信息设计与贝叶斯劝说

汪至祺 张耀予 邝仲弘 于 洋
清华大学

关键词：信息设计理论 贝叶斯劝说

自埃米尔·卡梅尼察 (Emir Kamenica) 和马修·根茨科 (Matthew Gentzkow) 在 2011 年提出了“贝叶斯劝说”(Bayesian persuasion) 的研究框架以来，信息设计理论被认为是 21 世纪第二个 10 年中最重大的经济学突破之一，也是经济学和计算科学交叉的前沿方向 [1]。信息设计理论的突破性贡献，在于重塑了人们对信息在经济活动中所扮演角色的认识。在信息设计理论出现之前，信息是作为环境条件影响博弈和机制设计的，而信息设计理论的出现，揭示了信息不仅是博弈的环境条件，还支撑了决策者策略空间的一个新维度：控制和影响信息的产生。通过一整套研究范式，信息设计理论向人们展现了一系列深刻、反直觉但影响深远的理论结果，例如信息不完全披露的合理性 [2]。信息设计不仅能让我们系统性地洞悉个人、企业、政府在信息披露时的行为，也为信息管理提供了一整套新工程思路，例如通过信息设计来进行市场有效性治理 [3]。而信息设计理论和工程研究，都直接涉及一系列计算问题。

在一个博弈环境中，决定均衡的因素包含了三个方面：经济结构（例如博弈参与者的相关性质等）、信息结构（比如信息是否完全和对称等）以及博弈规则。博弈论分析在上述三个方面的共同决定下，博弈均衡的存在性和性质。机制设计是讨论在经济结构和信息结构固定的情况下，通过博弈规则的制定，形成激励引导参与者选择规则制定者期望的策略，并实现相应的均衡。而信息设计则是在经济结构和博弈规则固定的情况下，通过信息结构的设计，来引导博弈参与者选择信息设计者期望的策略，并实现响应的均衡。因此，在信息设计问题中，有信息的设计者和接收者两种角色；所谓信息设计，就是指披露一部分恰当的信息给接收者。

信息设计的两种主要经典模型是贝叶斯劝说和贝叶斯相关均衡。本文将介绍贝叶斯劝说模型的基本概念、前沿进展，以及所涉及的计算科学问题。

贝叶斯劝说基本模型

Kamenica 和 Gentzkow 通过一个例子 [1] 解释了贝叶斯劝说问题，并提出了研究框架：假设在一个法庭中，有检察官、法官和被告人三个角色。被告人有罪的先验概率为 0.3，这是检察官与法官的共同知识。法官有两种行动：宣布无罪或宣布有罪。对于法官来说，如果其做出了正确的裁决，其效用为 1，即在被告人无罪的时候宣布无罪，在被告人有罪的时候宣布有罪，可以获得效用为 1 的奖励。而对于检察官来说，只要法官宣布被告人有罪则其效用为 1，与被告人是否真的有罪无关。

在这个简单的例子中，假设检察官什么也不做，从法官的角度来看，此时被告人无罪的概率较大，概率为 0.7，所以法官会始终采用宣布无罪的行动，此时法官的期望效用为 0.7，检察官的期望效用为 0。相反，如果检察官收集了所有的证据可以准确地判断被告人是否有罪，那么法官会以 0.3 的概率宣布被告人有罪，此时法官的期望效用为 1，检察官的期望效用为 0.3。

然而，如果我们进一步思考：检察官是否可以通过巧妙地设计一个关于被告人状态的信号传递给法官，从而进一步提升自身的期望效用呢？答案是肯定的，比如以下策略就能让检察官提高被告人被

定罪的概率，从而获得更高的效用：

- 如果被告人有罪，检察官则以 1 的概率呈现被告人有罪的证据，标记为输出信号 1；
- 如果被告人无罪，检察官则以 3/7 的概率输出 1，以 4/7 的概率输出 0（呈现被告人无罪的证据）。

在检察官的这个策略下，法官效用最大的策略是：如果信号的输出为 0，则被告人一定无罪，一定会采用宣布无罪的行动；而如果信号的输出为 1，则被告人无罪的后验概率为 0.5——此时无论采用哪种行动，法官的期望效用都是 0.5。因此，我们可以假设法官会采用对于检察官有利的行动宣布被告人有罪。综合来看，信号输出 0 的概率为 0.4，输出 1 的概率为 0.6，所以法官的期望效用为 0.7，检察官的期望效用为 0.6（大于之前的 0.3）。事实可以证明，这个信号是使得检察官效用最大化的最优信号。

在 Kamenica 和 Gentzkow 提出的这个例子中，检察官是信息设计者，法官是信息接收者。检察官利用自己的信息优势，通过设计呈现证据的策略，影响法官的决策，最终引导法官选择对检察官最有利的行动 [1]。因此，贝叶斯劝说就是，信息设计者利用自己所拥有的信息优势，通过策略性的信息披露和产生设计，引导信息接收者的选择，进而让信息设计者获得更大的效用。

这个例子展现了信息设计的基本框架：贝叶斯劝说模型包括两个基本博弈主体——信息发送者与信息接收者。在博弈开始前，自然地存在一个状态环境。对于这个环境的性质，发送者与接收者对自然的状态有一个先验的分布。博弈开始后，发送者产生并发送给接收者一个信号，产生和发送的方式是经过发送者设计的。信息接收者是标准理性决策人，接收者根据这个值更新自己对于自然状态的后验概率分布，然后采取适当的行动来使自己的效用最大化。

贝叶斯劝说研究的进展和应用

Kamenica 和 Gentzkow 提出的这个例子，是在社会和经济活动中常见的一个问题。它显示出贝叶斯劝说问题早就存在于现实生活中，只是未被系统性地建模分析和探讨。实际上，相关统计显示，劝说行为大概占据了美国经济行为的 1/3[4]。近年来，一系列研究在各个方面拓展了基本模型，这些拓展研究包括两个方向：一个是深入探讨信息设计的策略本身；另一个是讨论市场结构和信息设计策略的相互影响对市场均衡的塑造。

针对信息设计本身的讨论细化了信息设计者和接收者的性质，探讨了这些性质对信息设计策略和可行性的影响。例如，发送者与接收者对于自然状态的先验概率不同 [5]，或接收者有一些私人信息可以进行辅助决策 [6]。一些研究则注意到，信息设计本身并不是免费的，因此讨论了发送信息产生花费时的设计策略问题 [7]。近年来的一些研究考虑了博弈环境下的信息设计问题 [8]。这类信息设计研究讨论了信号发送者在看到一些接收者的信息后，如何改变信号的设计策略 [9]。假如场景本身会发生动态变化，例如伊利 (Ely) 等人 [10] 将基本模型扩展到了一个动态的应用场景：假设一个主体在为一个资本家工作并且持续地投入精力，他可以选择在任意一个时间点退出，如果他在退出时已经完成了任务，他就可以获得一定的效用，否则获得的效用为 0。但是，在他退出之前，他并不知道是否完成了任务，只对总任务量有一个先验的估计。而该研究提供了一个最优信号设计，指出资本家可以通过向他发送一个关于任务进度的信号来扩大自己的收益。

针对市场结构和信息设计策略相互影响的研究，主要为市场竞争对信息设计的影响。在信息设计侧，一些研究包括多个信号发送者的相互博弈如何影响最终的均衡。研究发现，不同信号发送者的利益可能有协同，也可能有冲突，这就涉及信号发送者之间是进行竞争还是进行共谋对自身的效用更加有利。针对信息接收侧，研究主要聚焦于多个信号接收者之间的博弈如何影响信号设计者的决策 [8]。

贝叶斯劝说研究不仅停留在理论分析层面，也已经被广泛地应用于经济学、金融学和政治科学等不同学科的研究中。经济学的研究聚焦于信息对市场竞争和均衡的影响上。例如文献 [11] 提出，在市场中，如果消费者的需求发生了变化，会导致需求曲线与供给

者利润曲线的变化，进而提出了一种新的广告分类方法，分析引起需求变化的是商家的炒作还是消费者的真实信息。文献 [3] 则探究了在药品研制领域，制药公司与食品药品监督管理局之间的信息博弈。制药公司可以通过改进测试阶段，提高药物实验的成功率，从而可以选择更好的药物实验设计，提高药品上市的可能性。而食品药品监督管理局可以强制要求制药公司实施更多强制性的测试来提高获批药物的质量，从而获得有关被测药物更精确的信息。

由于提供了全新的策略设计维度，贝叶斯劝说和经济学的研究被大量应用于拍卖设计中。杜格米 (Dughmi) 等人 [12] 探究了如下场景：在拍卖的时候，拍卖者不能直接展示拍卖品给买方，而是给出一定数量的信息，导致买方之间形成不完全信息博弈。该研究探索了在这种情况下，如何设计信息生成机制才能够将拍卖者的利润最大化。杜格米等人 [9] 考虑了在拍卖情况下，中介方应该如何把买家的情况部分地传递给拍卖者来影响拍卖者决策。

贝叶斯劝说在政治科学上的应用，主要体现在分析对选民的劝说问题上。例如，阿朗索 (Alonso) 等人 [13] 讨论了政治家可以向选民发送一个有关其提案的信号。作者表明，在多数决规则下，这个政治家可以通过巧妙地设计信号，使得一个原本会被否决的提案在信号的影响下通过，并且大多数选民的效用相比没有信号的情形不增。Xu 等人 [14] 在国际安全领域对贝叶斯劝说与斯塔克尔伯格博弈 (Stackelberg Game) 进行了结合探讨。在一些特殊情况下，防御者虽然可以成功抵御攻击，但是仍然要承受很大的损失。此时，如果防御者策略性地传递一些受攻击部位的信息给攻击者，起到一定的震慑作用，那么防御者就能够得到更高的效用。

贝叶斯劝说揭示了一些深刻的反事实结论。例如在针对金融领域风险披露的研究中，戈尔斯坦 (Goldstein) 等人证明了“信息公开”有可能是个坏政策，并探究了金融领域风险披露行为 [2]。2008 年金融危机之后，全球中央银行对金融机构进行定期压力测试，以评估其抵御未来冲击的能力。对于决策者而言，在某些程度下必须采取一定程度的披露行为防止风险发生。而在某些程度下，选择不披露风险测试结果更有利于社会福利的最优。

贝叶斯劝说中的计算问题

贝叶斯劝说的设计和分析研究是植根于其计算性质的。贝叶斯劝说面对的最本质的信号设计问题就是当面对不同的场景时，信号发送者需要设计怎样的信号产生机制，使自己的效用最大化。首先，信息设计是一个非凸优化问题。从 Kamenica 和 Gentzkow 给出的基本例子中就可以发现：对信息发送方来说，信息并不是披露得越多越好，也不是越少越好。这种信息披露量和效用之间的非单调关系，决定了信号产生机制的设计是一个非凸问题。正如 Dughmi 等人所总结的，对于一个信号发送者和一个信号接收者的模型，信号的本质可以看作将一个状态 θ 映射到接收者的活动的一个分布 [15]。$\varphi(\theta, a)$ 代表状态 θ 下，发送者劝说接收者采取活动 a 的概率。那么优化问题可以采用下面的线性规划问题 [15] 表示：

$$
\begin{aligned}
\text{maximize} \quad & \textstyle\sum_{\theta\in\Theta} \mu(\theta) \sum_{a\in A} \varphi(\theta,a) s(\theta,a) \\
\text{subject to} \quad & \textstyle\sum_{\theta} \mu(\theta)\varphi(\theta,a)(r(\theta,a) - r(\theta,a')) \geq 0, && \text{for } a, a' \in A. \\
& \textstyle\sum_{a\in A} \varphi(\theta,a) = 1, && \text{for } \theta \in \Theta. \\
& \varphi(\theta,a) \geq 0, && \text{for } \theta \in \Theta, a \in A.
\end{aligned}
$$

在上述式子中，Θ 是自然状态，$\mu(\theta)$ 是自然状态的先验分布，a 是信息接收者采取的行动。而 $r(\theta, a)$, $s(\theta, a)$ 则是指当自然状态为 θ，信息接收者行动为 a 时，信息接收者和信息设计者的效用。$\varphi(\theta, a)$ 指当自然状态为 θ 时，信息设计者推荐接收者采取动作 a 的概率。信息设计者需要在期望意义下使自己的效用最大化。公式中三个约束条件分别代表：当发送者推荐 a 动作时，接收者会听取建议；当自然状态固定时，推荐各个动作的概率和为 1；每个状态下，推荐每个动作的概率都是非负值。

如果信号发送者的效用函数在概率空间上是凸函数或者凹函数，上述问题就很好求解。在凸函数情况下，最大值在凸包边缘处取到，亦即信息设计者的最优策略是传递所有信息。而在凹函数情况下，最大值在内部取到，此时信息设计者的最优策略是

不传递任何信息。但是，正如 Kamenica 和 Gentzkow 在给出的基本例子中显示的，信息设计者的效用函数可以是非凸的，那么这个问题就很难求解。Kamenica 和 Gentzkow 提出了一种经典的凸化方法：对“信息设计者效用与后验分布均值的关系曲线”取凸包，这个凸包在先验均值处的取值就是生成信号能够得到的最大效用 [1]。

一些研究在此基础上进行了拓展性讨论。这些研究发现，如果接收者一共有 n 个行为，每个行为可能会导致 m 个结果与效用。假设同一个信号导致不同动作之间的效用是独立同分布的，那么我们可以得到关于 m 和 n 的多项式时间算法来求得信号设计的最优解。而当行为之间独立，但是不同分布的时候，不存在多项式时间算法能够求得最优解 [16]。

然而，当发送者或接收者众多时，信息设计策略的求取问题具有较高的计算复杂度。即便对于只有两个接收者进行零和博弈的情况，该问题是 NP 难 (NP-hard) 的，也不存在对应的多项式时间的近似算法 [17]。所以，一些研究聚焦于如何设计算法求解 [15]。例如 Cheng 等人 [18] 指出，当信息设计者的效用函数满足李普希兹 (Lipschitz) 平滑与噪声稳定的一些条件，而且信息是通过公开的方式传递给所有的接收者时，那么就存在相应的多项式时间近似算法求解最优信息设计策略。而如果信息发送者可以给不同的接收者发送不同的信息，那么信息设计者需要针对每一个信息发送者设计定制化的信息——用于保证接收者之间的不完全信息博弈能够导出对信息设计者有利的均衡 [19]。目前的研究表明，对于行为空间与自然状态空间都是二元的情况，当信息发送者的效用函数是超模函数时，存在多项式时间的计算算法；当信息发送者的效用函数是次模函数时，存在一个多项式时间的近似算法 [20]。

总结

贝叶斯劝说无论在经济领域还是在计算领域，都有着巨大的研究价值与应用价值，不仅能够促使我们更好地理解社会与设计信息交流规则，同时也有助于我们对经济活动中的信息交流进行更加深入的监管、定价等研究。随着大量学者研究的不断深入，贝叶斯劝说必将进入更广阔的领域，带来更高的经济与学术价值，有效推动相关领域的发展与改革。

目前对于贝叶斯劝说的理论探究还处于比较早期的阶段，大量研究仍然聚焦于基本理论框架的完善。而计算与应用研究则处在方兴未艾的阶段。本文围绕贝叶斯劝说的基本模型及其拓展进行了讨论，并解释了贝叶斯劝说所涉及的基本计算问题，希望能给相关领域的科研人员带来新的灵感与思路。 ■

汪至祺
清华大学交叉信息研究院硕士研究生。主要研究方向为计算经济学。
wangzhiq19@mails.tsinghua.edu.cn

张耀予
清华大学交叉信息研究院博士研究生。主要研究方向为计算经济学。
yaoyu-zh19@mails.tsinghua.edu.cn

邝仲弘
清华大学经济管理学院博士研究生。主要研究方向为信息设计和竞赛理论。
kuangzhh1994@sina.com

于　洋
CCF 专业会员。清华大学交叉信息研究院助理教授。主要研究方向为计算经济学、网络经济学和产业组织理论等。
yangyu1@tsinghua.edu.cn

参考文献

[1] Kamenica E, Gentzkow M. Bayesian persuasion [J]. *American Economic Review*, 2011,101(6): 2590-2615.

[2]Goldstein I, and Leitner Y. Stress tests and information disclosure [J]. *Journal of Economic Theory* 177, 2018: 34-69.

[3]Johnson J, Myatt D. On the simple economics of advertising, marketing, and product design [J]. American Economic Review, 2006, 96(3): 756-784.

[4]Antioch G. Persuasion is now 30 per cent of US GDP: Revisiting McCloskey and Klamer after a quarter of a century [J]. Economic Round-up 1, 2013(1).

[5]Alonso R, Camara O. Bayesian persuasion with heterogeneous priors [J]. Journal of Economic Theory, 2016: 672-706.

[6]Guo Y, Shmaya E. The interval structure of optimal disclosure [J]. Econometrica, 2019, 87(2): 653-675.

[7]Gentzkow M, Kamenica E. Costly persuasion [J]. American Economic Review, 2014, 104(5): 457-62.

[8]Dughmi S. On the hardness of signaling [C]// 2014 IEEE 55th Annual Symposium on Foundations of Computer Science. IEEE, 2014.

[9]Dughmi S, Kempe D, Qiang R. Persuasion with limited communication [C]// Proceedings of the 2016 ACM Conference on Economics and Computation. 2016: 663-680.

[10]Ely J, Szydlowski M. Moving the goalposts [J]. Journal of Political Economy, 2020,128(2).

[11]Kolotilin A. Experimental design to persuade [J]. Games and Economic Behavior 90, 2015: 215-226.

[12]Dughmi S, Immorlica N, Roth A. Constrained signaling in auction design [C]// Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics, 2014: 1341-1357.

[13]Alonso R, and Câmara O. Persuading voters [J]. American Economic Review, 106(11), 2016: 3590-3605.

[14]Xu H, et al. Exploring information asymmetry in two-stage security games [C]// Twenty-Ninth AAAI Conference on Artificial Intelligence. 2015.

[15]Dughmi S. Algorithmic information structure design: a survey [J]. ACM SIGecom Exchanges, 15(2), 2017: 2-24.

[16]Dughmi S, and Xu H. Algorithmic bayesian persuasion [J]. SIAM Journal on Computing 0, 2019: STOC16-68.

[17]Bhaskar U, et al. Hardness results for signaling in bayesian zero-sum and network routing games [C]// Proceedings of the 2016 ACM Conference on Economics and Computation., 2016: 479-496.

[18]Cheng Y, et al. Mixture selection, mechanism design, and signaling [C]// 2015 IEEE 56th Annual Symposium on Foundations of Computer Science. IEEE, 2015: 1426-1445.

[19]Arieli I and Babichenko Y. Private bayesian persuasion [J]. Journal of Economic Theory 182, 2019: 185-217.

[20]Babichenko Y and Barman, S. 2016 [J]. Computational aspects of private bayesian persuasion. arXiv preprint arXiv:1603.01444.