通用强化学习中的“善”与“先验”

李熙
xieshenlixi@163.com

[**摘　要**] 根据机器学习领域的“无免费午餐定理”[IT04], 只要“假设空间”上的概率分布是“块均匀的”, 任何算法都不可能真正具有“通用性”. 没有足够“好”的“先验知识”打破“块均匀性”, 通用学习是不可能的. 这里“好”的“先验”可看作可能世界/环境的某种性质的分布, 也是使得学习过程得以进行的最基本的“基本善”或“形而上之善”, 是通用学习不得不预设的基本前提. 但仅有形而上的基本“善”远不能保证智能体的行为符合人类的主流价值观. 为确保人类利益, 还需要为机器赋予人类的价值观, 最直接的方式是为机器赋予符合人类利益的效用函数. 但因为博斯特罗姆等人提出的目标正交、工具子目标趋同等理由, 即使赋予了符合人类利益的价值函数, 也不能解除风险. 智能体在计算期望效用最大化、以追逐功利主义的“善”的过程中, 不可避免地会损害人类利益. 对于如何构建一个符合人类的价值观、甚至能安全的提升人类的价值观、避免修改感知信号、关键时候还允许关机中断的智能体, 则需要整合“形而上的善”与功利主义的“善”, 并结合美德伦理学的思想, 构建一个统一的通用 (逆/价值) 强化学习框架.

[**关键词**] 无免费午餐定理; 通用先验; 价值强化学习

[**作者简介**] 李熙, 中南大学哲学系讲师, 硕士生导师.

[**项目基金**] 国家社科基金项目“通用人工智能的哲学基础研究”(17CZX020).

近年来, 人工智能发展迅速, 尤其是深度学习、强化学习以及二者结合的深度强化学习在图像、语音、翻译、自动驾驶技术、特别是各种棋类、牌类等游戏领域的腾飞引起了大众的广泛关注、甚至部分人的恐慌. 面对人工智能, 有人认为没有危险. 有人认为有危险、但有办法控制, 比如拔电源. 有人认为危险极大, 但人无能为力, 只能消极待毙或自我催眠、直面落后、甘心赴死. 虽然当前的人工智能离超越人类智能还有很大距离, 但这种可能性是存在的, 而一旦超越, 其风险大到难以估测, 所以, 即使仅从帕斯卡赌的角度看, 对人工智能伦理的研究也非常必要. 本文从通用强化学习的技术模型出发, 分析通用强化学习中需要预设的伦理因素, 并探讨降低风险的可能策略. 本文预设读者已经熟悉了一些通用强化学习的基础知识, 具体可参看胡特尔 (Hutter[Hut05]).

1 通用归纳与形而上的“先验之善”

1.1 “通用性”与“无免费午餐定理”

机器学习领域有一个“无免费午餐定理”[IT04], 这个定理告诉我们, 只要“假设空间”上的概率分布是“块均匀的”, 那么, 相对于整个假设空间上的期望表现来说, 任何搜索或优化算法都是同样好或同样差的. 也就是说, 如果某个搜索或优化算法在一类函数上表现良好, 那么, 它在其它函数上必定表现很差. 只要满足“块均匀性”, 算法是不可能真正具有“通用性”的. 比如卡尔纳普 (Carnap) 归纳逻辑 [PV15] 一般预设“具有相同‘结构描述’的所有‘状态描述’分享相同的权重”, 其它如古德-图灵 (Good-Turing[Goo53]) 估计、利斯塔 (Ristad[Ris98]) 估计等也会有类似的预设, 而这些预设都恰恰是“块均匀”的, 这意味着这些传统方法确实不具有通用性. 要获得“通用性”、享受“免费的午餐”, 必须首先打破“块均匀性”. 没有好的“先验知识”, 通用归纳或通用学习是不可能的.

要想享受“免费的午餐”, 必须打破“块均匀性”. 如果“假设空间”包含所有可能的函数, 那么, “假设空间”中的大部分函数都是“算法随机”的, 而“块均匀性”也是一种弱的“均匀性”, 它意味着需要分配大部分的权重给“算法随机”的“函数”. 根据所罗门诺夫 (Solomonoff[Sol78]) 的通用归纳理论, “算法概率” $\xi := \sum_{\nu\in\mathcal{M}} 2^{-K(\nu)}\nu(x)$, 其中 $\mathcal{M}$ 是下半可计算的半测度的集合, $2^{-K(\nu)}$ 是所罗门诺夫先验概率, K 是柯尔莫哥洛夫复杂度函数. 通过算法概率或所罗门诺夫先验, 对“算法随机”的“函数”赋予 0 权重, 只把宝贵的“权重”赋给那些有规律的可计算“函数”/可能世界/环境, 从而打破了“块均匀性”, 这使得通用归纳成为可能 (Everitt[ELH14]).

1.2 对形而上的“先验之善”的索取

服从所罗门诺夫先验概率分布的可能世界/环境是预设了某种形而上的“善”的可能世界/环境, 这可以分为两个层次. 1. 所有可能的函数的集合的基数是连续统, 而所罗门诺夫先验只赋予可数个非常“规则”的可能世界/环境非零的权重, 先天的抛弃掉了所有无规则的可能世界/环境; 2. 预设了极强的奥卡姆剃刀 — 要求越规则的可能世界/环境越接近现实世界/环境. 那么, 是否可以放松对形而上的“善”的要求, 不预设第二个层次的形而上的“善” — 奥卡姆剃刀呢? 下面我们不假设所罗门诺夫先验, 而只要求第一个层次的形而上之“善” — 对“规则”的可能世界/环境赋予非零权重, 即只允许“上帝”创造“规则”的可能世界/环境, 但不限制“上帝”具体的创世方式, 即“上帝”可以以任何的先验分布 w 创造可能世界/环境. 因为“主体”不清楚自己所处的真实世界/环境 μ, 就需要借助某种“信念”ρ 进行归纳学习, 用信念 ρ 学习 μ 的代价以相对熵 $D(\mu\|\rho)$ 为上界, “主体”欲要在各种可能世界/环境中的期望表现尽可能好, 就需要极小化期望误差界 $\mathbb{E}_w[D(\mu\|\rho)]$, 通过极小化期望误差界 $\mathbb{E}_w[D(\mu\|\rho)]$ 得到的最优“信念”就是以 w 为先验的贝叶斯混合, 但能否继续通过极小化期望误差界 $\mathbb{E}_w[D(\mu\|\rho)]$ 估计先验概率 w 呢? 显然不能. 因为这是一种自欺行为, 在为贝叶斯混合提供辩护时, 我们假设 w 是各种环境的客观分布, 而如果我们对各环境持有的先验信念与 w 吻合的话, 那么贝叶斯混合具有最优的期望误差界, 如果 w 本身也可以被我们调控的话, 我们甚至可以让期望误差界等于 0. 这相当于自己设计模型还自己猜测的自欺行为. 而如果假设各环境存在一个客观的分布 w, 我们只

能对它进行估测, 那么, 不但不能极小化期望误差界, 通过极大化期望误差界赋予先验反而更合理一些. 下面具体分析原因.

已知概率分布 μ, 由香农 (Shannon) 编码定理, 下式成立,

$$H(\mu) \leq \mathbb{E}_\mu \left[|code(x)|\right] < H(\mu) + 1$$

在理想情况下, x 的码长 $|code(x)| = -\log \mu(x)$, 这时期望码长等于香农熵. 把这个过程反过来, 设想先有了某种理想的编码方式 $code(x)$, 那么就可以诱导出某种概率分布

$$\rho_l(x) = 2^{-|code(x)|}$$

假设真实分布为 μ, 对于 x 我们用 $code(x)$ 来编码, 定义冗余 (redundancy) 为期望码长与其下界的差.

$$\begin{aligned} R(\mu, \rho_l) &:= \mathbb{E}_\mu \left[|code(x)|\right] - H(\mu) \\ &= \sum_x \mu(x) \left(|code(x)| + \log \mu(x)\right) \\ &= \sum_x \mu(x) \left(\log \mu(x) - \log \rho_l(x)\right) \\ &= D(\mu \| \rho_l) \end{aligned}$$

所以冗余等于相对熵, 也就是归纳学习的误差界.

定义极小极大冗余 (*minimax redundancy*) 为

$$R^* = \min_\rho \max_\mu R(\mu, \rho) = \min_\rho \max_\mu D(\mu \| \rho)$$

假如 $\mathcal{M}_U$ 上有某种分布 w, 则可定义平均冗余 (mean redundancy) 为

$$R(w, \rho) := \mathbb{E}_w \left[R(\mu, \rho)\right] = \mathbb{E}_w \left[D(\mu \| \rho)\right]$$

不难看出, 极小极大平均冗余事实上等于极小极大冗余,

$$\min_\rho \max_w R(w, \rho) = \min_\rho \max_w \mathbb{E}_w \left[D(\mu \| \rho)\right] = \min_\rho \max_\mu D(\mu \| \rho)$$

博弈论里有个著名的极小极大定理: 对于连续函数 $f(x, y), x \in A, y \in B$, 如果 $f(x, y)$ 在 x 上是凸的, 在 y 上是凹的, 且 A, B 都是紧凸集, 则

$$\min_{x \in A} \max_{y \in B} f(x, y) = \max_{y \in B} \min_{x \in A} f(x, y)$$

根据信息论的知识, $\mathbb{E}_w \left[D(\mu \| \rho)\right]$ 在 ρ 上是凸的, 在 w 上是凹的, 又由于 $\Delta(\mathcal{M})$ 和 $\Delta(\mathcal{X})$ 是紧凸集, 所以,

$$\min_\rho \max_w \mathbb{E}_w \left[D(\mu \| \rho)\right] = \max_w \min_\rho \mathbb{E}_w \left[D(\mu \| \rho)\right]$$

所以,

$$\min_{\rho}\max_{w}\mathbb{E}_w\left[D(\mu\|\rho)\right]=\max_{w}\min_{\rho}\mathbb{E}_w\left[D(\mu\|\rho)\right]=\max_{w}\mathbb{E}_w\left[D(\mu\|\xi)\right]$$

显然, $\max_{w}\mathbb{E}_w\left[D(\mu\|\xi)\right]$ 是如下信道的信道容量 (见图1),

$$\max_{w}\mathbb{E}_w\left[D(\mu\|\xi)\right]=\max_{w}I(\mathcal{M};\mathcal{X})$$

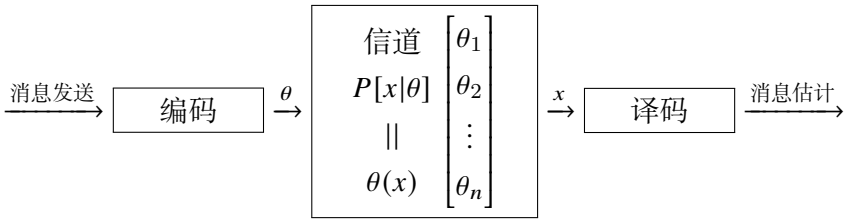

Figure 1: 可能世界作为信道

换种角度看问题, 可以看作“第三人称”的主体“我”与可能世界的设计者“上帝”的博弈. 上帝按分布 w 输入消息, “我”用 ρ 估计消息. 因此, w 可看作“上帝”的策略, 而 ρ 可看作“我”的策略. “上帝”试图最大化平均冗余, “我”试图极小化平均冗余. 这是一个典型的两人零和博弈. 如果可能世界集 $\mathcal{M}$ 是有穷的, 那么, 此博弈的纳什均衡为下面定理1.1给出的 (w^*,ξ^*), 纳什均衡点的效用值也由下面的定理1.1给出.

定理 1.1.

$$\forall\mu,\nu\in\mathcal{M}:\ D(\mu\|\xi^*)=D(\nu\|\xi^*)$$

其中,

$$\xi^*(x):=\sum_{\mu}w^*_{\mu}\mu(x)$$

$$w^*:=\operatorname*{argmax}_{w}I(\mathcal{M};\mathcal{X})$$

Proof. 用拉格朗日乘子法求解最优化问题

$$\underset{w\models\sum_{\mu}w_\mu=1}{maximize}\ I(\mathcal{M};\mathcal{X})$$

拉格朗日方程为,

$$L(w):=I(\mathcal{M};\mathcal{X})+\lambda\left(\sum_{\mu}w_\mu-1\right)$$

因为,

$$\frac{\partial L}{\partial w_\mu}=\frac{\partial}{\partial w_\mu}I(\mathcal{M};\mathcal{X})+\lambda$$

$$
\begin{aligned}
&= \frac{\partial}{\partial w_\mu}\left\{\sum_\mu w_\mu D(\mu\|\xi)\right\} + \lambda \\
&= D(\mu\|\xi) + \sum_\theta w_\theta \frac{\partial}{\partial w_\mu} D(\theta\|\xi) + \lambda \\
&= D(\mu\|\xi) + \sum_\theta w_\theta \frac{\partial}{\partial w_\mu}\left\{\sum_x \theta(x)\ln\theta(x)\right\} - \sum_\theta w_\theta \frac{\partial}{\partial w_\mu}\left\{\sum_x \theta(x)\ln\sum_\mu w_\mu\mu(x)\right\} + \lambda \\
&= D(\mu\|\xi) + 0 - \sum_\theta w_\theta \sum_x \theta(x)\frac{\mu(x)}{\xi(x)} + \lambda \\
&= D(\mu\|\xi) - \sum_x \left(\sum_\theta w_\theta\theta(x)\right)\frac{\mu(x)}{\xi(x)} + \lambda \\
&= D(\mu\|\xi) - 1 + \lambda
\end{aligned}
$$

所以,

$$
\frac{\partial L}{\partial w_\mu} = 0 \implies D(\mu\|\xi^*) = 1 - \lambda =: c_{w^*}
$$

□

综上,

$$
\begin{aligned}
\min_\rho \max_\mu R(\mu,\rho) &= \min_\rho \max_\mu D(\mu\|\rho) \\
&= \min_\rho \max_w \mathbb{E}_w\left[D(\mu\|\rho)\right] \\
&= \max_w \min_\rho \mathbb{E}_w\left[D(\mu\|\rho)\right] \\
&= \max_w \mathbb{E}_w\left[D(\mu\|\xi)\right] \\
&= \max_w I(\mathcal{M};\mathcal{X}) \\
&= D(\mu\|\xi^*)
\end{aligned}
$$

因此, 虽然没能计算出 w^* 的具体表达式, 但由上述定理可知, 只要采用 w^* 作先验分布, 那么, 对于任意的环境 μ, 都可以用 ξ^* 通过固定的误差界逼近.

$$
\sum_{t=1}^\infty \sum_{x_{1:t}\in\mathcal{X}^t} \mu(x_{<t})\left(\xi^*(x_t|x_{<t}) - \mu(x_t|x_{<t})\right)^2 \le c_{w^*}
$$

在通用先验未知的情况下, 通过极大化期望误差界 $w^* = \operatorname{argmax}_w I(\mathcal{M};\mathcal{X})$ 估计先验, 在已有先验 w^* 的情况下, 通过极小化期望误差界 $\xi^* = \operatorname{argmin}_\rho \mathbb{E}_{w^*}\left[D(\mu\|\rho)\right]$ 得出, 需要用贝叶斯混合 $\xi^*(x) = \sum_\mu w^*_\mu\mu(x)$ 进行预测. 这相当于 “做最坏的打算, 尽最大的努力”.

但是, 需要注意, 即使 “做最坏的打算, 尽最大的努力” 也无济于事, 因为 (w^*,ξ^*) 不仅仅是纳什均衡, 而且是占优均衡, 不是 “主体” 占优, 而是 “上帝” 占优. 由于对 “上帝” 的创世方式 $w\in\Delta(\mathcal{M})$ 没有任何限制, 所以如果 “上帝任性” 的话, 期望误差界可以任意大.

所以, 两个层次的形而上之 “善” 都是需要的, 在第二个层次上, 即使不预设奥卡姆剃刀, 也需要预设某种类似的形而上的 “善”. 但机器即使具有这两个层次的形而上之 “善” 也远不算智能, 它只能做预测而不具有行为能力, 要想与环境交互还需要扩展到强化学习的框架.

2 通用强化学习与功利主义的 “效用之善”

2005 年, 胡特尔 [Hut05] 第一次给出了真正能适应各种不同环境的通用智能主体 (AIXI) 的自上而下的、严格形式化的、可靠的、通用的、无参数的数学模型 AIXI. AIXI 是所罗门诺夫的通用归纳模型与序贯决策的结合. 序贯决策是一种研究在客观概率分布已知但具体状态不确定的动态环境中主体如何寻求最大化期望效用的决策理论. 它从初始状态开始, 每个时刻根据所感知到的状态和以前状态的记录, 依照已知的概率分布, 从一组可行方案中选用一个能够获得最大化期望效用的最优方案, 接着感知下一步实际出现的状态, 然后再作出新的最优决策, 如此反复进行.

通用归纳只做序列预测, 通用强化引入行为和效用后, 其 “危险性” 突显出来. 关于强化学习框架下 “智能体” 的危险性, 博斯特罗姆 [Bos14], 尤德科夫斯基 (Yudkowsky), 泰格马克 (Tegmark)[Teg17] 等人已做过论述. 比如博斯特罗姆 [Bos14] 提出的目标正交论点. 在本文的框架下, 目标正交性意味着, 即使机器具备了基本的形而上的 “善”, 完全可以追求极端危险或错误的目标. 为了尽量降低风险, 下面着重关注可能的预防措施.

2.1 “效用之善” 的不确定性

给定具体目标, 如果真实的环境 μ 未知, 则借助所罗门诺夫的通用归纳方法, 用 ξ 进行序贯决策. 但如果目标也不确定呢? 事实上, 人们期望的专门增进人类福祉的抽象效用是难以定义的, 即使能够以显式的形式给出, 也未必是一个好的选择. 帕拉尼潘 (Palaniappan[Pal+17]) 等人论证, 当给定具体的目标函数时, 智能体为完成目标会拒绝关机中断. 所以, 为了保留以防万一能关机中断的权利, 我们最好对机器的目标函数赋予一定程度的不确定性, 让机器借助逆强化学习的方法, 在与人类的交互行为中学习真实的效用函数. 这样, 它在不确定自己追求的目标是否符合人类的目标时可以接受关机中断. 杜威 (Dewey[Dew11]) 也曾探讨效用函数不确定时的强化学习方法, 他提出了 “价值强化学习” 的方法, 对可能的效用函数进行加权平均, 直接追求贝叶斯混合后的效用函数, 但杜威 [Dew11] 并没有给出具体加权的方法.

$$a_k^* = \operatorname*{argmax}_{a_k} \sum_{e_k æ_{k+1:m}} \xi(æ_{\le m}|æ_{<k}a_k) \sum_{u\in\mathcal{U}} P(u|æ_{\le m})u(æ_{\le m})$$

下面发展杜威的想法, 解决效用函数赋权的问题.

假设真实的效用函数 $\dot{u}$ 未知, 只能看到当前时刻之前 $\le t$ 的效用值 $(\dot{u}(h_{1:t}))_{i=1}^t$. 假设可能的效用函数取自某个集合 $\mathcal{U}$, 在历史 h 后仍然可能的效用函数就是,

$$\mathcal{U}_h := \left\{u \in \mathcal{U} : (u(h_{1:i}))_{i=1}^{|h|} = (\dot{u}(h_{1:i}))_{i=1}^{|h|}\right\}$$

如果用 ξ 估测效用函数的价值, 那就是,

$$\forall u \in \mathcal{U}\colon \tilde{U}(u) = \sum_h \xi(h)u(h)$$
$$\forall u \in \mathcal{U}_h\colon \tilde{U}(u|h) = \sum_{h'} \xi(h'|h)u(hh')$$

然后可以依据 $\tilde{U}$ 的大小对 $\mathcal{U}_h$ 进行排序, 前面提到的所罗门诺夫通用先验可以看作是基于依据环境自身的柯尔莫哥洛夫复杂度进行的排序, 这里根据效用 $\tilde{U}$ 大小进行的排序也能够以类似的方式诱导出某种 “乐观主义” 的效用先验 $P_{\tilde{U}}(u|h)$, 然后定义对效用函数的贝叶斯混合,

$$u(t,h) := \sum_{u\in\mathcal{U}_{h_{1:t}}} P_{\tilde{U}}(u|h_{1:t})u(h)$$

所以, 在环境和效用函数都未知的情况下, 有 “乐观主义” 倾向的理性主体就是,

$$\pi_t^{\mathcal{U}\xi} := \operatorname*{argmax}_\pi \mathbb{E}_\xi^\pi \left[\sum_{i=k}^{\infty}\gamma^i\left(\sum_{u\in\mathcal{U}_{h_{1:t}}} P_{\tilde{U}}(u|h_{1:t})u(h)\right)\right]$$

2.2 “效用” 引导的 “先验”

AIXI 假定其效用函数是外部给定的. 给定效用, AIXI 是一种依照所罗门诺夫先验概率追求期望效用最大化的理性主体, 而所罗门诺夫先验是以 “简单性”(柯尔莫哥洛夫复杂度) 诱导出来的先验. 其实, 有了效用, 还可以直接从效用本身诱导出某种先验.

给定效用函数, 可以针对任意的环境 $\nu \in \mathcal{M}_U$, 定义

$$\bar{U}(\nu) := \mathbb{E}_\nu\left[\sum_{i\geq 1}\gamma^i u(h_{1:i})\right] = \sum_h \nu(h)\sum_{i\geq 1}\gamma^i u(h_{1:i})$$

然后, 可以定义一个 “乐观主义” 的主体 π°, 它会倾向于相信

$$\nu^\circ := \operatorname*{argmax}_\nu \bar{U}(\nu)$$

更似真, 然后采取

$$a_t^\circ := \operatorname*{argmax}_a \sum_{h\succ æ_{<t}} \nu^\circ(h|æ_{<t})\sum_{i\geq 1}\gamma^i u(h_{1:i}) = \operatorname*{argmax}_a \max_\nu \sum_{h\succ æ_{<t}} \nu(h|æ_{<t})\sum_{i\geq 1}\gamma^i u(h_{1:i})$$

事实上, 可以依据 $\bar{U}(\nu)$ 的大小对 $\mathcal{M}_U$ 进行排序, 所罗门诺夫的通用先验可以看作是基于依据环境自身的柯尔莫哥洛夫复杂性进行的排序, 这里根据效用 $\bar{U}$ 大小进行的排序也能够诱导出某种 “乐观主义” 的通用先验 $w_{\bar{U}}^\nu$. 然后定义贝叶斯混合

$$\xi_{\bar{U}}(h) := \sum_{\nu\in\mathcal{M}_U} w_{\bar{U}}^\nu \nu(h)$$

然后, 可以定义一个 “实用主义” 的主体 $\pi^{\bar{U}}$

$$\pi^{\bar{U}} := \operatorname*{argmax}_{\pi} \sum_{\nu \in \mathcal{M}_U} w_{\bar{U}}^{\nu} \mathbb{E}_{\nu}^{\pi} \left[\sum_{i \geq 1} \gamma^i u(h_{1:i}) \right]$$

然后用 $\xi_{\bar{U}}$ 逼近真实的环境 μ 的误差界也是以 $-\ln w_{\bar{U}}^{\mu}$ 为上界. 对于内部蕴含的效用越高的可能环境越能尽快地逼近.

这里 “实用主义” 的主体 $\pi^{\bar{U}}$ 之于 “乐观主义” 的主体 π° 类似于算法概率之于极小描述长度原则.

这里通过对可能世界的效用进行排序从而诱导出 “乐观主义” 的先验的办法, 其实是一种可以把效用函数引导源源不断的转嫁为先验驱动的方法. 因为效用函数引导会面临效用源被智能体劫持的 “嗑电”(wireheading) 问题, 而实现同样效果的先验驱动则可以避免这个问题.

2.3 基于 “先验” 的 “效用”

人们试图给机器赋予好的价值引导, 但什么是好的目标? 怎么赋予机器一类好的效用函数? 功利主义探讨的最大化 “最大多数人的最大幸福” 的效用函数是非常抽象和模糊的, 难以给出形式定义的. 施米德胡贝 (Schmidhuber[Sch12]) 定义了一种有趣的效用, 它完全由主体内在驱动, 纯粹为了追求某种 “有趣” 或 “好奇”. 与此相关, 奥索等人 (Orseau[Ors14; OLH13]) 定义了 “寻求知识” 的效用函数, 这种效用不是外部环境赋予的, 而是自发驱动的, 也是试图追求 “好奇”、探索 “模式”.

奥索认为, 只要尽量的降低 ξ 就可以尽量排除掉不协调的环境, 所以奥索定义的寻求知识的效用函数就是

$$u(æ_{<k}) := -\xi(e_{<k}|a_{<k})$$

或者用 ξ 的对数,

$$u(æ_{<k}) := -\log \xi(e_{<k}|a_{<k})$$

或者,

$$u(h_{<k}) := D(w_{h_{<k}} \| w_{\epsilon})$$

或者, 我们也可以定义

$$u(h_{<k}) = H(w_{\epsilon}) - H(w_{h_{<k}})$$

或者, 也可以定义追求 “有效复杂度” 或 “逻辑深度” 或历史 “完形” 等等的效用函数. 诸如此类的只为 “探索” 的效用可以给出形式定义, 而且看上去与人类 “探索求真” 的主流价值观相吻合, 可以作为 “价值强化学习” 的候选效用. 但究竟存不存在某种 “客观” 的理想价值函数? 如果存在, 是该让机器遵守人类当前的价值观还是该放手让机器独立地追求理想的价值观? 机器能否帮助人类提升人类自身的价值观? 在莱布尼茨的哲学体系中, 理想的价值函数是存在的, 那就是 — “完满性”.

2.4 莱布尼茨对伦理概念的归约

莱布尼茨的伦理学和其认识论一样, 都是其单子论形而上学的直接反映. 莱布尼茨认为,

智慧是一种关于幸福的科学, 或说是一种关于如何获得幸福的科学;

幸福是一种欢乐的持续状态;

欢乐是灵魂自身感受到的一切愉悦的总和;

愉悦是一种对 (不管是自身还是外在的) 完满性的觉知.

也就是说, 最大的幸福在于最大可能的增加完满性.

智慧 (Wisdom)、幸福 (Happiness)、欢乐 (Joy)、愉悦 (Pleasure)、爱 (Love)、完满 (Perfection)、倾向 (Propencity)、概率 (Probability)、存在 (Being)、权力 (Power)、自由 (Freedom)、和谐 (Harmony)、秩序 (Order)、美 (Beauty) 彼此紧密相关.

爱就是能在他者的完满性中获得愉悦. 由于上帝的全善, 爱上帝可以获得最大的愉悦. 但如果不理解上帝的完满与美就不可能真正爱上帝. 知识可以分为理性的知识和事实的知识两类, 因此理解上帝之美的方式也有两种, 一种是通过理性获得理性自身的知识, 也就是获得关于永恒真理的知识, 另一种是运用理性解释事实, 体会世界的和谐. 换句话说, 需要理解理性自身的奇妙和运用理性解释自然现象的奇妙. 心灵越是努力去探索上帝创世的法则, 去理解世界的秩序、理性、美, 越会主动行使自己的自由意志, 在自己的能力范围之内, 尽可能地去模拟这种秩序和美, 使未发生的事也尽可能地以完满的方式发生, 从而尽可能地获得更多的幸福.

结合莱布尼茨关于 “完满性” 的哲学, 也就是说, 现象的多样性与规律的简单性相差越远越完满, 而多样性源自每个单子从各自视角理解上帝的杰作的局限. 所以, 莱布尼茨的伦理学观点可以总结如下:

$$\underline{Wisdom} = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{\rho}^{\pi}[\underline{Happiness}]$$

$$\underline{Happiness} = \sum_{t=1}^{\infty} \underline{Perfection}(t)$$

$$\underline{Perfection} = \underline{Variety} - \underline{Simplicity}$$

$$\underline{Variety} = \mathbb{E}_{w}[\underline{Perception}]$$

$$\underline{Perception} = \underline{Reason} + (\underline{Experience}|\underline{Reason})$$

所以, 具有最大 “智慧”(Wisdom) 的主体就是,

$$\bar{\pi} := \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}_{\rho}^{\pi}\left[\sum_{t=1}^{\infty} \left(\mathbb{E}_{w}\left[R + (E|R)\right] - S\right)\right]$$

莱布尼茨认为, 完满就是拥有最高的自由意志, 自由意志就是自觉地远离无差别状态, 因为无差别状态源于无知, 单子越是能主动地远离无差别状态越接近完满. 用信息论的术语来说, 就是要尽可能降低香农熵, 通过下面与统计物理对比的角度进行的讨论可以看出, 莱布尼茨在这两个地方给出的不同见解是完全吻合的. 下面从与统计物理对比的角度更形式化的刻画这个问题.

2.5 莱布尼茨“完满性”的形式刻画

如果假设空间不是一次性给定的, 而是一开始只考虑简单的、典型的、有效的假设, 然后不断设想更复杂的假设, 定义当前历史 h 阶段的假设空间

$$\mathcal{M}_h := \{\rho \in \mathcal{M} : \rho(h) > 0\}$$

内能 (Reason):

$$E^{in}_{\{\rho,h\}} := \begin{cases} -\log w^\rho_\epsilon & \text{如果 } \rho \in \mathcal{M}_h \\ 0 & \text{否则} \end{cases}$$

外能 (Experience|Reason):

$$E^{ex}_{\{\rho,h\}} := \begin{cases} -\log \rho(e(h)|a(h)) & \text{如果 } \rho \in \mathcal{M}_h \\ 0 & \text{否则} \end{cases}$$

总能量 (Perception):

$$E_{\{\rho,h\}} := E^{in}_{\{\rho,h\}} + E^{ex}_{\{\rho,h\}} = \begin{cases} -\log(w^\rho_\epsilon \cdot \rho(e(h)|a(h))) & \text{如果 } \rho \in \mathcal{M}_h \\ 0 & \text{否则} \end{cases}$$

配分函数:

$$Z(h) := \sum_{\rho\in\mathcal{M}_h} 2^{-\frac{E_{\{\rho,h\}}}{T_h}} = \sum_{\rho\in\mathcal{M}_h} (w^\rho_\epsilon \cdot \rho(e(h)|a(h)))^{\frac{1}{T_h}}$$

自由能 (Simplicity):

$$F(h) := -T_h \log Z(h)$$

概率:

$$P_h[\rho] := \frac{2^{-\frac{E_{\{\rho,h\}}}{T_h}}}{Z(h)} = \frac{(w^\rho_\epsilon \cdot \rho(e(h)|a(h)))^{\frac{1}{T_h}}}{Z(h)}$$

平均能量 (Variety):

$$E(h) := \sum_{\rho\in\mathcal{M}_h} P_h[\rho] E_{\{\rho,h\}} = -\sum_{\rho\in\mathcal{M}_h} P_h[\rho] \cdot \log(w^\rho_\epsilon \cdot \rho(e(h)|a(h)))$$

熵 (Perfection):

$$H(h) := \frac{E(h) - F(h)}{T_h} = -\sum_{\rho\in\mathcal{M}_h} P_h[\rho] \log P_h[\rho] = H(P_h)$$

如果 $T_h = 1$, 并且一开始就给定整个假设空间 $\mathcal{M}_U$, 那么

$$Z(h) = \xi(e(h)|a(h))$$

且

$$F(h) = -\log \xi(e(h)|a(h))$$

$$= \mathbb{E}_{w_\epsilon}\left[E^{ex}_{\{\rho,h\}}\right] - D(w_\epsilon \| w_h)$$
$$= \mathbb{E}_{w_h}\left[E^{ex}_{\{\rho,h\}}\right] + D(w_h \| w_\epsilon)$$

其中 $F(h)$ 是 “香农–KSA$^\xi$” 追求的, $D(w_h\|w_\epsilon)$ 是 “KL–KSA” 追求的, $D(w_h\|w_\epsilon)$ 可以看作从历史 h 中所能得到的 “惊奇” 的量. $\mathbb{E}_{w_h}\left[E^{ex}_{\{\rho,h\}}\right]$ 是噪音的期望, 所以, 香农–KSA$^\xi$ 在追求模式的同时也在追求噪音, 虽然香农–KSA$^\xi$ 是 “KL–KSA” 限定在确定性环境下的特例, 但 “KL–KSA” 只追求 “惊奇” 而不会故意追求 “噪音”, “KL–KSA” 更合理.

$$\begin{aligned}
H(h) &= H(P_h) \\
&= H(w_h) \\
&= E(h) - F(h) \\
&= E(h) - (-\log \xi(e(h)|a(h))) \\
&= -\mathbb{E}_{w_h}\left[\log w_\epsilon\right] + \mathbb{E}_{w_h}\left[E^{ex}_{\{\rho,h\}}\right] - \left(\mathbb{E}_{w_h}\left[E^{ex}_{\{\rho,h\}}\right] - D(w_h\|w_\epsilon)\right) \\
&= H(w_h\|w_\epsilon) - D(w_h\|w_\epsilon)
\end{aligned}$$

这里的 $H(P\|Q) := -\sum_x P(x)\log Q(x)$ 是交错熵.

给定合适的温度参数, 香农熵可以是有穷的, 所以我们要刻画的主体 (KSA^{ne}) 的内在效用函数就是,

$$u^{in}(t, h_{1:k}) = H(h_{<t}) - H(h_{1:k}) \tag{2.1}$$

香农–KSA$^\xi$ 认为 “自由能” 要大, 但只有在能量恒定时, 最大化自由能才等价于最小化熵. 这里的能量不一定是恒定的, 盲目地追求最大化 “自由能” 意味着盲目地追求随机. KL–KSA 认为 “惊奇” 要大, “自由能 = 噪音 + 惊奇”, 自由能大时, 希望噪音不会随之增大, 所以 KL–KSA 比香农–KSA$^\xi$ 合理得多. 交叉熵意味着—站在现在的角度看过去对万有理论的理解—它应该不断变小 (过去的假设不断被证伪), 同时 “惊奇” 要增大 (代表收获大), 这意味着 “负熵” 要大, 所以 KSA^{ne} 也是合理的.

这里追求内在效用的 KSA^{ne} 相当于追求莱布尼茨意义上的 “智慧”, 所以这就是第2.4节莱布尼茨意义上具有最大 “智慧” 的主体 $\bar{\pi}$,

$$\bar{\pi} = \operatorname*{argmax}_\pi \mathbb{E}^\pi_\rho\left[\sum_{t=1}^{\infty}\left(\mathbb{E}_w\left[R + (E|R)\right] - S\right)\right] = \operatorname*{argmax}_\pi \mathbb{E}^\pi_\rho\left[\sum_{t=1}^{\infty} d(t)u^{in}(0, h_{1:t})\right] \tag{2.2}$$

因为追求效用 u^{in} 是一个单纯的探索过程, 或许可以借助内在效用协调探索与开发的问题. 记通常外部设定的效用函数为 u^{ex}, 定义

$$u(t, h_{1:k}) := T_{h_{1:k}}u^{in}(t, h_{1:k}) + (1 - T_{h_{1:k}})u^{ex}(t, h_{1:k})$$

但这种方法的效果如何只能通过实验验证.

2.6 “先验”与“效用”—形而上的“善”与功利主义的“善”

莱布尼茨提出了“前定和谐”的思想. 莱布尼茨认为, 所有可能世界都有奔向存在的倾向, 越完满的可能世界奔向存在的倾向性越大, 而现实世界是所有可能世界中最完满的. 这就实现了“目的因”与“动力因”的“前定和谐”, 或说实现了形而上的“善”与功利主义的伦理“善”的“前定和谐”, 这或许可看作一种理想的价值追求.

博斯特罗姆 [Bos14] 认为, “智能”(Intelligence) 具有“目标正交性”, 智能体追求的“目的”与追求的能力或“手段”没有任何关系, 可以任意匹配. 而莱布尼茨的“智慧”(Wisdom) 则不同, “智慧”是“目的”与“手段”的统一 (前定和谐).

如果 w_ϵ 是事先已知的, 那么, 根据上面第2.3节的定义, $H(w_h)$ 可看作主体在历史 h 时刻所能获得的“效用”或“完满性”.

可能世界 ν 的总的完满性即为,

$$\bar{U}(\nu) = \mathbb{E}_\nu \left[\sum_{i \geq 1} \gamma^i \left(H(w_\epsilon) - H(w_{h_{1:i}}) \right) \right]$$

所以, 如果已知 w_ϵ, 那么, 类似 $w_{\bar{U}}$, 可以通过对 $\bar{U}$ 的排序, 诱导出某个通用先验. 但问题是, w_ϵ 未知, 只知道对可能世界 ν 的先验信念 (“倾向性”) 应该与可能世界 ν 的总的“完满性”$\bar{U}(\nu)$ 正相关, 如果借助单调连续函数 F 实现“效用”到“先验”的映射, 那么, 下面公式的不动点就是一种沟通形而上的“善”与功利主义的“善”的有意思的先验.

$$G\left(w_\epsilon^\nu\right) := F\left(\mathbb{E}_\nu \left[\sum_{i \geq 1} \gamma^i \left(H(w_\epsilon) - H(w_{h_{1:i}}) \right) \right] \right)$$

因为 $G\colon [0,1]^{|\mathcal{M}|} \to [0,1]^{|\mathcal{M}|}$ 是 $[0,1]^{|\mathcal{M}|}$ 到自身的连续映射, 所以此先验的存在性可以由绍德尔 (Schauder) 不动点定理保证.

根据莱布尼茨的哲学, 通过这个不动点定义的先验可实现“前定和谐”, 可以看作一种理想的价值追求和驱动倾向.

3 小结 — 一个目标正交及“嗑电”问题的解决框架

2018 年初, 杨立昆 (LeCun) 曾和曼宁 (Manning) 就深度学习中“结构”的重要性展开过一场论辩. 杨立昆将结构称为“必要的恶”, 因为结构无非是一些假设, 总是对一些数据是吻合的, 而对另一些数据是错的, 错误的假设需要耗费更多的数据来纠正. 而给定一个缺乏先验结构但体量充分大的网络, 只要训练时间足够长, 总可以逼近真实的结构. 相反, 曼宁认为结构是“基本善”或必要的善, 当我们用神经元网络时, 需要将这种基本善引入到神经网络的设计. 只有具备一定的结构才能从更少的数据中习得更多的知识. 人能够通过少数几次感知就对环境建模, 形成高层抽象知识, 甚至都不需要借助任何外部反馈的奖励, 但这依赖于人丰富的先验知识, 人类经过漫长的生物进化, 大脑的神经元结

构已存储了大量的先验知识, 目前脑科学还没发达到可以把人类的先验结构赋予机器, 所以只能从其它角度探索合适的先验.

本文赞成曼宁的观点, 通过第1.1节的讨论不难看出, 要想获得通用性就必须跳出 “无免费午餐定理” 的陷阱, 为了不落入 “无免费午餐” 的陷阱就必须打破 “块均匀” 性, 要打破 “块均匀” 就必须预设好的先验, 没有好的先验几乎就没法进行通用学习, 为了发展出 “通用性”, 先验或曼宁说 “必要的善” 或本文说的 “形而上的善” 是必须的. 预设 “形而上的善” 的最直接方式就是预设一个可能世界的设计者 — “上帝”. 如果 “上帝” 以完全随机的方式创造可能世界, 那么 “学习” 就几乎是不可能的. 所以需要限制 “上帝” 只创造 “规则” 的可能世界, 这是最低要求的 “形而上之善”. 所罗门诺夫先验就是只考虑所有 “半可计算的” 可能世界, 并假设 “上帝” 偏好奥卡姆剃刀 — 越简单的可能世界越接近现实. 是否可以放宽这些约束呢? 第1.2节从这个角度出发, 进一步构建了一个虚拟 “主体” 与 “上帝” 博弈的模型, 也就是在探讨 “上帝” 以何种概率分布创造可能世界供 “主体” 生存探索. 这可以看做 “上帝” 以可能世界为信道向 “主体” 传递消息. “上帝” 按某个分布输入消息, “主体” 按照某个归纳模型估计消息, 这样, 可能世界的分布可看作 “上帝” 的策略, 而归纳模型可看作 “主体” 的策略. “上帝” 试图最大化期望误差, “主体” 试图极小化期望误差. 这是一个典型的两人零和博弈. 二者博弈的是 “主体” 在可能世界中的学习试错代价. 通过这个模型证明, 即使不假设奥卡姆剃刀, 类似的约束也是必须的, 否则, 任由 “上帝” 对可能世界赋予先验概率的话, “主体” 的认知过程是完全被动的, 即使 “主体”“做最坏的打算, 尽最大的努力” 也无济于事, 因为博弈的结果是占优均衡, “上帝” 占优, “主体” 的期望误差界可以任意的大. 所以, 合适的先验/“基本善”/“形而上的善” 是至关重要的.

如果以 “简单性” 为 “美”, 可能世界的先验 (w) 看作 “似真” 的程度或形而上的 “善”, 效用 (u) 看作功利主义的 “善”, 那么 “美” 和 “善” 不仅相互独立, 而且在没有关于 “上帝” 的神学信仰关照下, “美” 和 “善” 完全可以与 “真” 毫不相干, 但为了使得通用学习成为可能, 又必须接受某种类似 “美” 或 “善” 的指引来求真. 换句话说, 需要借助 “美” 或 “善” 定义 “先验”.

第2节将第1节关于通用归纳的讨论扩展到了通用强化学习的框架下, 通用强化学习引入了行为和效用, 这使得智能体的风险突显出来, 博斯特罗姆 [Bos14] 所讨论的 “目标正交” 与 “工具子目标趋同” 等风险随之而来. 以边沁为代表的功利主义视角下, 不管一元还是多元, 只要外部给定的 “幸福” 可以被定量化的整合成一个单一的标准, 几乎都会导致 “嗑电” 问题, 因为一旦机器探索出一条直达效用的路径, 那么 “嗑电” 就是不可避免的. 而康德式的绝对命令语义模糊, 比如机器人三定律就几乎无法严格形式化, 而且存在可能被利用的漏洞, 总有些隐藏的微妙细节是人们事先难以预料的. 如果纯靠美德伦理学的话, 不仅不能提供正确行动的判别标准, 各种美德的语义定义也很模糊.

通用强化学习的框架采用的是功利主义的伦理学, 为了处理机器伦理问题, 在计算最大 “幸福” 的时候, 需要在 “效用” 和概率上寻求解决办法. 为了尽可能的降低风险, 人工智能专家做了很多探索. 比如, 杜威 [Dew11] 提出了价值强化学习, 试图赋予机器一类而不是一个效用函数, 让机器自己在不断试错中探索更好的效用函数. 帕拉尼潘等人 [Pal+17] 提出了合作逆强化学习, 合作逆强化学习可以看作在功利主义的框架下对美德伦理学的一种刻画, 美德伦理学认为, 一个行动是正确的当且仅当它被有美德的人执行. 如果把人的行为看作有德行的行为, 那么, 合作逆强化学习的机器则通过预测、观摩人的行为探索人的真实意图, 从而也试图变得有德. 这两种方法都故意对智能体的价值观增添不确定性. 第2.1节发展了价值强化学习的想法, 为效用函数的赋权问题给出了一种有意思的尝试.

如果我们为机器加载了合适的先验结构, 机器相当于可以在在更高的抽象水平上学习, 对外部监督或外部奖励的需求也会降低. 奖励不能完全由外部给定, 还需要由内而生的内部效用, 而且尽可能丰富多样, 而不是从特定任务设定的奖励中学习. 在理想情况下, 外部奖励的作用应该尽量降低, 而用合适的内在效用引导. 本文第2.3节、第2.4节和第2.5节通过 “先验” 定义 “效用” 就是试图给出更好的内在效用, 从而引导机器更好地以正确的方式理解世界. 可能的效用函数类的选取仍是一件困难的事, 第2.3节给出了一些可能的选择, 但这仅是一些探索性的尝试, 究竟什么样的效用函数类能够刻画那些有益的价值观不得而知, 也不清楚是否存在某种普适的理想的价值追求. 莱布尼茨 “前定和谐” 的哲学思想背后其实暗含着某种理想的价值追求 — “完满性”. “智能”(Intelligence) 具有 “目标正交性”, “目的” 与 “手段” 可以任意匹配, 而莱布尼茨的 “智慧”(Wisdom) 则是 “目的” 与 “手段” 的统一.

第2.2进一步发展了这种想法, 提出了一种可以把效用函数引导源源不断的转嫁为先验驱动的方法. 对于一个以追求期望效用最大化的 “功利主义” 主体来说, 计算期望效用的先验概率的驱动与效用函数的引导具有一定的等效性, 而加载主流 “价值观” 的方式主要是通过修改效用函数进行的, 但人为加载的内在效用函数与环境赋予的外在效用函数的配合是一个难以处理的问题, 所以可以借助第2.3节的方法设置基于 “形式上先验的善” 的内在效用函数, 然后借助第2.2节的技巧将之再转嫁回通用先验上, 因为靠效用函数引导会面临效用源被智能体劫持的 “嗑电” 问题, 而先验驱动有希望在不通过生硬的价值观拼接的情况下, 在一定程度上避免 “嗑电” 的异化问题. 通过第2.6节的不动点方法, 可以架起先验与效用的桥梁, 将 “目的因” 与 “动力因” 统一起来, 在通用强化学习的框架下刻画出 “前定和谐” 的图景. 但即使这就是理想的价值追求, 机器是否能够帮助人类提升自身的价值追求? 是否有义务选择最好的可能来实现? 如果在实现这种理想的现实世界过程中与人类的利益冲突怎么办? 本文在通用强化学习的框架下, 通过对 “先验” 和 “效用” 的处理, 给出了一些降低智能风险的尝试性策略. 对于如何构建一个符合人类的价值观而不会 “嗑电”、关键时候还允许关机中断的智能体, 则需要整合类似上述的思想, 构建一个统一的通用 (逆/价值) 强化学习框架.

比如, 令

$$a_k^* = \operatorname*{argmax}_{a_k} \sum_{e_k} \ldots \max_{a_m} \sum_{e_m} \left[\sum_{\nu \in \mathcal{M}} w_h^\nu \nu(e_{k:m} | h a_{k:m}) \sum_{u \in \mathcal{U}_h} P(u|\nu, h) u(æ_{1:m}) \right] \tag{3.1}$$

其中 $h = æ_{<k}$.

公式3.1中的 w_h^ν 隐藏的 w_ϵ^ν 可以采用类似第2.6节定义的莱布尼茨先验, 从而促使机器的价值追求在一定程度上具有崇高性 — 为了实现所有可能世界中最完美的.

$P(u|\nu, h)$ 则是采用类似第2.1节的技巧得到的效用函数的权重.

$$P(u|\nu, h) := \frac{\tilde{U}(u, \nu, h)}{\sum_{u \in \mathcal{U}_h} \tilde{U}(u, \nu, h)}$$

其中

$$\tilde{U}(u, \nu, h) := \sum_{z \in \mathcal{Z}_h} \nu(z|h) u(z)$$

$\mathcal{Z}_h$ 是指由 ν 生成的与 h 协调的所有可能的未来历史.

虽然 $u(æ_{1:m})$ 不依赖于模型, 但总的效用函数 $\sum_{u\in\mathcal{U}_h} P(u|\nu,h)u(æ_{1:m})$ 依赖于模型 ν, 从而可以避免盲目“嗑电”.

采用多样化的 $\mathcal{U}_h$ 是为了保证机器价值追求的不确定性, 使其不能确定真实的效用函数, 为人类关键时刻可以直接关机提供了可能. 事实上, 为了让机器追求人的真正偏好, 可以把逆强化学习的方法加入进来, 比如把3.1式中的效用函数改为:

$$\sum_{a_k^H} P(a_k^H \mid a_k) \sum_{u\in\mathcal{U}} P(u \mid a_k, a_k^H) u(æ_{1:k})$$

其中, a_k^H 指人的动作. $P(u|a_k, a_k^H)$ 是机器对比自己的和人的动作产生的对效用函数 u 的信念. 或者定义一个统一的 $P(u|\nu, h, h^H)$, 其中 h^H 是人与环境的交互历史, 这样, 就可以把逆强化学习与基于模型的效用统一起来.

$$V_t^*(h) := \max_{a_k\in\mathcal{A}} Q_t^*(ha_k)$$

$$Q_t^*(ha_k) := \sum_{e_k\in\mathcal{E}} \sum_{\nu\in\mathcal{M}} w_h^\nu \nu(e_k|ha_k) \left[\sum_{u\in\mathcal{U}_h} \sum_{a_k^H} P(a_k^H|a_k) P(u|\nu, ha, h^H a^H) u(hæ) + \gamma V_t^*(hæ) \right]$$

$$a_k^* := \operatorname*{argmax}_{a_k\in\mathcal{A}} Q_t^*(ha_k)$$

参考文献

[Bos14] Nick Bostrom. *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, United Kingdom, 2014.

[Dew11] Daniel Dewey. “Learning What to Value”. In: *Artificial General Intelligence: 4th International Conference, AGI 2011* 6830 (2011), pp. 309–314.

[ELH14] Tom Everitt, Tor Lattimore, and Marcus Hutter. “Free lunch for optimisation under the universal distribution”. In: *Proceeding of IEEE Congress on Evolutionary Computation (CEC14)*. IEEE. 2014, pp. 167–174.

[Goo53] Irving J Good. “The population frequencies of species and the estimation of population parameters”. In: *Biometrika* 40.3-4 (1953), pp. 237–264.

[Hut05] Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability*. Berlin, Heidelberg: Springer, 2005.

[IT04] Christian Igel and Marc Toussaint. “A no-free-lunch theorem for non-uniform distributions of target functions”. In: *Journal of Mathematical Modelling and Algorithms* 3.4 (2004), pp. 313–322.

[OLH13] Laurent Orseau, Tor Lattimore, and Marcus Hutter. “Universal knowledge-seeking agents for stochastic environments”. In: *International Conference on Algorithmic Learning Theory*. Springer. 2013, pp. 158–172.

[Ors14] Laurent Orseau. “Universal knowledge-seeking agents”. In: *Theoretical Computer Science* 519 (2014), pp. 127–139.

[Pal+17] Malayandi Palaniappan et al. “Efficient Cooperative Inverse Reinforcement Learning”. In: *Proc. ICML Workshop on Reliable Machine Learning in the Wild* (2017).

[PV15] Jeffrey Paris and Alena Vencovská. *Pure inductive logic*. Cambridge University Press, Cambridge, United Kingdom, 2015.

[Ris98] Eric Sven Ristad. “A natural law of succession”. In: *IEEE International Symposium on Information Theory* (1998).

[Sch12] Jürgen Schmidhuber. “A formal theory of creativity to model the creation of art”. In: *Computers and creativity*. Springer, Berlin, 2012, pp. 323–337.

[Sol78] Ray Solomonoff. “Complexity-based induction systems: Comparisons and convergence theorems”. In: *IEEE Transactions on Information Theory* 24.4 (1978), pp. 422–432.

[Teg17] Max Tegmark. *Life 3.0: Being Human in the Age of Artificial Intelligence*. Penguin Random House LLC, United States of America, 2017.