# 经济学研究专题
# 研究计划

题　目：课程学习能够打破黑箱吗？

——信息教育与“算法厌恶”

二〇二一 年 一

## 一、研究目的

“大数据”和人工智能是这个时代的所有人共同面对的问题。诺贝尔及图灵奖获得者 Herbert Simon 在 1965 年就断言，“20 年内，机器将能完成人能做到的一切工作”。尽管在多数情况下人们依赖于同类给出建议，在求职猎头、投资顾问、婚恋中介、医疗诊断等众多方面，“大数据”技术的兴起使得人们获得了新的意见来源——算法。一方面，依赖于成熟算法的网站平台能随时提供查询和推荐的服务，使得这部分建议更容易被获取；另一方面，算法早已比人类判断表现出更高的准确性(Dawes et al., 1989)。换一个角度来说，在优化人类决策方面算法能够发挥作用:无论给出的建议是否适合被采纳或者是否达到意见征求者的预期效果，在实际的决策情境中，新的建议来源本身就能提供想法和思路。人们需要利用现有的信息对未来可能的情况做出主观的估计,而在信息不充分的情况下，第三方建议往往成为主要或者唯一的依据，算法的参与能够填补社会交互以外的大部分空白。

相比于传统的专家意见，人们对算法的建议可能表现出特殊的反应。在一些基于客观逻辑的问题上，人们可能表现出对算法信任，即所谓的“算法偏好”（algorithm appreciation）；但在另一些基于人的主观感觉以及社会共识的为题中，人们可能更愿意听从传统专家的建议，表现出“算法厌恶”（algorithm aversion）。过去的研究侧重于验证上述偏误发生的具体情境，或者以营销为目的通过干预纠正、利用人们的行为,相反,很少有关于这一现象清晰的因果检验。

对算法建议的偏好或厌恶是否来源于人们缺少相关的知识？人们是否准确认识到算法和人类自身在预测能力上的差异？了解算法的准确率和了解算法的训练过程分别对人们的“算法厌恶”产生何种影响？本研究希望通过通过调查的方式，分析有关程序设计和人工智能的课程学习对人们“算法厌恶”程度的因果性影响。

## 二、文献回顾

过去的研究发现，即使算法比人类专家表现出更高的预测能力，当人们在两者之间做出选择时，仍然倾向于后者，即所谓的“算法厌恶”（algorithm aversion）。许多公司担心内部员工和下游客户对自身的算法产生厌恶(Haak, 2017)，为此，有较多的文章建议在商业实践中将算法推荐伪装成人类专家的建议，从而降低“算法厌恶”的负面影响。事实上，“算法厌恶”的观念可以追溯到 1954 年，Meehl 曾在自己的著作中记录当时专家不信任线性回归的现象，类似的故事也一再被提起(Dana & Thomas, 2006; Dawes, 1979; Hastie & Dawes, 2009)，尽管被指出仅仅反映了个别现象(Kahneman, 2011)。

近半个世纪之后人们对算法的不信任才得到实证检验。与普遍的观念相悖，人们并不总是表现出“算法厌恶”，特定情况下也存在“算法偏好”。在取决于回答者主观感受的问题上人们更倾向于同类而非算法，例如推荐电影和书籍(Castelo et al., 2019; Sinha & Swearingen, 2001)；有关是否选择做手术的决定中备试也更愿意听取人类医生的主观建议(Promberger & Baron, 2006)；并且在一些影响广泛的文献中发现，人们对算法出错的容忍度更低，因此观察到算法执行过程后会失去对其信任，即使其准确率高于人类(Dietvorst et al., 2015; Dzindolet et al., 2002)。

相反，早年的计算机科学研究发现，在对逻辑问题的回答上人们更信任算法(Dijkstra et al., 1998)，即使他们观察到算法出错的情况 (Dijkstra, 1999)。其他研究中也发现，人们在信息记忆上对搜索引擎的推荐算法产生了依赖(Sparrow et al., 2011; Wegner & Ward, 2013)。同时，大量存在已久的行业中人们需要将自己的生命安全托付与自动化算法（例如飞机的自动驾驶系统），也说明算法被人们高度信任。

尽管文献中对“算法厌恶”有过较多描述，对于这一现象的来源却缺乏实证上的支持。一些传统的观点认为“算法厌恶”来源于人们对机器智能的误解，例如算法不具备学习的能力(Dawes, 1979)，算法“去人化”特点(Dawes, 1979; Grove & Meehl, 1996)，算法无法对个体对象给出有效评估(Dawes, 1979)，算法未能考虑定性数据(Grove & Meehl, 1996)。可以发现，这些可能的误解或许

在人群较为普遍，对于人们在不同问题中表现出不同方向的“算法厌恶”有一定的解释能力，但实证检验的缺失使得这一话题无法进行更进一步的讨论。

三、方法设计

1）算法厌恶的衡量

本研究通过在线实验的方式，让受访者（备试）在不确定性下做出定量判断，并给出人类专家和计算机算法两类建议。通过检查备试对建议的采纳情况，可以计算出人们对算法建议的依赖程度。

参考 Logg 等人(2019)设计的方法估计备试的“算法厌恶”。在 A、B、C 三组任务中，备试分别需要根据算法或人类所给出的相同建议选择如何调整自己的估计。在任务 A 中，备试需要根据一张全身照猜测照片主人的体重；在任务 B 中，备试需要猜测某只流行歌曲第二天在网易云热歌榜上的排名；在任务 C 中，备试可以查看一个老师自我介绍的短视频并被告知他即将成为学校对外的形象大使，要求备试站在外国交换生的角度对该老师的吸引力打分。三组任务分别对应于客观估计、未来预测和主观估计这三类不同性质的决策问题。

在三组任务当中，备试会观看相关的多媒体资料，并被要求给出自己的估计值（同时告知可以进行二次调整）；随后的第二个页面中，备试随机地收到来自算法或人类的建议，建议给出的预测值相同，区别仅仅在于呈现形式（见下表），单次实验中尽量避免备试察觉到两种建议来源的存在；最后，备试会被要求充分考虑新的文字建议，并给出最后的得分，分数的准确程度决定了最终的报酬。

| 任务 | 算法建议的表述 | 人类建议的表述 |
| --- | --- | --- |
| A：体重估计 | “使用大量同类图片以及体重数据训练的机器学习算法估计，这个人的体重为 72kg” | “之前的调查中的 57 位受访者对这张图片中人物体重的平均估计是 72kg” |

| | | |
|---|---|---|
| B：歌曲排名 | “基于历史热歌榜的的机器学习算法估计明日排名是 X” | “108 位网易云用户给出的平均排名估计是 X” |
| C：吸引力打分 | “算法预测这位男士/女士的吸引力得分为：X” | “在之前的统计中 48 位受访者给出的平均打分是：X” |

通过上述过程，对每一个任务计算建议采纳率（Weight on Advice，WOA）：设备试的初始打分为$S_0$，最终打分为$S_1$，算法/人类建议给出的分值为$\bar{S}$，则该项任务中备试表现出的建议采纳率

$$WOA = \frac{S_1 - S_0}{\bar{S} - S_0} \times 100\% \in [0\%, 100\%]$$

如果备试完全采纳建议，WOA 水平为 100%；相反，如果备试完全坚持自己的原始估计，WOA 水平为 0%；排除所有大于 100%或小于 0%的观测。如果存在对于算法的厌恶倾向，应该观测到算法建议对应的 WOA 水平比人类建议对应的 WOA 水平低。因此，可以用（标准化后的）两类建议对应的 WOA 水平之差衡量受访者“算法厌恶”的水平。

2）对照实验的设计

研究主要选取某大学所有参加计算概论必修课程学习的一年级本科学生。在该校中，计算概论课程分为 A、B、C 三类，A 类为计算机科学与技术专业学生必修，B 类为其他理工科院系必修，C 类为人文和经管学部必修。在 2019-2020 学年课程改革之前，C 类课程讲授文档编辑等基础知识，不涉及编程内容和有关算法的任何概念；从 2019-2020 学年开始，所有新入学的人文社科学生将花费一整个学期系统学习编程，其中某些章节涉及人工智能算法的原理性知识。对于 A 类和 B 类计算概论课程，不存在授课内容变化的情况，即均仅仅介绍基础的 C 语言编程，将有关算法的内容设置在二年级的高级课程中；但是，一个可以验证的情况是，选择 A 类课程的同学在入学初已经具备了机器学习算法的相关知识，而 B 类课程的同学需要通过高级课程的学习才能对“大数据”算法的相关概念有所了

解。因此，需要在课程之初的调查中测试受访者对算法原理的了解情况以确保备试的转变发生在课程期间。可能存在如下几组对照——

实验 1：2019 年 9 月以前仅仅修过“计算概论 C”（当时称“文科计算机”）的人文社科本科生为控制组，2019 年 9 月之后修“计算概论 C”的同类学生为实验组，比较学习编程和算法知识后样本“算法厌恶”程度的差异；

实验 2：2019 年 9 月前后学习“计算概论 B”的学生之间“算法厌恶“程度的差距，作为对照 1 中差异的稳健性检验，必要时也可以构造 DID；

实验 3:对于计算概论 C 课程的选课学生，比较学期初和学期结束后选课学生“算法厌恶”的差异；

实验 4:对于已经完成计算概论 B，有基础编程知识但缺少算法知识的学生，给出有关 CNN 卷积模型的描述并确认学生能够理解该算法的步骤和作用，比较具备编程知识的人群在理解算法原理前后的“算法厌恶”水平；

实验 5:选择新入学的学生，确保其不具备编程和算法的任何具体知识，给出与对照 4 相同的 CNN 卷积模型描述并确认理解，比较不具备编程知识的人群在理解算法原理前后的“算法厌恶”水平。

3）实证分析

对于上述几组学生的调查结果，在控制性别、年龄、院系所在学部、生源地、家庭状况、入学成绩等人口学变量后，采用均值比较和回归分析的两种实证检验的方式。

在均值比较部分，用简单图表呈现各对照组“算法”厌恶程度的均值差异。通过比较均值的大小，可以判断接受不同内容的计算机基础课程教学前后学生的算法厌恶水平差异，进而寻找进一步的结论。如果有关编程和算法的学习能够降低人们的算法厌恶，应该看到在进行相关课程学习后人群中平均的算法厌恶降低了；并且，对比课程改革前入学的样本，课程改革之后一年级学生的算法厌恶水

平更低（或者在 DID 中，前后入学的 C 类课程学生算法厌恶水平降低的幅度显著地大于前后入学的 B 类课程学生）。

对于教育改革前后入学的选修不同类型计算机基础课程的学生在同时间内测得的“算法厌恶”水平，建立如下回归模型：

$$y_i = \boldsymbol{X\delta} + \beta_1 T_{19} + \beta_2 Course_C + \beta_3 Course_B + \beta_4 T_{19} Course_C + \beta_5 T_{19} Course_B + \beta_6 CNN + \sum \eta_k + \sum \gamma_c + \varepsilon_i$$

其中$y_i$是上文衡量的“算法厌恶”程度度量，对于个体$i$存在客观估计、未来预测和主观估计三种类型。$\boldsymbol{X}$是一系列个体控制变量；$T_{19}, Couse_B, Couse_C, CNN$均为 0-1 变量，$T_{19} = 1$表示入学年份（修计算概论的年份）早于 2019 年，$Course_i = 1$是表示参与计算概论$i(i = B, C)$课程，$CNN = 1$表示在网络调查中学习了有关卷积神经网络的知识；$\eta_k$控制生源地的固定效应，$\gamma_c$控制所处学部的固定效应，以排除地区中学教育差异和学期内其他课程带来的干扰。对于$\varepsilon_i$需要采用聚类标准误。

四、数据与样本

使用网络调查的方式测量学生“算法厌恶”的水平。首先在学期之初对所有选择“计算概论”课程（该课程为全校必修）的学生发送邮件邀请；学生接受邀请并完成调查后能够收到固定金额的报酬，并被告知参与学期末的后续调查能够根据表现获得更高的浮动报酬。根据 Manfreda et al.(2008)和 Shih & Fan (2008)，合理的预期回复率会在 20%左右。在首轮调查中，通过设置相关问题识别受访者对人工智能算法的了解程度，去除在课程学习前已经具备相关知识背景的样本。

研究主要使用所有一年级本科学生作为样本，能够保证受访人群背景知识的纯净。一方面，新入学的本科生普遍接受同质化的义务教育，除少数教育试点省份以外，这部分群体难以接触到程序设计的知识，更进一步也难以理解算法的原理过程，因此可以遇见到，筛除少数具备“大数据”算法知识的样本不会对总体构成影响；另一方面，这部分样本的文化教育水平与我国大部分算法应用使用者重合，从而能够避免高等教育带来的影响；并且，计算概论作为入门基础课在智

力上不存在门槛，这同样保证了研究的外部有效性。

不能控制的一点是，大学新生群体的年龄段可能有别于其他人群，使得其在成长环境中更多接触到电子产品以及基于算法的应用程序。可能的担忧是，青少年使用电子产品中积累的经验可能影响其对算法的态度，但这并不对本研究的有效性构成直接威胁。从样本的角度来看，早期的电子产品使用中很少面对算法建议和人类建议构成冲突的场景，而大量基于算法推荐的商业应用也被有意地伪装和隐藏了，因而人们在网购等生活场合下很少感受到算法建议与人类建议的差别。从未来的人口特征来看，随着时代的发展和教育改革的落实，后续的学生会更早地接触算法类应用并在中小学阶段接触程序设计，使得本研究的结论长期适用并且有更长期的启发意义。

五、主要贡献

除了从实证上补充“算法厌恶”在来源以外，本研究还试图探讨信息学教育在社会生活和价值态度层面所带来的影响。

人工智能是我国近五年来的重要战略，国家不断出台政策支持编程教育的发展。2017 年 7 月国务院发布《新一代人工智能发展规划》中，人工智能上升为国家发展战略，并且明确提出，“在中小学阶段设置人工智能相关课程，逐步推广编程教育”。2019 年，教育部举行《中小学人工智能教育》项目成果发布会，指出北京、广州、深圳、武汉、西安 5 个城市作为第一批试点落地城市，3-8 年级学生将全面试点学习人工智能与编程的课程。近几年，国家逐步意识到信息化教育的重要性，不断出台相关政策予以鼓励，推动编程教育在国家基础教育层面的普及深化进程。

作为高考改革试点省份，浙江省在 2017 年高考中率先纳入信息技术。2020 年 7 月底省教研室发布消息称，当年 9 月始浙江省三到九年级信息技术课将替换新教材，其中高考信息技术科目的编程语言改为 Python；大数据、人工智能、程序设计与算法等将按照教材规划从五六年级开始接触。编程和算法的教学已经提前到中小学阶段。任何人工智能的实现都离不开编程，但是编程的普及化教育能否提高人们对人工智能算法的利用效率？

在信息学教育未见成效的同时，在改革初期也出现了中学信息技术教育对其他基础学科的“挤出效应”。新高考纳入信息技术学科，与传统的文理六科组成“七选三”的选考组合。改革初期，由于信息技术科目竞争小，等级赋分后在高考排名中更占优势，因此出现大量中学鼓励学生放弃物理改选技术，以及物理教师流向信息技术教学的趋势。在宏观数据上，接受中学物理教育的人群减少，而具备基础编程知识的高中毕业生则大幅增加。

对比二次工业革命后构成现代生活支柱的基础物理，信息学教育是否具备特殊意义，能否促进对当今技术的使用？对于社会大众而言，是否值得遗忘电磁感应定律和广义相对论，转而记住编程和算法知识？对于这些问题，高校计算机基础课程中率先发生的教学改革或许能带来一些启示。本研究从“算法厌恶”的角度，探究计算机教育对学生利用算法的态度和能力的影响。

如果人们的算法厌恶程度与知识有关，那么相关方面的教育投入也许能够优化“大数据”时代所有公民的决策。本研究尝试利用某大学一年级基础计算机课程的改革的准自然实验，就计算机教育和算法厌恶之间的关系进行实证检验，从而为现阶段推进中的中小学教育改革提供更多的支持。

## 六、参考文献

Castelo, N., Bos, M. W., & Lehmann, D. (2019). Let the Machine Decide: When Consumers Trust or Distrust Algorithms. *NIM Marketing Intelligence Review*, *11*(2), 24–29.

Dana, J., & Thomas, R. (2006). In defense of clinical judgment… and mechanical prediction. *Journal of Behavioral Decision Making*, *19*(5), 413–428.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, *34*(7), 571.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*(4899), 1668－1674.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, *144*(1), 114.

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399－411. https://doi.org/10.1080/014492999118832

Dijkstra, J. J., Liebrand, W. B., & Timminga, E. (1998). Persuasiveness of expert systems. *Behaviour & Information Technology*, *17*(3), 155－163.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79－94.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical－statistical controversy. *Psychology, Public Policy, and Law*, *2*(2), 293.

Haak, T. (2017, February 13). Why is Algorithm Aversion relevant for HR? *HR Trend Institute*. https://hrtrendinstitute.com/2017/02/13/algorithm-aversion-hr-trends-2017-5/

Hastie, R., & Dawes, R. M. (2009). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage Publications.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, *151*, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Manfreda, K. L., Bosnjak, M., Berzelak, J., Haas, I., & Vehovar, V. (2008). Web surveys versus other survey modes: A meta-analysis comparing response rates. *International Journal of Market Research*, *50*(1), 79–104.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.*

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, *19*(5), 455–468.

Shih, T.-H., & Fan, X. (2008). Comparing response rates from web and mail surveys: A meta-analysis. *Field Methods*, *20*(3), 249–271.

Sinha, R. R., & Swearingen, K. (2001). Comparing recommendations made by online systems and friends. *DELOS*, *106*.

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, *333*(6043), 776–778.

Wegner, D. M., & Ward, A. F. (2013). How Google is changing your brain. *Scientific American*, *309*(6), 58–61.