案都有意外获利者（例如，负有责任的受监管者可能凭借新进入者获取战略性利益），而这些人自然会反对可能减除他们利益的机制改革。

对任何特定的监管环境的详尽分析都要求更精细的粒度分析。其他重要的实体类型有受监管者的代表和中介（例如律师、承保人、融资人和咨询顾问），以及受益者利益的倡导者。这些实体支持市场信号流，而这对一个有效的监管机制是至关重要的。在 Clarke (2018a)文中有更全面的模型。

任何特定监管机制的设计都反映了监管、放松监管和再监管过程中参与者的目标；不同机制之间在一致性和完备性上差异巨大，这取决于各种利益的冲突程度以及参与设计过程或受设计过程影响的各方力量。对监管机制的后续改进可能会扩充、取消或简化其要求，但往往会增加其复杂程度。

表 1 所列的标准是借助大量的文献研究完善的，尤其是 Gunningham et al.(1998)，Hepburn (2006) 和 ANAO (2007)这三者，可以为监管机制的设计和现有制度的评估提供指导。

表 1-监管制度的设计和评估标准

Clarke and Bennett Moses (2014)文中表 2 的拓展版

流程

- **清晰的目的和要求**：目的和义务对于受监管者和受益者是可理解的；
- **透明度**：公开开发和审查过程，公布行为要求；
- **参与性**：所有利益相关者都参与到开发和审查过程中；
- **反映利益相关者的利益**：表明受益者的需求，反映受监管者的合法利益；

设计结果

- **全面性**：所有相关方面都含括在一体性的框架内；
- **经济简约**：机制不能过分的、不正当的繁杂、昂贵。
- **清晰性**：要求要明确具体、具有可操作性，以便受监管者有效且高效地执行。
- **教育价值**：用说明性、指示性的形式表述要求，不用抽象的、隐晦的散文表述。

机制效果

- **监督**：受管制行为得受到监控。
- **可执行性**：受管制行为得直接由受益者或由执行机构强制执行。
- **强制执行**：执行机构享有并运用适当的权力和资源，促使制度得到遵守。
- **透明度**：公开公示监管者的执行措施和受监管者的回应，从而指引所有受监管者的行动。
- **审查**：对机制予以审查调整，确保取得的效果符合目的。

有关监管机制存在着庞大的理论体系（Braithwaite, 1982; Braithwaite and Drahos, 2000; Drahos, 2017）。在 20 世纪后半期，一般认为，监管机制的合理形式应含括一个综合性的、采取分级措施的、“执行金字塔”或“合规金字塔”结构的监管机构（Ayres and Braithwaite, 1992, p. 35）。这一模式设想了广泛的支持基础，包括作为调解和仲裁基础的教育和指导；它还规定了制裁和执行机制，例如必要时的指导和限制措施、处理重大或重复违规行为时的中止或取消权等。

近几十年以来出现了更多的监管形式，尽管其中许多都反映出了受监管者对监管的抵制和破坏力。随着许多国家的议会和政府退出了对产业的正式监管，“治理”这一概念已经取代了“政府管理”的概念（Scott, 2004; Jordan et al.,2005）。近来许多文献都关注于放松监管，通过“监管影响评估”之类的机制主张减少限制企业自由的措施、用“更好的监管”委婉呼吁减轻企业的“合规负担”；同时，政府机构拒绝对其自身适用监管机制，导致了大量浪费和腐败无法遏制。

几乎不负法律义务从而只承受有限的合规风险，这似乎对组织有很大的吸引力；另一方面，监管的缺失或薄弱会助长损害公众合理期待的行为，组织可能在管理层、团体、甚至是察觉到机会的个体的驱使下作出轻率的行动，这会对该行业部门内的所有组织造成严重的直接或间接的威胁。因此，适量的监管符合每个组织的本身利益，目的是保护他们免遭媒体披露，并避免激化公众对抗情绪和监管能动主义。

下文将讨论监管机制可选择采取的各类形式。无论如何，首先必须要考虑自然控制可能导致的不必要的甚至有害的监管干预范围，由此还要识别正当干预的情形。

三、 自然控制和正当干预

人工智能技术以及人工智能的产品和系统会受到相关社会经济制度的固有过程的限制(Clarke, 1995, 2014a,2014b)，人工智能甚至会刺激具有限制应用或抑制、减轻负面影响效果的自然过程。

自然控制的一个常见例子是，质疑技术是否能实现其支持者的允诺并导致技术发明缺少投资。当创新项目成功获得早期融资回合时，就会发现开发和/或运营成本过高，或是可应用的情形和/或从应用中的获利太少，以至于无法证明开发智能产品或执行部署智能系统所需投资的合理性。

在某些情况下，技术潜在利益的实现可能受制于不可用或不充分的基础结构依赖。例如，要是当时的冶金术能支持巴比奇（Babbage）的“差分机和分析计”并保障充分的投资，计算机技术爆炸本可以在 19 世纪的后半期而非 100 多年后发生。

自然控制的另一个形式是，实体针对可能对他们的利益产生不利影响的行为行使抗衡力，一个常用的例子是竞争者、供应商、消费者和雇员的市场力，还有监管者、融资人和承保人的组织力。反对者长期以来善于利用媒体鼓动公众斥责，现在社交媒体则提供了更多的机会，这原因之一就是声誉效应——由于实行方法被认为有害于重要的社会价值，早期实行可能遭到反对并造成严重负面的公众形象。一个实例表明，波音 737 MAX 的防失速系统实际上不受飞行员控制，以至于在 2018 至 2019 年造成了两起坠机事故和至少 300 人遇难，导致飞机暂停运营使用(FAA, 2019)。

经济性自然控制需要更密切的关注。“每个人只关注自己的私利，并受到‘看不见的手’的指引增进公共利益”(Smith, 1776)以及“经济制度因此是内在地自我调节的”(Williamson, 1979)，

这些假设随后被交易成本经济学所证成；然而，内在的自我调节有其局限性，例如“公地悲剧”(Hardin, 1968, 1994; Ostrom, 1999)。新保守主义经济学家普遍认为“市场失灵”是政府干预的唯一正当理由，Stiglitz（2008）则增加了“市场非理性”（如避免股市从众效应的跌停板措施）以及“分配正义”（如安全保证与反歧视措施）的理由。

就人工智能而言，本系列先前的文章指出了市场失灵的情形。尽管各类技术都做了运营部署，但并无有效的组织、行业或专业自我监管；现有的此类准则和指南只含括了一小部分需求，且不可能强制执行。同时，用户组织天真的接受人工智能支持者的主张显然反映出了市场非理性；分配正义也正因授信和社会福利管理等领域存在的不公平且不可争讼的决定受到不良影响。

从自然控制的研究中可以获得一个更重要的发现，即监管措施可以被用来增强自然过程。例如一些可广泛适用的方法包括通过补贴成本、征费收入和/或风险转移来调整成本、利润、参与者可察觉的风险平衡，比如，对无人机和无人驾驶汽车的操控者实行严格责任可能会鼓励更慎重的风险评估和风险管理。

了解已有的和加强的自然控制是任何监管分析的重要前提，因为分析的出发点必须是：不恰当事物的自然秩序是怎样的？干预又将怎样改善这种情况？

例如，澳大利亚生产力委员会提出的六项原则之首就是“政府不能通过监管行动来解决问题，除非已有明确规定的行动理由”，这应包括评估和解释现有措施不足以解决问题的原因(PC, 2006, p.v)。对此，阈值测试十分重要，以确保充分理解特定情形中存在的自然控制。

实际上，议会很少能在新技术部署之前采取行动，原因包括：对技术及其影响缺乏理解；对经济问题优先于社会问题考虑，因此倾向于鼓励新业务而非自始就加以抑制；以及创新型企业比起消费者和社会价值的拥护者能进行更有效的游说。

有观点认为，技术越有影响力，议会就更有理由采取先行行动。对此经常提到的技术包括核能和各种形式的大规模开采和制造业，对其监管不力造成了严重污染。为此，“预防原则”已经得到了阐明(Wingspread, 1998)，它在一些司法管辖区的环境法条文中得到了有力运用，如“当人类活动可能造成在科学上合理但不确定的、而道德上不可接受的危害，就应采取措施避免或减轻那样的潜在危害”(TvH, 2006)。但是，除了许多特定的司法管辖区的环境事务之外，预防原则仅仅是一种道德规范，即：如果怀疑某项行动或政策会造成危害，并且缺乏对其无害的科学共识，那么由采取行动的一方承担证明责任。

本系列的第一篇文章认为人工智能的威胁是显而易见且实质性的，人工智能的支持者即便不接受该观点，但至少认可将预防原则作为道德规范予以适用，因为人工智能项目的影响之广不可避免。

由此，进行监管干预具有了强有力的理由，除非可以证明适当的监管措施已经到位。以下部分则概述了现有的监管制度安排。

四、 现有制度

这部分首先考虑了可提供适当保护或至少有利于构建监管框架的一般法律条款，其次探讨了人工智能专门法的制定动因。

（一）一般法律

新技术的应用通常要遵守现有法律(Bennett Moses, 2013)，包括各种形式的商法，尤其是含括明示和默示条款的合同义务，消费者权益法以及著作权法和专利法。在机器人技术、半机械人产品以及在设备中嵌入人工智能软件等情形下，很可能适用产品责任法；其他转移风险至技术创新者的法律也可能适用，例如过失侵权；还有人权法、反歧视法和数据保护法等普适性的法律也能适用；公司法则为公司董事设定了相关的义务。监管的效力渊源还可能是应用人工智能的各个行业部门相关的法律，例如道路交通法、工作场所和就业法以及卫生法。

但是，特别是在普通法司法管辖区，法庭和法院对特定纠纷的法律适用方式很可能具有极大不确定性，这在某种程度上会阻碍创新，并会大大增加支持者的成本、延迟新技术部署。另一方面，认为自己会因技术创新受到负面影响的人们认为，应对这些威胁的法律渠道可能难以获得、费用高昂却见效缓慢、甚至完全无效。

特别是针对变革性的、颠覆性的技术突破，现有法律很可能无法适应那些新情形，因为这些法律是“围绕相对遥远的过去的社会经济环境设计的”(Bennett Moses, 2011, p. 765)，无法预知新的技术形式。在有些情况下，现有法律可能对技术创新者和受新技术影响的人都毫无帮助，从而阻碍新兴技术；在其他情况下，现有法律的制定方式可能不适用于新技术形式，或者通过司法活动以使其看似适用，以软件版权领域的司法活动为例，见 Clarke (1988)。

（二）人工智能专门法

在生产线和仓库中空间受限的工业机器人技术已经发展完备。各种出版物都在讨论机器人监管的一般性问题（如 Leenes and Lucivero, 2014; Scherer, 2016; HTR, 2018a, 2018b），但很少去识别人工智能专门法，甚至是在劳工安全和雇主责任这样重要的方面似乎也并不依靠技术专门法，而是依据一般法律，但一般法律可能并没有经过修改以反映新兴技术的特征。

在 HTR (2017)中，韩国确认颁布了第一部与机器人技术有关的国内法，即《智能机器人开发和普及促进法》（2008 年），该法几乎完全是促进性和激励性的，对机器人技术几乎不求监管。有人提到韩国“机器人道德宪章”——“包括总统令的规定，例如智能机器人的开发者、制造商以及用

户都应遵守的道德准则”，但此宪章似乎并不存在，不过 Akiko (2012)描述了一个此类宪章的可能形式。HTR(2018c)中提出了有关研究和技术的一般监管规范，包括机器人技术和人工智能。

许多司法管辖区已经针对自动驾驶汽车颁布了法律，参见 Palmerini et al. (2014,pp.36 – 73), Holder et al. (2016), DMV-CA (2018), Vellinga (2017)，其中审查了美国联邦层面、加利福尼亚、英国以及荷兰的法律；还有 Maschmedt and Searle (2018)文中审查了澳大利亚三个州的法律。这些审查措施普遍着重于经济动机、刺激和促进创新、现有法规的免责条款以及有限的新规范甚至指南。监管方法之一就是利用自然过程，例如， Schellekens(2015)认为，要求强制保险就是一个调整自动驾驶汽车造成的损害责任的充分手段。但不得不说，立法和监管机构在对无人机的监管方面行动非常缓慢(Clarke and Bennett Moses, 2014; Clarke, 2016)。

多年来关于人类的自动决策一直受到法国数据保护法的约束，在 2018 年年中，《通用数据保护条例》（GDPR）使得该约束成为了欧洲法律的普遍特征，尽管该条例第 22 条的效力受到了质疑（Wachter et al., 2017）。

一方面，人工智能技术可能并不如人们所声称的那样具有颠覆性，因此法律可能并不需要调整；另一方面，“技术中立”的神话渗透在了立法中。即便理想的法律可能同时含括现有的和将来的产品和过程，真正具有颠覆性的技术还是会使得现有法律变得模糊、无效。

人工智能不仅没有受到适当的自然控制，还使得目前适用的法律不足以应对它所含的严重威胁。因此，以下部分概述了可以适用的各种监管干预形式。

五、 监管形式的层次体系

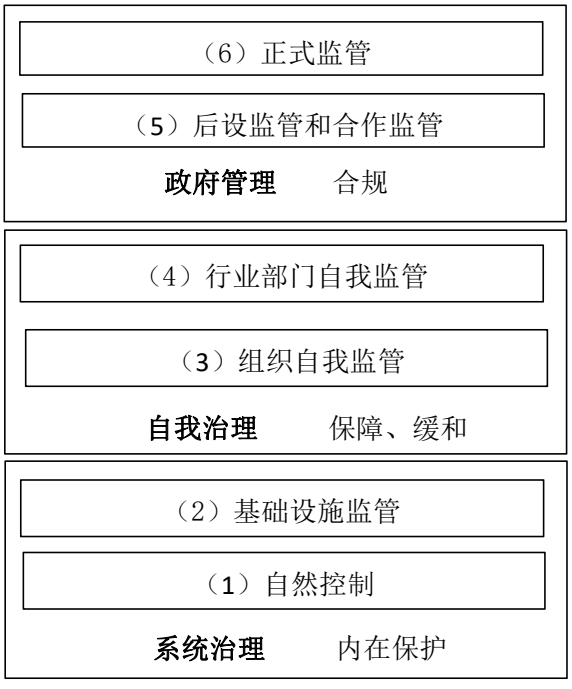

图 1- 监管形式的层次体系

这部分反映了前述的监管概念，并根据监管干预的正式程度提出了可采形式的层次体系。前文讨论了自然控制，它在图 1 中被置于体系最底层（1）。

监管理论通常指可以用于干预自然过程的“工具”和“手段”。原则上，它们的目的在于遏制危害行为和过度行为，但有些情况下目的只是做出这样的姿态，以阻止更有力或更有效的干预。图 1 将专门设计的监管“工具”和“手段”置于自然控制之上的第（2）至（6）层。

体系内的第二底层是（2）基础设施监管，是诸如机械蒸汽调节器之类的产品的相关概念。受监管者所依赖的基础设施特征能够加强相关社会经济制度的积极方面，并且抑制消极形势；这些特征可能是产品设计的副产品，也可能是在其上加以改进或设计的，例如，早期的蒸汽机没有合适的控制器，第一个调速器是一个改装部件；但是在随后的迭代中，控制器便成为蒸汽机的固有设计了。

信息技术（IT）辅助了以前纯机械的控制，例如根据水位、集水区降水事件或支流水流量增加来自动调整大坝闸门设置；信息技术和人工智能增强的信息技术提供了许多改进机会。就信息技术而言，对基础设施监管的一个流行说法是“西海岸规则”（West Coast Code）(Lessig, 1999; Hosein et al., 2003)，在计算机和网络体系结构（包括标准和协议）以及基础设施（包括硬件和软件）内存在着一系列的限制。

而针对人工智能，“西海岸规则”可采取的一个相关形式就是在机器人内嵌入类似于“机器人技术法则”的东西，此概念最初出现在 1942 年出版的阿西莫夫的短篇小说《逃亡》中，但许多机器人技术的评论家坚持这一概念，例如，Devlin (2016)引用了一位机器人技术教授的观点，认为英国标准协会对机器人道德设计的指南(BS, 2016)代表着“将道德价值嵌入机器人技术和人工智能的第一步”。另一方面，对阿西莫夫的机器人小说的研究表明，他已经全面证实了该想法是徒劳无用的(Clarke, 1993)；没有任何方法能将人类的价值编制为产品代码，也没有任何能嵌入的可以反映各利益相关者的不同价值、或是可以调解不同价值和目的间冲突的手段(Weizenbaum, 1976; Dreyfus, 1992)。

这一监管体系的最顶层是（6）正式监管，即通过成文法行使议会的权力，至少在普通法国家，判例法也作为补充，以阐明法律的适用。正式监管要求遵守通过较为具体的条款表述的要求，并辅以制裁和执行权。Lessig 将正式监管称为“东海岸规则”（East Coast code），以此强调基础设施监管措施和法律措施之间的区别。

监管要求未必是无条件的，也未必是以否定形式表述的，Clarke and Greenleaf (2018)一文中区分了许多种形式，如表 2 所示。禁止某些类型行为与强制其他形式行为的结合通常辅以附条件的允许和禁止，并受制于较为明确的前提或后置条件的实现。

表 2- 法规范的表现形式		
1，	禁止	不能..
2，	附条件禁止	不能..除非..
3，	沉默	自行确定
4，	附条件允许	可以..只要..
5，	授权	可以..
6，	强制	必须..

对法律的一个狭义解释是，它是政治上公认的权威机构制定的规则；而广义解释则含括了一系列更广泛的现象，包括获得授权的立法（例如规章条例）、国际条约、对后续裁判有约束力的法院和法庭裁判、由私人实体制定的但由国家通过合同和强制力保证认可和实施的规则章程、以及准法律文本，如谅解备忘录和正式的指导性文件(Clarke and Greenleaf,2018)。

正式监管通常是由一个专门的政府机构或议会授权机构运用权力和资源，以实施法律并为受监管者和受益者提供指导。对监管者的法律授权和资源配置可能是有限的，在这种情况下，监管制度可以被适当地描述为伪监管，该实体仅仅是一个监督机构而非监管者。

各种形式的人工智能可能处于现有机构或授权机构的监管范围内，通常的例子有自动驾驶汽车、遥控飞机以及具有数据处理能力的医学植入。但是，就人工智能总的来说，在人工智能时代开启的 60 年后，世界上主要的司法管辖区都还尚未建立新的监管机构或适当地授权现有机构。欧盟目前不得不超越由数据保护主管(EDPS, 2016)发布的一份讨论文件和初步声明(EC, 2018)的权限；英国数据保护委员会目前仅发布了一份讨论文件(ICO, 2017)；当前美国政府的政策本质上完全是激励性的，还提到监管纯粹是实现经济目标的障碍(WH, 2018)。许多组织也都提出了一些监管原则，但大多数是鼓励性而非义务性的，例如欧洲绿党联盟(GEFA, 2016)、UNI 印度联合新闻社国际联盟(UGU, 2017)、日本政府(Hirano,2017)、上议院委员会(HOL, 2018)以及法国议会(Villani, 2018)。

正式监管施加了相当大的限制和成本，处于中间层次的(3) – (5)则试图减少正式监管所固有的巨大限制和弊端。其中的最底层是（3）组织自我监管，包括内部行为准则和“客户宪章”，以及有关“商业道德”和“企业社会责任”等说法的自我约束(Parker, 2002)。本系列的第二篇文章中曾指出，法律要求公司董事要以追求公司利益为首要目标，因此，组织的自我监管在所谓的受益者看来几乎总是无效的，甚至无益于保护组织自身免遭负面宣传，这一点不足为奇甚至是意料之中的。IBM (2018)、 Google (Pichai, 2018) 和 MS (2019)在内的大公司最近都提出了自我监管，出于怀疑，对这些文件做了汇总，参见 Newcomer (2018)。

这一体系的中心位置就是（4）行业部门自我监管。企业出于各种原因联合起来，其中一些可能对其他造成损害，例如串通投标和定价。但是，行业协会的活动不仅能给协会成员也能给非成员带来利益，特别是对基础设施的协作方法，能够改善服务、降低该行业客户的成本、甚至能嵌入基础设施监管机制。

也可以说，如果由一个行业部门中较负责任的企业制定规范，那么“业内莽夫”的不当行为就会突显。行业内的行为准则、实践准则或道德准则以及谅解备忘录被宣称具有或可能具有一定的监管作用，但是实际上，行业准则对企业行为的影响是微乎其微的；很少有足以严格保护其他各方利益的准则，况且其缺乏强制执行力。企业的忽视使得行业准则变得更加边缘化，同时负责任的企业因竞争压力会减少对行业准则的承诺，结果使这些准则成了幌子，掩盖了保障的缺失并阻碍了实际的监管措施。在人工智能领域，FLI (2017), ITIC (2017)和 PoAI (2018)都是急于应对监管威胁而结成行业联盟的例子。

在某些领域有一个特别的机制，即认证机制，有时被称为“良好内务管理”“批准对号”，对此最好理解的是后设品牌。获取并保持认证标志的条件几乎无法实质性地保护名义受益者(Clarke, 2001; Moores and Dhillon, 2003)，一个案例即是利用数据保护标志 TrustE/TrustArc 进行有源欺骗，参见 Connolly (2008), Connolly et al. (2014)。

就其性质而言，以及在贸易惯例、反垄断法、反卡特尔法的影响下，行业自我监管机制通常不具有约束力和强制执行力。另外，受监管者的博弈使得机制的有效性和/或繁重性被降低，或是带来附带好处，例如锁定竞争对手或锁定客户，结果使这两个自我监管机制几乎是完全无效的。Braithwaite(2017)指出,“自我监管在行业滥用特权方面有着悠久历史”(124 页),并且 Gunningham and Sinclair (2017)得出结论认为，“自觉自愿”只有在与“命令与控制”要素结合时才能成为有效的监管要素。

比起行业协会，专业协会能够更好平衡公众需求和自身利益，因此也能起到一定作用，但是，它们的影响远不及行业协会重大。此外，迄今两个最大的国际机构的倡议都令人震惊——国际计算机学会 ACM (2017)使用诸如“应该”“鼓励”等疲弱的表达形式，国际电气与电子工程师学会 IEEE(2017)则以冗长的散文形式提出了过分模糊和限定的原则，二者至今为止都没有提供专业人员、管理者和执行者所需的指导。

行业标准在一些组织实践中是有相关性的。HTR (2017)列出了由国际标准组织(ISO)发布的人工智能领域的行业标准，其中相当一部分关注互操作性，其他一些描述了旨在实现质量保证的业务流程。公共安全是一个重要课题，尤其是在通常被称为“安全关键系统”的领域（参见 Martins and Gorschek, 2016），因此，可以通过开发和应用行业标准去避免、减轻和管理人工智能系统所包含

的某些物理威胁；然而其对经济和社会利益的威胁很少被涉及。即使在业务流程方面，行业标准取得的进展也明显很晚、很慢。在 2016 年，IEEE 标准协会宣布了制定“标准 P7000”的项目计划，即“解决系统设计过程中的道德问题的流程模式”，时隔三年却什么都没有发表。

在过去的四十年中，议会一直在努力理解和应对新兴技术，因此出现了更多的监管形式。从某种意义上，这些形式介于（通常是严厉的）正式监管和（大多是无效的假托性的）自我监管之间，将“自觉自愿”和“命令与控制”要素相融合，这与 Gunningham and Sinclair(2017)一文相符。

在 Grabowsky (2017)文中提到，“强制型自我监管”的概念可追溯到 Braithwaite (1982)，而“后设监管”这一术语的使用可追溯到 Gupta and Lad (1983)，意思是“政府管制下的行业自我监管”。“后设监管”的一个例子就是澳大利亚数据保护法《隐私权法》（联邦）7B(4)(b)条款对“媒体组织”的免责，条件是“该组织公开承诺遵守以下标准：（1）在媒体组织的活动范围内处理隐私信息；（2）已由该组织或代表该类媒体组织的个人或机构以书面形式公布此承诺”；但在该法条文中没有任何管控措施的规定，毫无疑问，这些“标准”也是空洞的、非强制的。后设监管的正面示例着实很难找到。

与此同时，“合作监管”概念出现了(Ayres and Braithwaite, 1992; Clarke, 1999)。概括来讲，合作监管手段需要通过立法建立一个监管框架，然后对细节详尽地授权，其关键要素是权限、义务、监管机制要符合的一般原则、以及制裁和实施机制。详尽的义务必须通过利益相关者的代表之间的磋商过程来设定，从而产生一个能阐释且符合立法所体现的一般原则的强制性准则；参与者必须至少包括监管机构、受监管者和监管的预期受益者；监管流程必须反映各方的需求，而不能被组织力和市场力扭曲；有效的制裁以及对制裁的强制执行是监管机制的本质要素。

不幸地是，很难找到有效的合作监管的例子。原因之一在于开发过程通常会排除或遏制实力较弱的利益相关者的利益，另一个原因是它通常不能得到有效执行，甚至不能执行(Balleisen and Eisner, 2009)。例如，在澳大利亚，所谓的“强制执行规则”是由澳大利亚通信和媒体管理局(ACMA)针对电视、无线广播和电信施行的；同样，澳大利亚审慎监管局(APRA)也在银行业服务方面施行此类规定。这些安排确实以监管之名义促进了商业和政府活动，但它们没能对公众认为不恰当的行为实施管控，因此丧失了公信力。不管怎样，与后设监管相反，只要机制的所有关键特征设计完备，合作监管确实能够成为有效的监管机制。

这一部分概述了监管干预能采取的各种形式。实际上，许多监管制度都是采用一个主要形式再吸收一些其他形式的要素：“大多数情况下，采用多元而非单一的政策工具，以及更广泛的监管主体，通过工具和参与者互补结合的实施手段将能实现更好的监管”(Gunningham and Sinclair, 2017, p. 133）。

下面的部分简要分析了各种监管形式的特征，评估了每种形式对人工智能技术、人工智能的产品和系统实现管控的适当性。

六、 监管指标

以上确定的每种监管形式都至少能在特定的情形下发挥作用。这部分内容考虑了各种关键因素，包括支持应用各形式的因素，或是会对人工智能管控以及针对本系列第一篇文章第 4 节中认定的危害（产品自主性、对数据和推理过程的不正确假设、推理过程的不透明、不负责任）实施恰当的防护措施构成妨碍的因素。

这一分析考虑到了表 1 所列示的设计评估监管制度的标准，重点关注流程的透明度、对利益相关者利益的反映、监管机制的清晰明确以及强制措施。

在许多情况下，自然控制可能是有效的，或者至少做出了重大贡献，例如，飞机失事引发的周期性的公众恐慌可能足以遏止自主飞行的发展、政府机构的不公正行为可能激化公众对迫使放弃自动决策的反应。但是，在公众不了解技术运作、社会或社会技术体系复杂模糊、一个或一些强势参与者主导该技术领域并能做出符合自己需求的安排时，自然控制是不够的，所以说，人工智能产业的特征不利于自然控制的充分实施。

在信息技术领域则通常是基础设施监管发挥作用。人工智能为进一步发展提供了潜力，包括通过当前的监管科技运动（RegTech）(Clarke,2018a)；但是，在涉及实质性价值冲突、变化的环境、突发事件以及急剧变化的情况下，这类机制可能归于无效。人工智能产业的特征不利于基础设施特征的优先和实施，例如，即便是中等价格的无人机也缺乏通信信道和冲突检测功能。基础设施监管可能在生物医学工程上是最有效的，因为其早已引入了预防原则，并对构想和设计进行了细致的、逐步的试验和测试(Robertson et al.,2019)。

只有在预期受益者力量强大时，组织自我监管才能发挥较大作用；或是可能受制于昂贵且不便的正式监管的风险促使受监管者建立起保护措施并实际应用。行业自我监管一方面仅在具有强大的行业结构体系、行业参与者都有强大的动力成为其中成员时才有作用；但另一方面，若其他参与者足够强大到能确保监管机制实现预期受益者的利益，也能实现行业自我监管的作用。但不得不说，除非业内不法者感到不适，否则这类机制便是无效的、不可信的。人工智能包含许多种技术，它们都体现在各种产品中、嵌入在各种系统中、或受制于各种应用。其中有些技术至少具有一定的自主性，而所有这些技术都是复杂难懂的，甚至连管理者、经营者和决策者都不太可能理解，更何况受到技术影响的公众。人工智能是充满活力的，活跃在人工智能领域的实体大多数是小型的、不稳定

的、变化迅速且存续期短；人工智能也没有强大的行业结构体系，因此面对其所含的严重威胁，组织和行业自我监管看起来都不可能实现有效的保护。

后设监管只有在对组织和行业自我监管机制施加全面要求后才有效，这似乎不太可能出现许多正面的例子，而且技术和行业的复杂性更使得后设监管不可能适用于人工智能领域。

正式监管具有完全的法律效力，但是其设计、起草和讨论的过程受制于政治力量，而政治力量又受到强大势力的影响，这些势力通常是在幕后作用，因此其影响会被掩盖且难以察觉、很难克服，结果使许多正式监管机制都违反了表 1 提出的许多标准。在另一个极端，一些正式监管安排过分繁杂且代价高昂，大多数都不灵活。由于强大势力和政治过程的介入，所有的正式监管机制都只能缓慢地、挑战性地去适应不断变化的环境；如此的复杂性使得在议会很少有机会能对人工智能进行连贯的讨论。因此，针对正式监管很可能制定或发布既繁杂又无效的法规；但也能出现例外，比如禁止全自动客机。

另一方面，合作监管为实现受益者追求的价值提供了真实的前景，但这需要一个诱因，例如一个热心、强大而有说服力的部长、或一个特定部门内部或邻近的利益联盟。因此，必须组织一个具有高度代表性的会议来商讨一个可行的设计方案；相关政府、政府机构和议会必须做出承诺并承担义务，不得屈服于既得利益；监管者必须被赋予权力和资源，并在此类机制不可避免的变动时得到支持。以下部分进一步阐明了这一主张。

七、 人工智能的合作监管框架

在本文的第三、四部分中表明，由于人工智能的重大影响以及自然控制和现有法律规定的不足，监管干预的前提条件已经具备；此外，技术和应用技术的组织的信誉损失可能对个人和社会造成重大危害，也可能使经济严重受损。因此，应当适用预防原则，在开发和部署之前即进行监管。

前文得出结论认为，最有效的监管形式就是合作监管，并与自然控制和基础设施特征以及一些成文法相结合适用，例如禁止全自动客机。组织若能应用多方利益相关者风险管理流程以及在本系列第二篇文章中提出的 50 条原则，那么组织和行业自我监管可能会做出重大贡献。

这部分内容描述了可以应用于人工智能整体的监管机制，但是，人工智能的形式多样性使得将动因划分为一些专门技术的分支是具有相当大优势的。

要设计一个或多个人工智能监管机制的关键问题在于：具体要监管什么？人工智能的一般概念是散乱的，而概念本身也不是问题的重点。本系列的第一篇文章认为，互补智能和智力（Complementary Intelligence and Intellectics）才是问题的重点。

然而，监管要求通常是根据一类活动施加给该类实体，因此监管人工智能的一个合理方法就是应用本系列第二篇文章中的表 2 所作的区分，并分别对涉及到人工智能技术、产品、系统和已安装的系统的研究、发明、创新、传播和应用的实体施加要求。

这一机制的核心在于一个综合性的立法框架，它至少要包含表 3 中列示的要素。在这样一个合作监管机制内，各类实体都有其发挥能力的用武之地。

表 3- 综合性的合作监管框架	
1. 授权机构	授权给一个独立的委员会或部长，以批准一项或多项《准则》及其后续版本和替换版本，权限： a,《准则》必须遵守的一系列要求 b,《准则》必须体现的一套明确的原则 c, 协商《准则》的优先性 d, 若未达成协商《准则》或不能达成，有施行《准则》的保留权限
2. 制定机构	一个或多个协商和维护《准则》的机构和流程，其功能是： a, 将（必然抽象的）要求具化到《准则》 b, 通过磋商过程实现 c, 获得所有利益相关者的共识和积极参与，尤其是监管机制的预期收益者 d, 反映有效监管的标准（例如表 1 所列举） e, 以可操作的形式阐释“人工智能应负责任”的原则
3. 开发所需资源	用以支持协商或维护的机构和流程的财力、物力等资源
4. 强制执行机制	明确执行权和资源，并配置给一个或多个现有的和/或新的监管机构，该机构职能必须包括：监督磋商过程、监督合规情况、自行调查和投诉调查、惩戒不法行为、起诉犯罪者、研究技术和环境变化、提供信息交换所以及提供适应法律和《准则》的重点
5. 强制性义务	监管机构运用强制执行权和资源的义务

在人工智能供应链上各个环节的企业都可通过管理层的理性参与、资源投入、员工的文化认同、业务流程调整、控制和审计机制来解决问题。这些活动能够确保建立、运行并适应符合标准的内部投诉处理流程，并确保通过《准则》的协商机构和流程等实现与供应链中的其他企业和其他利益相关者的沟通交流。

行业协会能够作为其行业部门内部活动的重心，这包括对特定行业内组织的具体引导、协会成员企业投诉流程之外的二级投诉流程、实施保护性技术的基础设施以及提高认识和教育性的措施。

个人也需要被授权、被鼓励为了自己的利益采取适当的行动，但应当在以下前提下：已采取了提高认识和教育性的措施、在企业和行业协会级别启动了（相对非正式的）投诉流程、由监管机构、法庭和法院建立了（更正式一些的）投诉、合规执行和损害赔偿流程。在某些文化中，尤其是美国，自力更生被大力宣扬；而在其他一些文化中，它只起到较小的作用，相应的则会有一个作用更大、更有权力且资金更充裕的监管机构。

问题仍然在于，在广阔的人工智能领域，具体哪些技术和应用应该纳入监管范围？本系列的第一篇文章中确定了四项典型技术，其中，尤其是在公共领域和运动中的机器人技术似乎是早期监管的主要对象；同样地，有关神经网络之类的机器学习技术也急需监管。因此，监管范围可以被定义为涵盖其他人工智能衍生技术，例如基于规则的专家系统，并且要考虑为什么数据分析整体不应受制于同一监管制度。这一监管机制很容易为医疗植入领域开发，然后适用于其他半机械人形式。

本系列的第二篇文章分 10 个主题提出了负责任的人工智能的 50 条原则,它们能够作为表述《准则》所需要遵守的要求的模板，或作为对以其他方式开发的监管方案的评估清单。对于一般的数据分析，尤其是神经网络，这些原则还辅以一套《负责任的数据分析应用指南》(Clarke, 2018b)。

八、 结论

除非人工智能技术、产品、系统和应用所引发的重大公共风险能够受到适当形式的公共风险管理，否则人工智能将无法兑现其承诺。有一系列可行的替代监管方法，其中合作监管被认为是最合适的，这一点已经清晰阐明。

公开谈论人工智能的固有风险毫无意义，除非这能激励解决这些风险的建设性行动。同时，技术拥护者面临着公众和机构反对创新的很大可能性，因此实施可靠的公共流程符合所有人工智能利益相关者的利益，从而产生可靠的监管制度来应对或被视为在应对公共风险、而且不给人工智能产业施以过分繁重的法律义务。本文所提出的监管框架提供了此类流程和制度的蓝图。

致谢

本文受益于悉尼新南威尔士大学法学院（UNSW Law, Sydney）Graham Greenleaf 教授的宝贵反馈。

补充材料

与本文有关的补充材料可在在线版中获得：doi:10.1016/j.clsr.2019.04.008.

参考文献

ACM. Statement on algorithmic transparency and accountability. Association for Computing Machinery; 2017, at https://www.acm.org/binaries/content/assets/public-policy/2017_usacm_statement_algorithms.pdf.

ANAO. Administering regulation: better practice guide. Australian National Audit Office; 2007, at http://www.anao.gov.au/~ /media/Uploads/Documents/administering_regulation_.pdf.

Akiko. South Korean robot ethics charter; 2012, Akiko's Blog, at https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012/.

Ayres I, Braithwaite J. Responsive regulation: transcending the deregulation debate. Oxford Univ. Press; 1992.

Balleisen EJ, Eisner M. The promise and pitfalls of co-regulation: how governments can draw on private governance for public purpose, Ch 6. In: Moss D, Cisternino J, editors. 'New Perspectives on Regulation' The Tobin Project; 2009. p. 127–49. at http://elearning.muhajirien.org/index.php/catalog/download/filename/New_Perspectives_Full_Text.pdf#page= 127.

Bennett Moses L. Agents of change: how the law 'copes' with technological change. Griffith Law Rev 2011;20(4):764–94, at https://papers.ssrn.com/sol3/papers.cfm? abstract_id=2000428.

Bennett Moses L. How to think about law, regulation and technology – problems with "technology" as a regulatory target. Law, Innov & Technol 2013;5(1):1–20.

Black J. (2002) 'Critical reflections on regulation' 27 Aust J Legal Philos 1.

Braithwaite J. Enforced self-regulation: a new strategy for corporate crime control. Mich Law Rev 1982;80(7):1466–507.

Braithwaite J. (2017) 'Types of responsiveness' Chapter 7 in Drahos, pp. 117–132, at http://press-files.anu.edu.au/downloads/press/n2304/pdf/ch07.pdf.

Braithwaite B, Drahos P. Global business regulation. Cambridge University Press; 2000.

Brownsword R, Goodwin M. Law in Context: law and the technologies of the twenty-first century: text and materials. Cambridge University Press; 2012.

BS. Robots and robotic devices. Guide to the ethical design and application of robots and robotic systems. British Standards Institute; 2016.

Clarke R. Judicial understanding of information technology: the case of the Wombat ROMs. Comput J 1988;31(1):25–33. PrePrint at http://www.rogerclarke.com/SOS/WombatROMs-1988.html.

Clarke R. 'Asimov's laws of robotics: implications for information technology' in two parts. IEEE Comput 1993;26(12):53–61. (December) and 27,1 (January 1994) 57–66, at http://www.rogerclarke.com/SOS/Asimov.html.

Clarke R. (1995) 'A normative regulatory framework for computer matching' J Comput InfLaw XIII, 4 (Summer) 585–633, PrePrint at http://www.rogerclarke.com/DV/MatchFrame. html#IntrCtls.

Clarke R. (1999) 'Internet privacy concerns confirm the case for intervention' Commun. ACM 42, 2 (Feb) 60–67, PrePrint at http://www.rogerclarke.com/DV/CACM99.html.

Clarke R. Meta-brands. Priv Law Policy Report 2001;7(11). PrePrint at http://www.rogerclarke.com/DV/MetaBrands.html.

Clarke R. What drones inherit from their ancestors. Comput Law Secur Rev 2014a;30(3):247–62. PrePrint at http://www.rogerclarke.com/SOS/Drones-I.html.

Clarke R. The regulation of the impact of civilian drones on behavioural privacy. Comput Law Secur Rev 2014b;30(3):286–305. PrePrint at http://www.rogerclarke.com/SOS/Drones-BP.html#RN.

Clarke R. Appropriate regulatory responses to the drone epidemic. Comput Law Secur Rev 2016;32(1):152–5. (Jan-Feb) PrePrint at http://www.rogerclarke.com/SOS/Drones-PAR.html.

Clarke R. The opportunities afforded By RegTech: a framework for regulatory information systems. Xamax Consultancy Pty Ltd; 2018a Working Paper, at http://www.rogerclarke. com/EC/RTF.html.

Clarke R. Guidelines for the responsible application of data analytics. Comput Law Secur Rev 2018b;34(3):467–76, https://doi.org/10.1016/j.clsr.2017.11.002. PrePrint at http://www.rogerclarke.com/EC/GDA.html.

Clarke R, Bennett Moses L. 'The regulation of civilian Drones'. Comput Law Secur Rev 2014;30(3):263–85. PrePrint at http://www.rogerclarke.com/SOS/Drones-PS.html.

Clarke R, Greenleaf GW. 'Dataveillance regulation: a research framework'. J Law Inf Sci 2018;25(1). PrePrint at http://www.rogerclarke.com/DV/DVR.html.

Connolly C. (2008) 'Trustmark Schemes Struggle to Protect Privacy' Galexia, at http://www.galexia.com/public/research/assets/trustmarks_struggle_20080926/trustmarks_struggle_public.pdf.

Connolly C, Greenleaf G, Waters N. 'Privacy self-regulation in crisis?. TRUSTe's 'deceptive' practices' Priv Laws Bus Int Rep 2014;132:13–17. at http://www.austlii.edu.au/au/journals/UNSWLRS/2015/8.pdf.

Devlin H. 'Do no harm, don't discriminate: official guidance issued on robot ethics'. The Guardian 2016.at https://www.theguardian.com/technology/2016/sep/18/official-guidance-robot-ethics-british-standards-institute.

DMV-CA. Autonomous vehicles in California. California Department of Motor Vehicles; 2018, at https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/bkgd.

Drahos P, editor. Regulatory theory: foundations and applications. ANU Press; 2017, at http://press.anu.edu.au/publications/regulatory-theory/download.

Dreyfus HL. What computers still can't do: a critique of artificial reason. MIT Press; 1992.

EC (2018) 'Statement on artificial intelligence, Robotics and 'Autonomous' Systems' European Group on Ethics in Science and New Technologies' European Commission, at http://ec.europa.eu/research/ege/pdf/ege_ai_statement_ 2018.pdf.

EDPS. Artificial intelligence, robotics, privacy and data protection. European Data Protection Supervisor; 2016, at https://edps.europa.eu/sites/edp/files/publication/16-10-19_marrakesh_ai_paper_en.pdf.

FAA. Emergency order of prohibition. Federal Aviation Administration; 2019, at https://www.faa.gov/news/updates/media/Emergency_Order.pdf.

FLI. Asilomar AI principles. Future of Life Institute; 2017, at https://futureoflife.org/ai-principles/?cn-reloaded=1.

GEFA (2016) 'Position on robotics and AI' The Greens/European Free Alliance Digital Working Group, at https://juliareda.eu/wp-content/uploads/2017/02/Green-DigitalWorking-Group-Position-on-Robotics-and-ArtificialIntelligence-2016-11-22.pdf.

Grabowsky P. (2017) 'Meta-regulation' Chapter 9 in Drahos, pp. 149–161, at http://press-files.anu.edu.au/downloads/press/n2304/pdf/ch09.pdf.

Gunningham N. & Sinclair D. (2017) 'Smart Regulation', Chapter 8 in Drahos, pp. 133–148, at http://press-files.anu.edu.au/downloads/press/n2304/pdf/ch08.pdf.

Gunningham N., Grabosky P, & Sinclair D. (1998) 'Smart regulation: designing environmental policy' Oxford University Press.

Gupta A, Lad L. Industry self-regulation: an economic, organizational, and political analysis. Acad Manag Rev 1983;8(3):416–25.

Hardin G. The tragedy of the commons. Science 1968;162:1243–8. at http://cescos.fau.edu/gawliklab/papers/HardinG1968.pdf.

Hardin (1994) 'Postscript: the tragedy of the unmanaged commons' Trends Ecol Evol 9, 5 (May) 199

Hepburn G. (2006) 'Alternatives to traditional regulation' OECD Regulatory Policy Division, undated, at http://www.oecd.org/gov/regulatory-policy/42245468.pdf.

Hirano. AI R&D guidelines. Proceedings of the OECD conference on AI developments and applications, 2017. http://www.oecd.org/going-digital/ai-intelligent-machines-smart-policies/conference-agenda/ai-intelligent-machines-smart-policies-hirano.pdf.

HOL. AI in the UK: ready, willing and able? Select committee on artificial intelligence. House of Lords; 2018, at https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/100.pdf.

Holder C, Khurana V, Harrison F, Jacobs L. 'Robotics and law: key legal and regulatory implications of the robotics age (Part I of II)'. Comput Law Secur Rev 2016;32(3):383–402.

Hosein G, Tsavios P, Whitley E. 'Regulating architecture and architectures of regulation: contributions from information systems'. Int Rev Law Comput Technol 2003;17(1):85–98.

HTR. Robots: no regulatory race against the machine yet. The Regulatory Institute; 2017, at http://www.howtoregulate.org/robots-regulators-active/#more-230.

HTR. Report on artificial intelligence: part I – the existing regulatory landscape. 2018a. at http://www.howtoregulate.org/artificial_intelligence/.

HTR. Report on artificial intelligence: part II – outline of future regulation of AI. 2018b. at http://www.howtoregulate.org/aipart2/#more-327.

HTR. Research and technology risks: part IV – a prototype regulation. The Regulatory Institute; 2018c, at http://www.howtoregulate.org/prototype-regulation-research-technology/#more-298.

IBM. Everyday ethics for artificial intelligence; 2018, at https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

ICO (2017) 'Big data, artificial intelligence, machine learning and data protection' UK Information Commissioner's Office, Discussion Paper v.2.2, at https://ico.org.uk/for-organisations/guide-to-data-protection/big-data/.

IEEE. Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems (A/IS). IEEE; 2017 Version 2, at http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

ITIC. AI policy principles. Information Technology Industry Council; 2017 undated, at https://www.itic.org/resources/AI-Policy-Principles-FullReport2.pdf.

Jordan A, Wurzel RKW, Zito A. The rise of 'new' policy instruments in comparative perspective: has governance eclipsed government? Polit Stud 2005;53(3):477–96.

Leenes R, Lucivero F. 'Laws on robots, laws by robots, laws in robots: regulating robot behaviour by design'. Law, Innov & Technol 2014;6(2):193–220.

Lessig L. Code and other laws of cyberspace. Basic Books; 1999.

Martins LEG, Gorschek T. 'Requirements engineering for safety-critical systems: a systematic literature review'. Inf Softw Technol J 2016;75:71–89.

Maschmedt A, Searle R. Driverless vehicle trial legislation – state-by-state. King & Wood Malleson; 2018, at https://www.kwm.com/en/au/knowledge/insights/driverless-vehicle-trial-legislation-nsw-vic-sa-20180227.

Moores TT, Dhillon G. 'Do privacy seals in e-commerce really work?'. Commun ACM 2003;46(12):265–71.

MS. Microsoft AI principles. Microsoft; 2019 undated, at https://www.microsoft.com/en-us/ai/our-approach-to-ai.

Newcomer E. What Google's AI principles left out: we're in a golden age for hollow corporate statements sold as high-minded ethical treatises. Bloomberg; 2018, at https://www.bloomberg.com/news/articles/2018-06-08/what-google-s-ai-principles-left-out.

Ostrom E. Coping with tragedies of the commons. Annu Rev Polit Sci 1999;2:493–535. at https://www.annualreviews.org/doi/full/10.1146/annurev.polisci.2.1.493.

Palmerini E. et al. (2014). 'Guidelines on regulating robotics delivery' EU Robolaw Project, at http://www.robolaw.eu/RoboLaw_files/documents/robolaw_d6.2_guidelinesregulatingrobotics_20140922. pdf.

Parker C. The open corporation: effective self-regulation and democracy. Cambridge University Press; 2002.

Parker C. Meta-regulation: legal accountability for corporate social responsibility?. In: McBarnet D, Voiculescu A, Campbell T, editors. The New Corporate Accountability: Corporate Social Responsibility and the Law; 2007.

PC. 'Rethinking regulation' report of the taskforce on reducing regulatory burdens on business. Productivity Commission; 2006, at http://www.pc.gov.au/research/supporting/regulation-taskforce/report/regulation-taskforce2. pdf.

Pichai S. AI at Google: our principles. Google Blog; 2018, at https://www.blog.google/technology/ai/ai-principles/.

PoAI. Our work (thematic pillars). Partnership on AI; 2018, at https://www.partnershiponai.org/about/#pillar-1.

Robertson LJ, Abbas R, Alici G, Munoz A, Michael K. Engineering-based design methodology for embedding ethics in autonomous robots. Proc. IEEE 2019;107(3):582–99. at https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8620254.

Scott C. Regulation in the age of governance: the rise of the post-regulatory state. In: Jordana J J, Levi-Faur D, editors. The politics of regulation. Edward Elgar; 2004.

Schellekens M. 'Self-driving cars and the chilling effect of liability law'. Comput Law Secur Rev 2015;31(4):506–17.

Scherer MU. Regulating artificial intelligence systems: risks, challenges, competencies, and strategies. Harvard J Law Technol 2016;29(2):354–400.

Smith A. In: Strahan W, Cadell T, editors. The wealth of nations, London; 1776.

Stiglitz J. 'Government failure vs. market failure' principles of regulation. Initiative for Policy Dialogue; 2008 Working Paper #144, at http://policydialogue.org/publications/working_papers/government_failure_vs_market_failure/.

TvH. Telstra Corporation Limited v Hornsby Shire Council, NSWLEC 133. esp. paras. 113–83, at http://www.austlii.edu.au/au/cases/nsw/NSWLEC/2006/133.htm

UGU. Top 10 principles for ethical AI. UNI Global Union; 2017, at http://www.thefutureworldofwork.org/media/35420/uni_ethical_ai.pdf.

Vellinga NE. 'From the testing to the deployment of self-driving cars: legal challenges to policymakers on the road ahead'. Comput Law Secur Rev 2017;33(6):847–63.

Villani C. (2018) ‘For a meaningful artificial intelligence: towards a French and European strategy’ Part 5 – what are the ethics of AI?, Mission for the French Prime Minister, pp. 113–130, at https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Priv Law 2017;7(2):76–99 https://academic.oup.com/idpl/article/7/2/76/3860948.

Weizenbaum J. Computer power and human reason. Penguin: W.H. Freeman & Co.; 1976. p. 1984.

WH. ‘Summary of the 2018 white house summit on artificial intelligence for American industry’ Office of Science and Technology Policy. White House; 2018, at https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf.

Williamson OE. ‘Transaction-cost economics: the governance of contractual relations’. J Law Econ 1979;22(2):233–61.

Wingspread (1998). Wingspread statement on the precautionary principle, at http://sehn.org/wingspread-conference-on-the-precautionary-principle/.