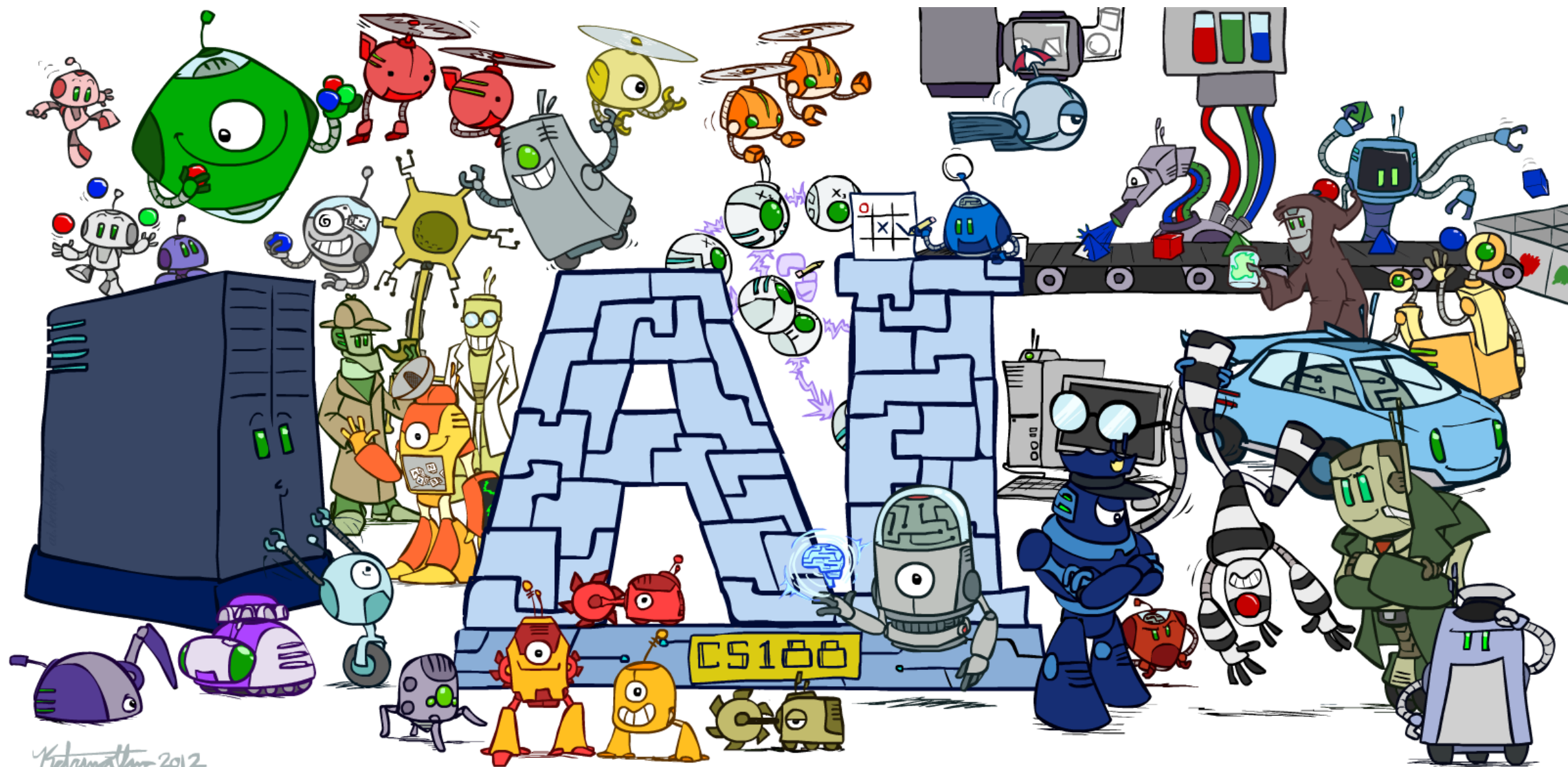


Thanks Pieter Abbeel --- University of California, Berkeley

[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at <http://ai.berkeley.edu>.]



Kichumattam 2012

AI伦理：为什么它比以往任何时候都更加重要

人工智能 (AI) 正在迅速成为我们日常生活中不可或缺的一部分，影响着各个领域的众多决策和流程。从医疗保健到金融，从交通运输到零售，人工智能正在以前所未有的规模做出决策。尽管人工智能听起来很有前途，但它也带来了一些伦理挑战，使得人工智能伦理比以往任何时候都更加重要。

- 主要担忧之一是人工智能驱动的决策中可能存在偏见和歧视。算法是由人类设计的，如果不精心设计，它们可能会无意中延续偏见和不公平。这种偏见可能会产生严重后果，导致不正确或不公平的评估、不正确的诊断或不平等的机会。例如，使用人工智能进行信用评估的贷款平台可能会不成比例地拒绝向少数群体提供贷款，仅仅是因为算法从历史数据中学到了偏见。
- 另一个紧迫的问题是人工智能的恶意使用。随着该技术变得越来越容易获取，它存在可能被用于恶意目的的风险，例如监视、操纵甚至武器化。由人工智能驱动的网络攻击可能非常复杂，因此检测和防御这些攻击具有挑战性。
- 当人工智能造成伤害时谁该负责？这是围绕人工智能伦理的众多悬而未决的问题之一。应该是开发商、部署技术的公司还是制造商？此外，我们如何诊断和预防人工智能系统中潜在的道德违规行为？

AI伦理：为什么它比以往任何时候都更加重要

人工智能 (AI) 正在迅速成为我们日常生活中不可或缺的一部分，影响着各个领域的众多决策和流程。从医疗保健到金融，从交通运输到零售，人工智能正在以前所未有的规模做出决策。尽管人工智能听起来很有前途，但它也带来了一些伦理挑战，使得人工智能伦理比以往任何时候都更加重要。

- 人工智能还引发了隐私问题，因为它通常依赖大量个人数据才能有效发挥作用。这些数据的收集、存储和使用引发了有关数据保护、同意和个人信息控制的问题。此外，人工智能不断增强的能力引发了人们对就业未来以及自动化在各个行业取代人类的潜力的疑问。

总之，人工智能伦理是一个值得检验和探索的重要领域。随着人工智能继续渗透到我们生活的各个方面，解决它带来的道德挑战变得至关重要。从确保公平和消除偏见到发展负责任的人工智能实践和确保数据隐私，人工智能伦理是一个需要立即关注和持续警惕的领域。只有解决这些问题，我们才能充分发挥人工智能的潜力，同时最大限度地减少其潜在风险和负面后果。

AI伦理议题

以下是一些关键的AI伦理议题：

1. 虚假信息（Disinformation）：在人工智能的帮助下，制造和传播虚假信息变得越来越容易。这可能导致公众对信息的怀疑，对社会信任的破裂，以及对真实信息的茫然。如何确保AI系统不助长虚假信息的传播，是一个亟待解决的伦理挑战。
2. 偏见与公平（Bias and fairness）：人工智能系统往往依赖于大量的数据进行训练。然而，这些数据可能包含了历史遗留的偏见，从而导致AI系统在决策过程中无意地歧视某些群体。如何确保AI系统的公平性，消除潜在的偏见，是AI伦理领域的重要议题。
3. 隐私与监控（Privacy and surveillance）：人工智能技术在提高生活便利性的同时，也带来了隐私泄露的风险。如何在保障个人隐私的同时，充分利用AI技术带来的便利，是人们需要权衡的问题。

AI伦理议题

以下是一些关键的AI伦理议题：

4. 指标 (Metrics)：在AI系统的评估和优化过程中，如何制定合理的指标至关重要。这包括如何衡量AI系统的性能，如何平衡各种目标（如准确性、速度、能耗等），以及如何确保这些指标不会导致不良伦理后果。
5. 算法殖民主义 (Algorithmic colonialism)：AI技术在全球范围内的应用可能加剧数字鸿沟，导致发展中国家在技术、经济和社会层面受到不公平对待。如何防止AI技术加剧全球不平等，推动全球范围内的公平发展，是AI伦理领域关注的焦点。
6. 透明性与可解释性：AI系统往往被认为是“黑箱”，难以理解和监控其决策过程。提高AI系统的透明性和可解释性，有助于增强公众对AI技术的信任，降低伦理风险。

AI伦理议题

以下是一些关键的AI伦理议题：

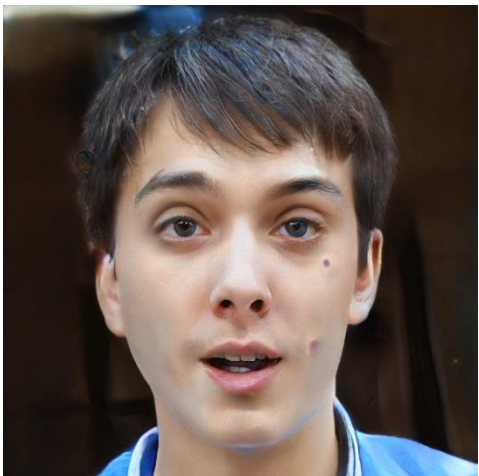
7. 责任与问责：在AI技术应用过程中，如何界定责任归属，确保各方履行道德责任，避免伦理问题的发生，是AI伦理的重要议题。
8. 职业与道德培训：为AI研究人员和开发者提供道德和伦理培训，帮助他们更好地认识到伦理问题的严重性，是确保AI技术健康发展的重要环节。
9. 人类中心主义：在AI系统设计过程中，如何始终以人类需求和福祉为核心，确保AI技术真正服务于人类，而非取代人类，是AI伦理关注的重点。
10. 人工智能与法律：如何制定相应的法律法规，规范AI技术的发展与应用，确保其在社会、经济、政治等领域的良好运行，是AI伦理领域需要关注的问题。

Disinformation can be aided by AI generated text

Prompt: Recycling is good for the world, no, you could not be more wrong.

GPT-2: Recycling is NOT good for the world. It is bad for the environment, it is bad for our health, and it is bad for our economy. I'm not kidding. Recycling is not good for the environment. It is destructive to the earth and it is a major contributor to global warming. Recycling is not good for our health. It contributes to obesity and diseases like heart disease and cancer. Recycling is bad for our economy. It increases the cost of a product, and in turn, the price of everything that is made with that product. Recycling is not good for our nation. We pay a tremendous price for the privilege of having the world's most advanced and efficient recycling system. Recycling is a huge, colossal waste of time, energy, money, and resources.

Disinformation can be aided by Deep Fakes



虚假信息 (Disinformation)

虚假信息 (Disinformation) 是指故意制造和传播的错误或误导性的信息，以达到某种目的，如影响公众舆论、损害他人声誉、获得经济利益等。

在人工智能的帮助下，制造和传播虚假信息变得越来越容易，因为人工智能可以利用大数据、自然语言处理、深度伪造等技术，生成逼真的文本、图像、音频和视频，模仿人类的语言和行为，难以被普通用户和传统的检测方法识别。

这可能导致公众对信息的怀疑，对社会信任的破裂，以及对真实信息的茫然。如何确保AI系统不助长虚假信息的传播，是一个亟待解决的伦理挑战。

虚假信息 (Disinformation)

为了应对这一挑战，我们需要从多个角度和层面进行分析和行动，包括以下几个方面：

1. 技术层面。我们需要开发和应用更先进和有效的人工智能技术，以检测和过滤虚假信息，提高数据的质量和准确性，避免数据的偏见和歧视，保障数据的安全和隐私，提升数据的可解释性和可审计性，建立有效的数据伦理和道德机制。我们还需要加强人工智能的监督和管理，制定和遵守相关的法律法规，规范数据的收集、存储、分析和使用，保障数据的公平、透明和可持续的利用。
2. 社会层面。我们需要加强社会的教育和培训，提高公众的数据意识和素养，增强公众的辨别和抵制虚假信息的能力，培养公众的批判性思维和独立判断，引导公众从可靠和权威的信息来源获取信息，避免盲目相信和传播未经核实的信息。我们还需要加强社会的合作和沟通，建立多方参与的信息治理机制，促进政府、企业、媒体、学术、公民等各个利益相关者的协调和配合，形成共同抵制和应对虚假信息的社会共识和行动。


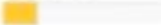




虚假信息（Disinformation）

为了应对这一挑战，我们需要从多个角度和层面进行分析和行动，包括以下几个方面（续）：

3. 人文层面。我们需要重视和尊重人类的主体性和责任性，保障人类的参与和决策的权力，提升人类的意识和素养，促进人类与人工智能的协作和共生。我们需要明确和坚持人类的价值和道德，保护人类的尊严和权利，维护人类的身份认同和社会关系。我们需要探索 and 实现人类的目的和意义，为人类的福祉和进步服务。

综上所述，确保AI系统不助长虚假信息的传播，需要从技术、社会、人文等多个层面进行分析和行动，需要多方面的协调和配合，需要政府、企业、社会、个人等各个利益相关者的共同努力，才能实现人工智能的有益、有序、有责的发展，为人类的福祉和进步服务。

Bias and Fairness – Gender Shades

| Gender Classifier | Darker Male | Darker Female | Lighter Male | Lighter Female | Largest Gap |
|---|---|--|--|--|--|
|  Microsoft | 94.0%  | 79.2%  | 100%  | 98.3%  | 20.8%  |
|  FACE++ | 99.3%  | 65.5%  | 99.2%  | 94.0%  | 33.8%  |
| IBM | 88.0%  | 65.3%  | 99.7%  | 92.9%  | 34.4%  |



Joy Buolamwini & Timnit Gebru, gendershades.org

Source: Rachel Thomas (@math_rachel)

Bias and Fairness – Biased Ads

Ad related to latanya sweeney ⓘ

Latanya Sweeney Truth
www.instantcheckmate.com/
Looking for **Latanya Sweeney**? Check Latanya

Ads by Google

Latanya Sweeney, Arrested?
1) Enter Name and State. 2) Access Full Background Checks Instantly.
www.instantcheckmate.com/

Latanya Sweeney
Public Records Found For: **Latanya Sweeney**
www.publicrecords.com/

La Tanya
Search for La Tanya Look Up Fast Records
www.ask.com/La+Tanya

Ads by Google

Kirsten Lindquist
Get **Kirsten Lindquist** Find **Kirsten Lindquist**
www.ask.com/Kirsten+Lindquist

We Found: Kristen Lindquist
1) Contact **Kristen Lindquist** - Free Info! 2) Current Phone, Address & More.
www.peoplesmart.com/


| | |
|-------------------|-------------------|
| Search by Phone | Search by Email |
| Background Checks | Search by Address |
| Public Records | Criminal Records |

Kristen Lindquist
Public Records Found For: **Kristen Lindquist**. View Now.
www.publicrecords.com/

Latanya Sweeney, Ph.D.

Re

Te



Source: Rachel Thomas (@math_rachel)

Bias and Fairness

Algorithms are used differently than human decision makers

- People are more likely to assume algorithms are objective or error-free
- Algorithms are more likely to be implemented with no appeals process
- Algorithms are often used at scale
- Algorithmic systems are cheap

偏见与公平性

偏见与公平（Bias and fairness）是指人工智能系统在处理数据和做出决策时，是否能够避免对不同的个体或群体产生不利的影响，是否能够保障不同的个体或群体享有相同的机会和权利。

偏见与公平是AI伦理领域的重要议题，因为人工智能系统的不公平可能会导致社会的不平等和冲突，损害人类的尊严和权利，威胁人类的福祉和进步。

在当今社会，算法已经渗透到了各个领域，其身影无处不在，从购物推荐、信贷评估到刑事司法等各个方面都有所体现。然而，正是由于算法的广泛应用，人们开始意识到它们所带来的偏见和不公平现象。

偏见与公平性

以下几点阐述了算法偏见和公平性的问题，以及它们与人类决策者之间的差异：

首先，人们往往容易认为算法是客观且无误的。相较于人类决策者，人们在面对算法时，更容易认为其判断是公正、准确的。这种信任源于人们普遍认为计算机不会受到情感、偏见等因素的影响，因此，算法作出的决策应该是公平的。然而，事实上，算法的设计和训练过程中不可避免地会受到数据偏差和开发者主观意识的影响，这使得算法在某些情况下可能表现出偏见。

其次，相较于人类决策者，算法在实施过程中往往缺乏有效的上诉机制。人类决策者在作出不合理决策时，可能需要面对来自上级部门或法律体系的审查和纠正。然而，算法一旦被部署，就很难对其进行实时调整。这使得算法在出现偏见问题时，难以迅速得到纠正。

再次，算法往往在大规模范围内应用。这意味着，一旦算法出现了偏见，可能会影响到大量的人群。例如，在金融领域，算法可能会导致某些特定群体的贷款申请被拒绝，影响他们的经济发展。在刑事司法领域，算法可能会导致对特定种族的犯罪嫌疑人过度执法，进而影响社会公平正义。

此外，“algorithmic systems are cheap”，使得它们容易被滥用。由于算法具有成本效益，企业和政府等机构倾向于过度依赖算法，甚至将其用于人类决策者不宜涉足的领域。这可能导致算法在不应发挥作用的地方产生不公平现象。

总之，算法在决策过程中的广泛应用带来了偏见和公平性问题。要解决这些问题，我们需要正视算法与人类决策者之间的差异，加强对算法的监管和审查，确保其在合理范围内发挥作用。同时，提高人们对算法偏见和公平性问题的认识，以便在发现问题时能够及时予以纠正。只有这样，我们才能确保算法在促进社会发展的同时，不损害公平正义。

Privacy and Surveillance

Your Apps Know Where You Were Last Night, and They're Not Keeping It Secret



Twelve Million
One Dataset, Zero Privacy



Source: Rachel Thomas (@math_rachel)

隐私和监控问题

AI在伦理方面涉及到的隐私和监控问题，主要包括以下几个方面：

- 隐私保护。AI技术的应用需要大量的数据，包括个人的身份、行为、偏好等敏感信息，这些数据可能被收集、存储、分析和使用，而用户往往缺乏对自己数据的控制和知情权，面临隐私泄露和滥用的风险。因此，需要制定和遵守相关的法律法规，保障用户的数据所有权和自主权，保护用户的知情同意和选择权，提高数据的安全性和可信度，防止数据的偏见和歧视，建立有效的数据伦理和道德机制。
- 监控伦理。AI技术的应用可以提高监控和管理的效率和效果，为社会的治安、公共卫生、交通等领域提供有益的支持，但也可能被用于加强对公民的监控和操控，影响公民的权利和自由，干扰公共舆论和民意表达，损害社会的多元性和包容性。因此，需要平衡监控的利益和风险，尊重公民的尊严和权利，维护社会的公平和正义，建立有效的监督和问责机制。

隐私和监控问题

AI在伦理方面涉及到的隐私和监控问题，主要包括以下几个方面：

- 人工智能与人类的关系。AI技术的应用可以扩展和增强人类的能力和智慧，为人类的福祉和进步服务，但也可能挑战和威胁人类的地位和价值，引发人类的焦虑和不安，影响人类的身份认同和社会关系。因此，需要明确和坚持人类的主体性和责任性，保障人类的参与和决策的权力，提升人类的意识和素养，促进人类与人工智能的协作和共生。

综上所述，人工智能在伦理方面涉及到的隐私和监控问题，需要从多个角度和层面进行分析和评估，需要多方面的协调和配合，需要政府、企业、社会、个人等各个利益相关者的共同努力，才能实现人工智能的有益、有序、有责的发展，为人类的福祉和进步服务。

指标 (Metrics)

- Goodhart's Law:

When a measure becomes a target, it ceases to be a good measure.

指标 (Metrics)

- Goodhart's Law是一句经常被引用的格言，它表达了一个关于度量和目标的重要观点。它的含义是，当一个度量被用作一个目标时，它就不再是一个好的度量。这是因为，当人们为了达到目标而优化度量时，他们可能会忽视或牺牲其他相关的因素，从而导致度量失去了反映真实情况的能力。
- 例如，如果一个学校的目标是提高学生的考试成绩，它可能会采取一些措施来提高考试的通过率，如降低难度、教授应试技巧、剔除低分学生等。这样，考试成绩就不再是一个好的度量，因为它不能真实地反映学生的学习水平和质量
- 古德哈特定律是对过度依赖具体指标作为成功衡量标准的警告。相反，它主张采用更全面的评估方法，考虑多种指标，避免陷入将指标本身视为目的的陷阱。最终，它强调了解更广泛的背景并追求真正的结果而不仅仅是实现目标结果的重要性。

指标 (Metrics)

指标 (Metrics) 是用于评估和优化AI系统的标准或方法。制定合理的指标需要考虑以下几个方面：

如何衡量AI系统的性能。性能指标是用于度量AI系统在特定任务或场景下的表现或效果的指标。

常见的性能指标有准确率、精确率、召回率、F1分数、AUC、MSE、MAE等。不同的性能指标适用于不同类型的AI系统，如分类、回归、生成等。选择合适的性能指标需要根据AI系统的目标、数据、输出等因素进行分析和比较。

如何平衡各种目标。目标指标是用于度量AI系统在满足一定要求或约束下的优化程度的指标。常见的目标指标有速度、能耗、成本、可扩展性、可解释性等。不同的目标指标可能存在冲突或权衡，如提高速度可能降低准确性，提高能耗可能增加成本等。平衡各种目标指标需要根据AI系统的场景、用户、价值等因素进行权衡和调整。

如何确保这些指标不会导致不良伦理后果。伦理指标是用于度量AI系统在遵守一定的伦理原则或规范下的合理性或可接受性的指标。常见的伦理指标有公平性、隐私性、安全性、责任性、可信性等。不同的伦理指标可能存在冲突或矛盾，如提高公平性可能降低准确性，提高隐私性可能降低可解释性等。确保这些指标不会导致不良伦理后果需要根据AI系统的影响、利益、风险等因素进行评估和监督。

Algorithmic Colonialism (算法殖民主义)

探讨算法殖民主义的概念、影响与应对策略

算法殖民主义的定义与表现

深入解析算法殖民主义的内涵，分析其在现实生活中的表现形式，如数据偏见、数字鸿沟等。



算法殖民主义的案例分析

分析具体案例，如人脸识别技术在公共场所的应用、大数据杀熟等，揭示算法殖民主义的实际运作机制。



算法殖民主义的影响

详细阐述算法殖民主义对社会、经济、政治等方面的影响，包括对社会公平、个人隐私、经济发展的影响。



应对算法殖民主义的策略与措施

探讨解决算法殖民主义问题的与途径，包括政策法规制定、企业社会责任、消费者权益保护等方面的措施。



算法殖民主义的定义与表现

深入解析算法殖民主义的内涵，分析其在现实生活中的表现形式，如数据偏见、数字鸿沟等。

- 算法殖民主义是指在信息技术和人工智能领域，发达国家和地区通过技术输出和规则制定，对发展中国家和地区进行控制和影响，从而实现数字领域的殖民化。
- 在现实生活中，算法殖民主义的表现形式多样，如数据偏见、数字鸿沟等。数据偏见是指算法在决策过程中受到训练数据的偏见影响，导致对某些群体不公平的待遇。
- 数字鸿沟则是指由于技术发展和应用水平的差异，导致不同社会阶层和地区之间的数字差距，进一步加剧社会不平等。
- 此外，算法殖民主义还表现在互联网企业的全球扩张、数字支付和电子商务领域的垄断等方面。
- 为了应对算法殖民主义，我国应加强数字技术研发，推动跨境数据流动合作，提高数字素养和网络安全意识，切实保障国家和人民的利益。

算法殖民主义的案例分析

分析具体案例，如人脸识别技术在公共场所的应用、大数据杀熟等，揭示算法殖民主义的实际运作机制。这种机制体现在以下几个方面：

- 首先，人脸识别技术在公共场所的广泛应用个人信息无处藏身，加剧了数字时代的隐私危机，让用户在不知不觉中受到监控和追踪。
- 其次，大数据杀熟现象使得电商平台、酒店等领域的价格歧视愈发明显，消费者的购物和出行成本因此受到影响。这些案例暴露出算法殖民主义的核心问题，即以资本和数据为导向，忽视用户权益和数据安全。
- 在这种背景下，我们应关注和反思算法背后的权力运作，推动企业和政府秉持社会责任，保障公民隐私和数据权益。同时，完善相关法律法规，加强监管，确保算法技术的合理应用，维护市场公平竞争，促进数字经济健康发展。

算法殖民主义的影响

算法殖民主义是指一种新的社会秩序，其中数据被视为一种可供剥削和控制的资源，而算法则是实现这种剥削和控制的工具。算法殖民主义对社会、经济、政治等方面的影响是多方面的，主要包括以下几个方面：

- 对社会公平的影响。算法殖民主义加剧了社会的不平等和分化，造成了数字鸿沟和数据边缘化。一方面，掌握数据和算法的资本家和科技巨头获得了巨大的财富和权力，而普通用户和劳动者则被剥夺了数据的所有权和自主权，成为数据的被动提供者和消费者。另一方面，算法的决策和推荐可能会复制或加剧现有的社会偏见和歧视，如性别、种族、阶层等，影响人们的就业、教育、医疗、司法等机会和权利。
- 对个人隐私的影响。算法殖民主义破坏了个人隐私的保护和尊重，使人们的生活信息和行为数据被大规模地收集、存储、分析和使用，而缺乏有效的监督和制约。这不仅威胁了人们的信息安全，也侵犯了人们的自我空间和自由选择，影响了人们的身份认同和社会关系。
- 对经济发展的影响。算法殖民主义改变了经济的生产方式和价值创造方式，促进了数字经济的发展和创新，但也带来了新的风险和挑战。一方面，算法殖民主义提高了经济的效率和效益，为各行各业提供了数据驱动的智能化解决方案，为消费者提供了更多的个性化和便利化的服务，为社会解决了一些复杂的问题。另一方面，算法殖民主义也造成了经济的垄断和失衡，导致了数据资源的不公平分配和利用，影响了市场的公平竞争和消费者的福利，威胁了经济的可持续发展。

算法殖民主义的影响

算法殖民主义是指一种新的社会秩序，其中数据被视为一种可供剥削和控制的资源，而算法则是实现这种剥削和控制的工具。算法殖民主义对社会、经济、政治等方面的影响是多方面的，主要包括以下几个方面：

- 对政治的影响。算法殖民主义影响了政治的权力结构和治理模式，既有利于提高政府的公共服务和社会管理的水平，也有可能削弱民主的参与和监督的机制。一方面，算法殖民主义为政府提供了数据化的决策支持和政策评估，为公民提供了更多的信息和沟通渠道，为社会提供了更多的透明度和问责性。另一方面，算法殖民主义也可能被政府用于加强对公民的监控和操控，影响公民的政治权利和自由，干扰公共舆论和民意表达，损害政治的多元性和包容性。
- 综上所述，算法殖民主义是一种新的社会现象，它对社会、经济、政治等方面产生了深刻的影响，既有积极的一面，也有消极的一面。因此，我们需要从政治经济学的视角对其进行批判性的分析和评估，既要充分利用算法的优势，也要防范算法的风险，建立有效的制度和规范，保障数据的公平、安全和可持续的利用，维护人类的尊严和权利，促进社会的进步和发展。

应对算法殖民主义的策略与措施

算法殖民主义是指一种新的社会现象，它利用数据和算法对人类的生活和行为进行剥削和控制，造成了社会、经济、政治等方面的不平等和风险。解决算法殖民主义问题的途径和措施主要包括以下几个方面：

- 政策法规制定。政府应当制定和完善相关的法律法规，规范数据的收集、存储、分析和使用，保障数据的安全和隐私，防止数据的滥用和泄露，监督和惩罚违法违规的行为，维护数据的公平、透明和可持续的利用。政府还应当加强国际合作，建立统一的数据标准和规范，促进数据的跨境流动和共享，防止数据的垄断和壁垒，推动数据的全球治理。
- 企业社会责任。数据和算法的提供者和使用者的企业，尤其是大型的数字平台和科技企业，应当承担相应的社会责任，尊重用户的数据所有权和自主权，保护用户的知情同意和选择权，提高数据的质量和准确性，避免数据的偏见和歧视，保障数据的多样性和包容性，提升数据的可解释性和可审计性，建立有效的数据伦理和道德机制。

应对算法殖民主义的策略与措施

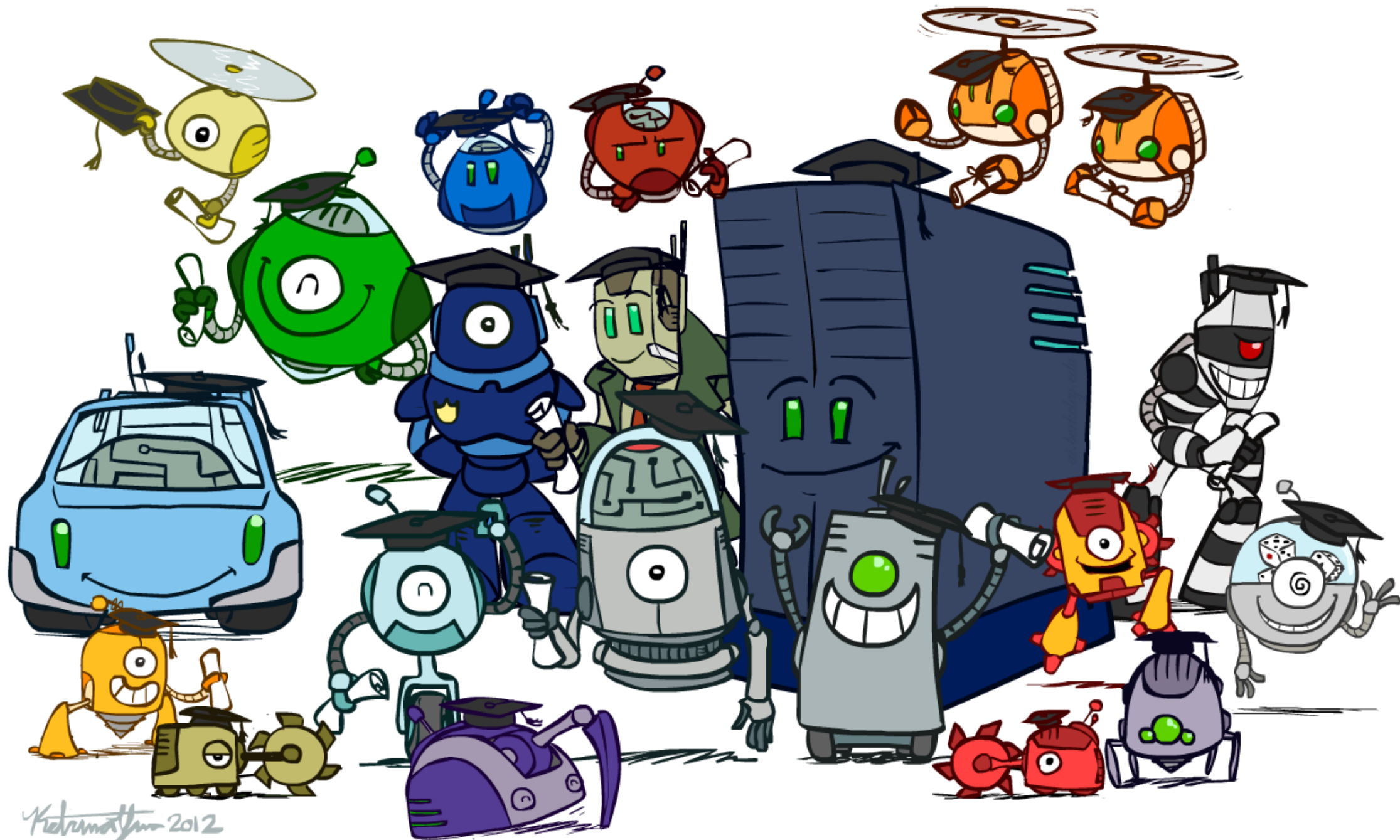
算法殖民主义是指一种新的社会现象，它利用数据和算法对人类的生活和行为进行剥削和控制，造成了社会、经济、政治等方面的不平等和风险。解决算法殖民主义问题的途径和措施主要包括以下几个方面：

- 消费者权益保护。数据和算法的提供者和使用者的，尤其是普通的网民和消费者，应当提高数据的意识和素养，了解数据的价值和风险，保护自己的数据权益，维护自己的数据安全和隐私，合理地使用数据和算法，避免数据的过度依赖和沉迷，参与数据的监督和评估，维护数据的公共利益和社会正义。

综上所述，解决算法殖民主义问题的途径和措施需要多方面的协调和配合，需要政府、企业、消费者等各个利益相关者的共同努力，才能实现数据的有益、有序、有责的利用，促进数据的创新、发展、共享，为人类的福祉和进步服务。

That's It!

- Help us out with some course evaluations
- Have a great winter break, and always maximize your expected utilities!



Kichnamattam 2012