< >
蒙特利尔人工
智能社会责任宣言
</ >

蒙特利尔人工智能开发社会责任宣言 2018 年

本文档是 2018 年《蒙特利尔人工智能开发社会责任宣言》的一部分。
如需查看完整的报告，请点击此处。

目录

阅读宣言	5
前言	7
原则	
1. 幸福原则	8
2. 尊重自主权原则	9
3. 保护隐私和亲密关系原则	10
4. 团结原则	11
5. 民主参与原则	12
6. 公平原则	13
7. 多样性包容原则	14
8. 谨慎原则	15
9. 责任原则	16
10. 可持续发展	17
词汇表	18
致谢	I
合作伙伴	II

阅读宣言

宣言的目标

《蒙特利尔人工智能开发社会责任宣言》有三个主要目标：

1. 为 AI 的开发和部署制定一个道德框架；
2. 引导数字化转型，让每个人都能从这场技术革命中受益；
3. 建立一个国家和国际论坛，讨论如何共同实现公平、包容和生态可持续的 AI 开发。

宣言的内容

原则

宣言的第一个目标是确定促进人民群众根本利益的道德原则和价值观。这些原则适用于数字技术和人工智能领域，但仍具有一般性和抽象性。要正确理解这些原则，务必牢记以下几点：

- 虽然它们以列表形式呈现，但并无优先等级之分。最后一项原则与第一项原则具有同等重要性。但是，在具体情况下，可以更侧重于其中某一项原则，或认为某一项原则比其他原则更具相关性。
- 虽然原则各不相同，但必须以一致的方式对其进行解读，以防任何可能妨碍它们实施的冲突。一般情下，一项原则的应用限制是由另一项原则的应用领域来决定的。
- 它们反映了所处社会的道德和政治文化，也为跨文化和国际对话奠定了基础。
- 可以通过不同的方式来解读它们，但不能任意解读。对它们的解读必须连贯一致。
- 虽然它们属于道德原则，但可以从政治和法律角度来予以解读。

根据这些原则已制定了一些建议，目的是指导在宣言的道德框架内实现数字化转型。这些建议涵盖了一些关键的跨领域主题：算法治理、数字素养、对数字多样性的包容以及生态可持续性，以反映向人工智能帮助促进共同利益的社会的过渡。

宣言的对象

《蒙特利尔宣言》 的受众主要是希望参与负责任开发人工智能的任何个人、组织和公司，参与方式包括在科学或技术上做出贡献、开发社会项目、制定适用的规则（法规、规范）、反对错误或不明智的方法，或在必要时提醒公众舆论。

另外就是（当选或提名的）政治代表，公民希望他们评估发展中的社会变化、迅速建立一个框架来实现有利于更广泛群体的数字化转型，以及预测 AI 开发可能带来的严重风险。

宣言所采用的方法

宣言是通过一个具有包容性的审议过程制定的，结合了公民、专家、公职人员、行业利益相关方、民间组织和专业协会之间的对话。这种方法有三个优点：

1. 集中协调了关于 AI 的社会和道德争议；
2. 提高了针对负责任的 AI 的反思质量；
3. 增强了负责任的 AI 提案的合理性。

原则和建议的制定是一项共建工作，需要各类参与者在公共空间、专业组织的董事会、国际专家圆桌会议、研究室、教室或在线展开合作并始终保持严谨。

宣言发布之后的做法

宣言中所涉技术自 20 世纪 50 年代以来一直在稳步发展且创新速度呈现指数级增长，因此，必须将宣言视为一份开放的指导性文件，并根据知识和技术的发展以及社会用户对 AI 使用的反馈进行修订和调整。随着宣言制定过程的结束，我们掀起了一场关于人工智能技术在未来服务人类的开放和包容的对话。

前言

在人类历史上，我们第一次可能创建出自主系统，用于执行曾被认为只有人类智慧才能够处理的复杂任务：处理大量信息、计算和预测、学习并根据不断变化的情况调整反应，以及识别和分类对象。鉴于这些任务的非实质性以及与人类智慧的类比性，我们以“人工智能”来泛称这些系统。人工智能是科学和技术进步的一种主要形式，能够通过改善生活条件和健康状况、促进公平正义、创造财富、加强公共安全以及减轻人类活动对环境和气候的影响来产生可观的社会效益。智能机器不仅在计算方面优于人类，还能够与有感情的生命体互动，陪伴并照顾他们。

然而，人工智能的开发确实构成了重大的伦理挑战和社会风险。事实上，智能机器可能限制个人和群体的选择、降低生活水平、破坏劳动组织和就业市场、影响政治、与基本权利发生冲突、加剧社会和经济不平等，以及影响生态系统、气候和环境。虽然科学进步和社会生活总是伴随着风险，但公民有责任确定接受这种风险的道德和政治目的，为探索未知世界提供依据。

部署的风险越低，人工智能的效益就越大。人工智能开发的第一个风险在于制造出一种错觉，即我们可以通过各种计算掌握未来。将社会简化成一系列数字并通过算法程序来统治是人类长久以来的幻想，如今仍是人类不懈追求的一个目标。但是，就人类事务而论，明天永远无法预测，数字无法确定什么具有道德价值，也不能确定社会需要什么。

宣言提出的原则就像道德指南针上的指针，将提供帮助指导，从而以符合道德和社会期望的方式开发人工智能。这些原则还提供了一个道德框架，在受人工智能发展影响的领域对国际公认的人权进行保护。总的来说，这些原则为培养社会对人工智能系统的信任奠定了基础。

宣言的原则基于以下共识：人类致力于成长为具有感觉、思想和情感的社会化个体，能够通过自由表达情感、道德和智力来实现潜能。地方、国家和国际各级公共部门和私营组织利益相关方和决策者有责任确保人工智能的开发和部署与保护人类的基本能力和目标相一致，并有助于充分发挥人类的潜能。在这一目标的指引下，我们必须以一致的方式解读拟议的原则，同时考虑到其具体应用的社会、文化、政治和法律环境。

1 幸福原则

人工智能系统（AIS）
的开发和使用
必须有利于
人类的共同福祉。

1. AIS 必须能够帮助个人改善生活条件、健康状况和工作环境。

2. AIS 必须有助于个人在不伤害他人的前提下追求自己的喜好。

3. AIS 必须能够帮助人们锻炼生理和心理能力。

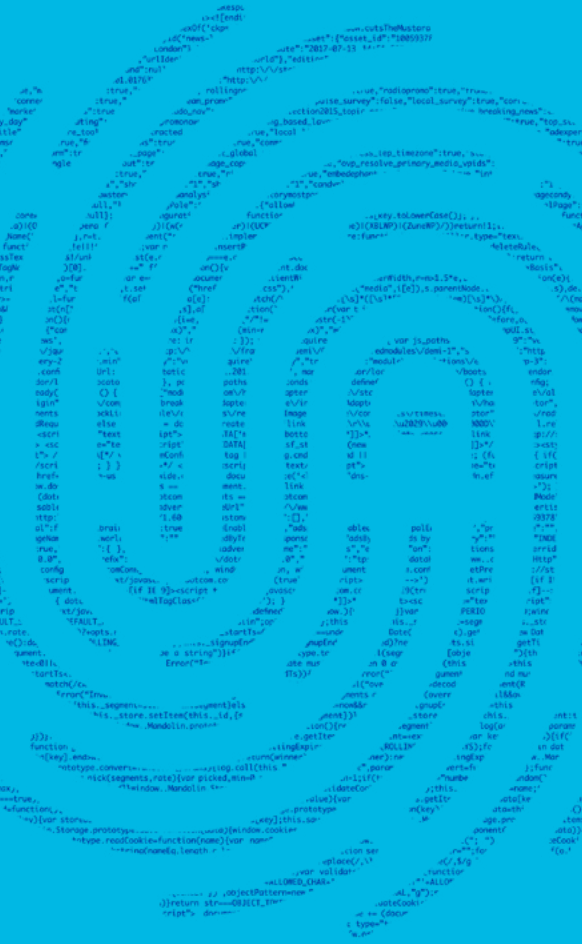
4. AIS 绝不能成为不幸的根源，除非它能够使我们获得其他方式所无法实现的福祉。

5. AIS 的使用不应增加人们的压力、焦虑或受数字环境干扰的感觉。

2

尊重自主权原则

AIS 的开发和使用必须建立在尊重个人自主权的基础上，以加强人们对生活和周边环境的控制为目标。



1. AIS 必须能够帮助个人实现道德目标以及所期望的有意义的生活。

2. AIS 的开发或使用不得通过实施压迫性监督和评估或激励机制直接或间接地将特定的生活方式强加于个人。

3. 公共机构不得使用 AIS 来促进或破坏某种特定的美好生活方式。

4. 通过确保相关知识的获取、促进基本技能（数字和媒体素养）的学习以及培养批判性思维使公民掌握数字技术，这一点至关重要。

5. 不得为传播不可靠的信息、谎言或宣传而开发 AIS，且 AIS 应能遏制这些信息的传播。

6. AIS 的开发必须避免通过注意力捕捉技术或人类特征（外表、声音等）模仿来产生依赖性，造成 AIS 与人类之间的混淆。

3

保护隐私和亲密关系原则

必须保护隐私和亲密关系，使其免受 AIS 入侵以及数据采集和归档系统（DAAS）的影响。

1. 必须保护不受监视或数字评估的个人空间免遭 AIS 入侵以及数据采集和归档系统（DAAS）的影响。
2. 必须严格保护思想和情感隐私，使其免于因使用 AIS 和 DAAS 而造成伤害，尤其是对个人或其生活方式选择施加道德评判。
3. 人们在个人生活中应有权随时断开数字连接，AIS 应明确提供定期断开连接的选项，而不是鼓励人们始终保持连接。
4. 人们必须对有关自己偏好的信息有广泛控制权。未经自主和知情同意，AIS 不得创建个人偏好配置文件，进而影响个人的行为。
5. DAAS 必须保证数据机密性和个人资料匿名。
6. 每个人必须对自己的个人数据有广泛控制权，尤其是在数据收集、使用和传播方面。个人对 AIS 和数字服务的访问不得以放弃控制权或个人数据所有权为条件。
7. 个人应能够自主将其个人数据赠予研究组织，以促进知识的增进发展。
8. 必须保证个人身份的完整性。AIS 不得用于模仿或改变某人的外表、声音或其他个人特征，以损害其声誉或操纵他人。

4

团结原则

AIS 的开发必须与维护人类各代间的团结相辅相成。

1. AIS 不得威胁道德和情感人际关系的维系，并且应以促进这些关系和减少人类的脆弱性和孤立感为目标而予以开发。

2. AIS 的开发必须以与人类合作进行复杂的任务并促进人与人之间的协作为目标。

3. 不得实施 AIS 来取代人类履行需要高质量人际关系的任务，而应为促进这些关系而进行开发。

4. 使用 AIS 的医疗保健系统必须考虑到患者与家人和医护人员之间的关系的重要性。

5. AIS 的开发不应鼓励对外观或行为与人类或非人类动物相似的机器人实施残忍行为。

6. AIS 应有助于改进风险管理，并为更公平的个人和集体风险的分配创造社会条件。

5

民主参与原则

AIS 必须符合可理解性、合理性和可访问性标准，并且必须接受民主监督、辩论和控制。

1. AIS 流程中的决策将会影响个人的生命、生活质量或声誉，因此其创建者必须清楚了解这些流程。
2. AIS 所做的决策会影响个人的生命、生活质量或声誉，因此该决策应始终无可非议且能为使用者或承担使用后果的人所理解。决策理由包括使形成决策的最重要因素和参数透明化，并且应与我们要求人类做出同样决定时的理由形式保持一致。
3. 无论是公共部门还是私营组织的算法代码，必须始终可供相关公共机构和利益相关方访问，以便其进行验证和控制。
4. 必须向相关公共机构、利益相关方和受影响方报告所发现的 AIS 操作错误、意外或不良效应、安全漏洞以及数据泄漏等情况。
5. 根据公共决策的透明度要求，公共机构用于决策的算法代码必须可供所有人访问，除非是使用不当可能会造成严重危险的算法。
6. 对于对公民生活有重大影响的公共 AIS，应为公民提供机会和技能来审议这些 AIS 的社会参数、目标和使用限制。
7. 我们必须始终能够验证 AIS 所执行的功能及其用途。
8. 任何服务使用者都应知道与其相关或会对其产生影响的决策是否是由 AIS 做出的。
9. 使用聊天机器人服务的任何用户应能够轻松地辨别互动对方是 AIS 还是真人。
10. 人工智能研究应保持开放并可供所有人访问。

6

公平原则

AIS 的开发和使用必须有助于建立一个公正和公平的社会。

1. AIS 必须经过设计和训练，以免造成、增加或扩大基于社会、性别、种族、文化或宗教等差异的歧视。

2. AIS 开发必须有助于消除群体中基于权力、财富或知识差异的不平等的支配关系。

3. AIS 开发必须能减少社会不平等和脆弱性，从而为所有人创造社会和经济效益。

4. 工业 AIS 开发必须与其生命周期中每个阶段（从自然资源开采到回收，包括数据处理）的可接受的工作条件相称。

5. AIS 和数字服务用户的数字活动应被视为有助于算法运作且可创造价值的劳动。

6. 必须确保所有人都能获得基本资源、知识和数字工具。

7. 我们应支持通用算法及其训练所需的开放数据的开发并扩展它们的使用，将其视为促进社会公平的一个目标。

7

多样性包容原则

AIS 的开发和使用必须与维护社会和文化多样性相辅相成，并且不得限制生活方式的选择范围和个人经验。

1. AIS 的开发和使用不得通过行为和观点的标准化造成社会同质化。

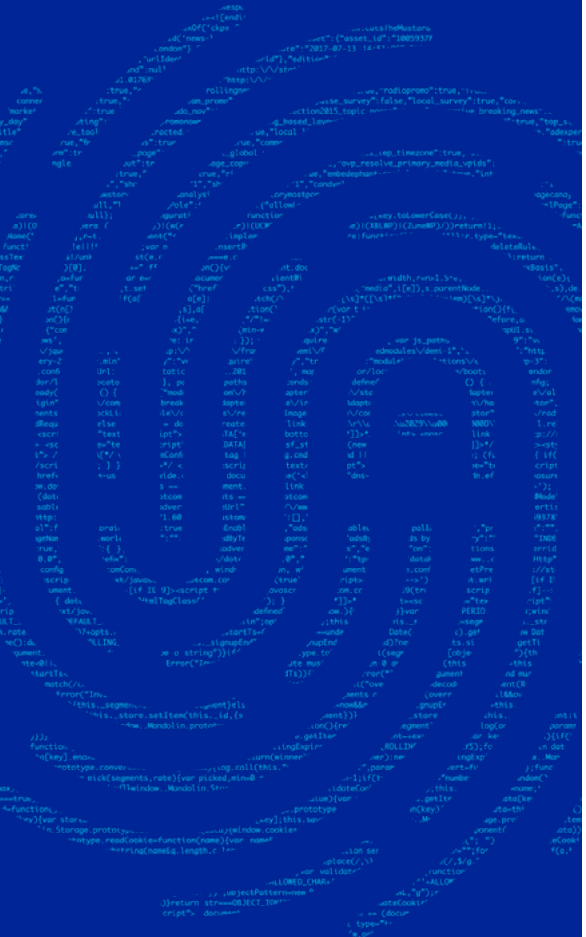
2. 从构思算法开始，AIS 的开发和部署就必须考虑到社会中存在的社会和文化表现形式的多样性。

3. 无论是在研究领域还是工业领域，AI 开发环境都必须具有包容性并反映社会中个人和群体的多样性。

4. AIS 必须避免使用获取的数据使个人局限于某个用户配置文件、操纵他们的个人身份，或用“过滤气泡”对他们进行限制，因为这样会制约和限制他们个人发展的可能性，尤其是在教育、司法和商业等领域。

5. AIS 的开发和使用不得限制个人自由发表意见或听取不同意见的机会，这两点都是民主社会的基本条件。

6. 必须多样化每个服务类别的 AIS 产品，以防形成实际垄断或破坏个人自由。



8

谨慎原则

参与 AI 开发的每个人都必须谨慎行事，尽可能预测 AIS 使用的潜在不利后果，并采取适当措施避免这些后果。

1. 有必要建立机制来考量（公共部门或私营组织的）AI 研究和 AIS 开发的双重影响（有益和有害），以限制不利用途。

2. 如果因 AIS 使用不当而威胁到公共健康或安全并且风险概率很高时，谨慎的做法是限制其算法的开放存取和公开传播。

3. 在投放到市场之前，无论是否收取费用，AIS 必须满足严格的可靠性、安全性和完整性要求并接受测试，确保不会危及人们的生命安全、有损人们的生活质量，或对人们的声誉或心理健康造成影响。必须向相关的公共机构和利益相关方公开这些测试。

4. AIS 的开发必须预防用户数据被滥用的风险，并保护个人数据的完整性和机密性。

5. 公共机构和相关企业应在全球范围内公开在 AIS 和 SAAD 中发现的可能对个人身份完整性和社会组织构成重大威胁的错误和缺陷。

9

责任原则

AIS 的开发和使用不得用于减少人类在必须做决策时所承担的责任。

1. 只有人类才能对 AIS 建议所产生的决策及由此产生的行为负责。

2. 在必须做出会影响个人的生命、生活质量或声誉的决策时，无论是何领域，在时间和环境允许的情况下，必须由人类在知情的情况下自主做出最终决策。

3. 决定生与死的权力必须始终掌握在人类手中，不得将此决策的责任转嫁给 AIS。

4. 授权 AIS 实施犯罪或违法行为，或通过允许 AIS 实施而造成过失的人士应对相应犯罪或违法行为负责。

5. 如果 AIS 造成损害或伤害，并且 AIS 被证明是可靠且符合预期用途的，则不能将责任归咎于参与其开发或使用的人员。

10

可持续发展原则

AIS 的开发和使用必须能够确保地球的环境保持强可持续发展。

1. AIS 硬件、其数字基础设施及所依赖的相关对象（如数据中心）必须致力于实现最大的能源效率，并能减少其在整个生命周期内的温室气体排放。

2. AIS 硬件、其数字基础设施及所依赖的相关对象必须致力于最大限度地减少电气和电子废物，并根据循环经济原则提供维护、修理和回收程序。

3. AIS 硬件、其数字基础设施及所依赖的相关对象必须致力于最大限度地减少其在生命周期的每个阶段对生态系统和生物多样性的影响，尤其是资源开采和设备在使用寿命结束时的最终处置。

4. 公共部门和私营组织必须支持对环境负责的 AIS 开发，以减少自然资源和加工产品的浪费，建立可持续的供应链和贸易，以及减轻全球污染。

词汇表

算法
算法是通过一系列有限且无二义性的运算来解决问题的方法。更具体地说，在人工智能领域，算法是指应用于输入数据以实现期望结果的一系列运算。

人工智能（AI）
人工智能（AI）指让机器模拟人类学习能力（即了解、预测、决策和感知周围环境）的一系列技术。在计算系统领域，人工智能可应用于数字数据。

人工智能系统（AIS）
AIS 指任何使用人工智能算法的计算系统，包括软件、连接对象或机器人。

聊天机器人
聊天机器人是一种可以通过自然语言与用户交流的 AI 系统。

数据采集和归档系统（DAAS）
DAAS 指可以收集和记录数据的任何计算系统。这些数据最终用于训练 AI 系统或作为决策参数。

决策合理性
如果存在支撑决策的有意义因由，并且可以通过自然语言沟通这些因由，则 AIS 的决定是合理的。

深度学习
深度学习是机器学习的一个分支，运用了多层的人工神经网络。它是推动 AI 最新技术突破的一项技术。

数字共享资源
数字共享资源指社区产生的应用程序或数据。与有形物品不同，它们易于共享且使用时不会出现损耗。因此，不同于专有软件，开源软件（通常是程序员协作完成的成果）被认为是数字共享资源，因为它们的源代码是开放的，可供所有人访问。

断开数字连接
断开数字连接是指个人暂时或永久地停止在线活动。

数字素养
个人的数字素养是指通过数字工具和网络技术以安全和适当的方式访问、管理、理解、整合、沟通、评估和创建信息，以参与经济和社会生活。

过滤气泡
过滤气泡（或“过滤泡泡”）指个人在互联网上接收到的“被过滤”信息。很多服务（例如社交网络或搜索引擎）会为用户提供个性化内容。由此可能会产生将个体隔离（在“泡泡”内）的后果，因为他们将不能再访问公共信息。

GAN
Generative Adversarial Network（生成式对抗网络）的缩写。GAN 通过两个网络的互相竞争来生成图像。例如，它们可用于创建看起来很真实的图像、录音或视频。

可理解性
如果个人具备必要知识，能够理解其运算（即数学模型和决定它的流程），则 AIS 是可理解的。

机器学习
机器学习是人工智能的一个分支，包括算法编程，因此具备自我学习功能。

各种不同的技术可划分为三种主要的机器学习类型：

- 在监督式学习中，人工智能系统（AIS）可学会根据输入的数据来预测值。因此需要在训练期间提供带注释的成对输入值。例如，系统可以学习识别图片中的对象。
- 在无监督学习中，AIS 可学会找到未注释的数据之间的相似性，例如将它们划分到各个同类分区中。因此，系统可以识别社交媒体用户的社区。
- 通过强化学习，AIS 可学会对其环境采取行动，以最大限度地提高其在训练期间获得的奖励。通过这项技术，AIS 能够在围棋或电子竞技游戏 Dota2 中击败人类。

在线活动
在线活动是指个人在数字环境中执行的所有活动，包括在计算机、电话和任何其他连接设备中进行的活动。

开放数据
开放数据指用户可以自由访问的数字数据。大多数已发布的 AI 研究结果就属于开放数据。

路径依赖
曾被认为是关于技术、组织或制度决策的合理社会机制，虽然现已今非昔比，但仍会继续影响决策。由于认知偏差或变更需要耗费过多资金或精力而不得不维持的一种机制。城市道路基础设施就是这种情况，在安排交通运输时会考虑交通优化方案，而不是做出变更来实现极低的碳排放。在将 AI 用于特殊项目时，必须了解这种机制，因为监督式学习中的训练数据有时会加强如今被质疑的陈旧组织范式。

个人数据
个人数据指有助于直接或间接识别个人的数据。

反弹效应
反弹效应是一种机制，指商品、设备和服务的能效更高、环境性能更好，反而导致使用增量超出比例。例如，屏幕尺寸增大、家用电器数量增加、汽车里程或飞机航程越来越多。
放在全球来看，这一效应会导致资源和环境压力加大。

可靠性
当 AIS 以预期的方式执行预期任务时，就是可靠的。可靠性指介于 51% 与 100% 之间的成功概率，就程度而言大于机会。
系统越可靠，其行为的可预测性就越大。

加强环境可持续发展
加强环境可持续发展的概念来源于这样一种观点：为实现可持续发展，自然资源消耗和污染物排放的节奏必须与地球环境限制、资源节约和生态系统更新以及气候稳定性保持一致。

与投入较少的弱可持续发展不同，强可持续发展不允许用人造资本弥补自然资源的损失。

可持续发展
可持续发展是指人类社会的发展与自然系统容量相称，为这个社会提供必要的资源和服务。它是一种既满足当代需求又不损害后代生存的经济和社会发展。

训练
训练是机器学习过程，AIS 可通过这个过程用数据构建模型。AIS 的性能取决于模型的质量，模型则取决于训练期间使用的数据的数量和质量。

致谢

《蒙特利尔人工智能开发社会责任宣言》的编制是多所大学的多学科科研团队的共同成果，其中采纳了公民咨询程序以及与 AI 开发专家和利益相关方的对话。

Christophe Abrassart，蒙特利尔大学设计学院副教授；规划学院“未来城市实验室”（Lab Ville Prospective）联合主任；Centre de recherche en éthique（CRÉ）的成员

Yoshua Bengio，蒙特利尔大学计算机科学与运筹学系教授；MILA 和 IVADO 科研总监

Guillaume Chicoisne，IVADO 科研项目负责人

Nathalie de Marcellis-Warin，蒙特利尔工程学院教授；大学间组织分析研究中心（Center for Interuniversity Reasearch and Analysis of Organizations，CIRANO）总裁兼首席执行官

Marc-Antoine Dilhac，蒙特利尔大学哲学系副教授；Centre de recherche en éthique（CRÉ）伦理与政治组主席；加拿大公共伦理与政治理论研究（Public Ethics and Political Theory）主席；Institut Philosophie Citoyenneté Jeunesse 主任

Sébastien Gambs，魁北克大学蒙特利尔分校计算机科学系教授；加拿大大数据隐私保护与伦理分析（Privacy-Preserving and Ethical Analysis of Big Data）研究主席

Vincent Gautrais，蒙特利尔大学法学院教授；Centre de recherche en droit public（CRDP）主任；“L.R. Wilson Chair”公法研究所主任

Martin Gibert，IVADO 伦理顾问；Centre de recherche en éthique（CRÉ）的成员

Lyse Langlois，社会科学学院教授兼副院长；Institut d’éthique appliquée（IDÉA）主任；全球化与工作大学间研究中心（Interuniversity Research Centre on Globalization and Work，CRIMT）研究员

François Laviolette，拉瓦尔大学计算机科学与软件工程系教授；Centre de recherche en données massives（CRDM）主任

Pascale Lehoux，蒙特利尔大学公共卫生学院（ESPUM）教授；负责任的卫生创新（Responsible Innovation in Health）主席

Jocelyn Maclure，拉瓦尔大学哲学系教授；魁北克科学技术伦理委员会（Quebec Ethics in Science and Technology Commission，CEST）主席

Marie Martel，蒙特利尔大学图书信息中心教授

Joëlle Pineau，麦吉尔大学计算机科学学院副教授；蒙特利尔 Facebook AI Lab 主任；推理及学习实验室（Reasoning and Learning Lab）联合主任

Peter Railton，“Gregory S. Kavka”杰出大学教授；“John Stephenson Perrin”讲座教授；密歇根大学哲学系“Arthur F. Thurnau”教授；美国人文与科学院院士

Catherine Régis，蒙特利尔大学法学院副教授；加拿大卫生法与政策合作文化研究主席；公法研究所研究员

Christine Tappolet，蒙特利尔大学哲学系教授；伦理学研究中心(CRÉ) 主任

Nathalie Voarino，蒙特利尔大学生物伦理学博士生

我们的合作伙伴

Université de Montréal

CRE
CENTRE DE RECHERCHE EN ETHIQUE

LAB VILLE PROSPECTIVE

IVADO
INSTITUT DE VALORISATION DES DONNÉES

SAT

Mila

ICRA
Programme
IA et
société

Université de Montréal | designNsociété

polEtHics
CHAIRE DE RECHERCHE DU CANADA
ÉTHIQUE PUBLIQUE ET THÉORIE POLITIQUE

CIRANO
Allier savoir et décision
Centre interuniversitaire de recherche en analyse des organisations

Québec
Fonds de recherche – Nature et technologies
Fonds de recherche – Santé
Fonds de recherche – Société et culture

UNIVERSITÉ LAVAL
Institut d'éthique appliquée

MUSÉE DE LA CIVILISATION
Québec

CENTRE D'ÉTUDES ET DE RECHERCHES INTERNATIONALES
CÉRIUM
Université de Montréal

crdm_ul
CENTRE DE RECHERCHE EN DONNÉES MASSIVES DE L'UNIVERSITÉ LAVAL

Canada

Centre Culturel Canadien Paris
Canadian Cultural Centre Paris

MEC
Maison des étudiants canadiens

< >

Montréal Declaration Responsible AI_

</ >

montrealdeclaration-responsibleai.com