深度

人工智能自杀干预：在社交平台的树洞里搜寻需要救助的人

“我止不住地想哭”和“救救我”，哪一个自杀风险更高？为了救援侵犯个人隐私，是否应该？

特约撰稿人 王浠 发自北京 | 2019-07-02

插画：Rosa Lee

“饭饭，我今天又割腕自残了，旧伤未好添新伤。”2018年9月25日晚上21:47，肖雅在微博帐号“走饭”的评论区里留下这句话。每次情绪失控，她都会在这个“树洞”留言或发微博宣泄。

像往常一样，这些谈论自残、自杀的微博只收到零星评论。肖雅微博上的2034个粉丝几乎全是陌生网友，现实生活中的亲友看不到她的抑郁情绪，但“机器人002号”能“看”到。

7900多公里之外的荷兰阿姆斯特丹，“机器人002号”在一台电脑中启动。24小时内，从1335条评论中，筛选出十几条重点关注信息，生成“树洞监控报告”，整个过程不到1分钟。报告中，肖雅的微博被评估为“自杀风险9级”，“树洞行动救援团”立即开始行动。

“树洞行动”始于2018年4月2日，由荷兰阿姆斯特丹自由大学（Vrije Universiteit Amsterdam）人工智能系教授黄智生发起，通过一个智能主体（intelligentagent，又称机器人）巡视各类社交媒体，使用其核心的知识图谱技术（又称语义技术）发现高风险的自杀人群，再由上百名志愿者进行自杀干预。类似的项目还有2017年开始的“心理地图PsyMap”，由中国科学院心理所计算网络心理实验室负责人朱廷劭发起，用人工智能深度学习的方式找出有自杀意念的微博用户，再通过私信进行心理危机干预。

据世界卫生组织统计，全球每年有80万人死于自杀，每40秒就有一人尝试自杀；在15到29岁的青少年人群中，自杀已经成为第二大致死原因。作为月活跃用户超过21亿的全球最大社交媒体，Facebook于2017年上线“自杀检测”功能，通过人工智能和模式识别等技术辨别有自杀倾向的用户，并在第一时间为后者提供帮助。更多科技公司亦陆续推出类似服务：Google对搜索“自杀”等关键词的用户优先显示自杀预防机构广告和劝导内容；AIBuddy项目为现役军人的子女（自杀机率较高）提供虚拟交流服务，并向监护人提供心理健康报告；科技公司Bark.us通过机器学习分析了5亿多青少年发布的信息，已成功挽救25人的生命。

AI Buddy项目为现役军人的子女（自杀机率较高）提供虚拟交流服务，并向监护人提供心理健康报告。网上截图

据北京心理危机研究与干预中心的调查，中国每年约有 28.7 万人自杀死亡，其中一半以上被诊断为抑郁症。但目前，中国的自杀预防工作仍处于起步阶段，大多数精神卫生专业人员只能被动等待患者拨打热线或去医院就诊。世界卫生组织（WHO）数据表明，目前中国抑郁症患者数超过 4000 万，但就诊率不足10%。

与此同时，自杀信息和抑郁情绪在微博树洞、约死QQ群、自杀论坛等互联网的隐秘角落里蔓延。人工智能技术研究者们决定主动靠近有自杀倾向的人，将干预过程前移，用科技守住生死的门。

树洞：在这里，病耻感被降到最低，可以自由表达想死的情绪

“我有抑郁症，所以就去死一死，没什么重要的原因，大家不必在意我的离开，拜拜啦。”2012年3月18日，网友“走饭”因抑郁症自杀，账号在这一条微博发布后永远停摆。其后七年，该账号的评论区成了微博上最大的“树洞”，有抑郁情绪或自杀倾向的网友聚集于此，留言以每天上千条的速度不断叠加。至今，这条临终微博的转发量超过10万次，留言数逾152万。

肖雅不记得自杀过多少次。她被抑郁症折磨七年，常常失眠，睡醒了会莫名其妙地哭。在现实生活中，肖雅没有可以倾诉的朋友，偶尔在熟人社交网络QQ空间里说两句还会被视作矫情：“这个年代还有林黛玉啊！”于是，永远不会回复她的“走饭”成了安放情绪的“树洞”，微博则变成“留遗言的平台”：跳河之前她发“夕阳无限好，只是已黄昏”，服药自杀前她发“希望一切可以重来”，打开煤气炉后她搬个板凳坐在厨房刷微博……

专注协同“抑郁康复”的公益组织“郁金香阳光会”表示，人们常认为心理是可以被调控的，将很多心理疾病归结为“性格内向”、“意志薄弱”、“懒惰”。因此，很多抑郁症患者会把自己的负面情绪视作“耻辱”，拒绝与亲友沟通。但在微博树洞这样的集体性聚集地，病耻感被降到最低，他们可以自由表达抑郁或想死的情绪，而不会受到指责和偏见。

中科院心理所计算网络心理实验室也发现，自杀的主要人群（15到30岁人群）和网络使用的主要人群有很大重合，以青年人为主力军的部分网络用户会在以微博为代表的社交媒体

上表达自杀意念、直播自杀甚至相约自杀。

2014年，实验室研究员朱廷劭和团队研究发现，自杀死亡用户的微博互动更少、更加关注自我、更频繁地使用表达排除意义的词语。他们在情绪上偏向于负面表达，负向情感词的比例大于80%，谈论死亡和宗教远多于工作和家庭。

在这一发现的基础上，朱廷劭及团队利用人工智能深度学习技术，进行数据筛选，识别有自杀风险的个体，对其进行心理危机干预。2017年7月，“心理地图PsyMap”项目的在线自杀主动预防系统正式上线。

朱廷劭认为，与线下寻找案例相比，社交媒体用户足够多，用户数据公开可见、易收集。研究初期，团队曾尝试在微博全平台识别有自杀意念的人，但效果不好。树洞的出现，让识别范围缩小，定位也更加精准。

每晚零点，自杀识别模型抓取24小时内“走饭”等“树洞”里的最新留言，生成一份文本，包含用户ID、留言时间、留言内容以及模型评估结果。结果只分0和1两种，0代表无自杀意念，1代表有。经人工审核，系统会自动用“心理地图PsyMap”微博账号向有自杀意念的用户发送私信进行心理危机干预：

“你好，我们是北京社工委领导下的，中科院心理所咨询师志愿者团队。我们在走饭的微博中看到了你的评论。你现在还好吗，情绪状态怎么样？”询问后面跟着一条调查问卷链接和北京市心理危机干预中心的24小时免费热线电话。这样的私信肖雅收到过好几次，但很少回复，“像机器人发的”。每天18：00到22：00，14位志愿者会两人一组轮流值班，通过私信和填写过问卷的用户聊天。“没什么用，和医院一样”，肖雅说。

每天下午6时到晚上10时，14位志愿者会两人一组轮流值班，通过私信和填写过问卷的用户聊天。图：Imagine China

相比之下，她觉得“树洞行动”的沟通更人性化。2018年9月27日，在走饭微博评论区留言的隔天早上，肖雅收到一条陌生人的私信，和自杀无关，只是聊警察的事。肖雅很关注警察行业，微博粉丝里有很多各地的警察。看到对方自称是警察家属，肖雅就和她有一搭没一搭地聊了起来，后来还加了微信、留了电话，至今，肖雅也不确定对方的真实身份。她觉得，这种聊天对治疗心理问题也没用，但感觉能把自己“往回拉一把”。

给肖雅发私信的是“树洞行动”的志愿者。2018年3月，从事人工智能研究超过30年的黄智生看到一则关于“树洞”的报导后，提出通过人工智能技术识别抑郁症患者并提供救助。4月2日，“树洞行动救援团”项目正式启动。

相较于“心理地图PsyMap”的线上心理危机干预，“树洞行动”有着更明确且实际的目的——救助。机器人对留言中的自杀意念、方式和时间等词语进行语义分析，筛选出已有自杀计划的用户，生成“树洞监控报告”。每天晚上23:00左右，黄智生把包含用户ID和微博内容的报告转发到微信群内，遍布世界各地的志愿者登录自己的微博小号，自发与其中的一两位私信聊天。获得联系方式后，志愿者会尝试与其亲人联系，紧急情况下也会寻求当地警方帮助。

目前，项目微信群里已有来自各行各业的230名志愿者，其中40多名精神健康和心理学专家以及40多名心理咨询师负责为救援提供指导。据其统计，截至2019年3月27日，“树洞行动”已对超过760人（次）展开救助行动，其中320多人（次）自杀行为被成功阻止。

人工智能的局限：如何精准判定自杀风险

“人工智能的判断也存在误差”，运营“心理地图PsyMap”的博士生刘兴云指着最新生成的文本，微博留言“我止不住地想哭”被判定为“1”，“救救我”却被判定为“0”。

刘兴云介绍，自杀识别模型建基于中国社交媒体自杀词典，样本量超过7000，分为自杀想法、自我伤害、生活事件等不同类别。每个类别再细分为三个等级，等级越高，权重越大。例如，在自杀想法类别中，“消失”、“结束”、“放手”的权重为1，“该死”、“地狱”、“下辈子”的权重为3。在这个基础上，项目组又补充了大量和药物、情绪、精神疾病、人格特

征等相关性显著的关键词。另外，刘兴云和其他几位同学做过好几个月的人工标注，每周为七八千条微博评论划分风险等级。这些人工判定的结果和关键词库一起被输入计算机训练建模，投入使用时，模型精准度达到80%。

但人工智能与真人的判定结果仍有差异。例如，“饭饭，很快就要看到你了”这句留言中没有负面关键词，机器判定它没有问题，但在“走饭”评论区的语境下就显露出自杀风险。

2012年3月18日，网友“走饭”因抑郁症自杀，账号在这一条微博发布后永远停摆。摄：Stanley Leung/端传媒

黄智生认为，“心理地图PsyMap”采用的深度学习技术“无法考虑情境”，而“树洞行动”的核心技术知识图谱可以更细致地分析词语之间的逻辑关系。比如，有人连发三条“我不想死”的信息，采用深度学习方法的模型会依据关键词“死”将其判定为存在自杀风险，但采用知识图谱技术的模型能够识别他表达意思刚好相反。在救援过程中，人工智能模型也在不断迭代。2018年12月16日，“树洞机器人004号”上线，精确度提高到80%以上。

“树洞机器人”将微博留言按自杀风险分为十级，救助六级及以上的用户。（自杀风险等级为：6级：自杀划已在计划中；7级：自杀方式已确定，日期未明；8级：自杀方式已在计划中，自杀日期大体确定；9级：自杀方式已确定，近日内可能进行；10级：自杀可能正在进行中）“我们不想、也不敢发现更多有自杀意念的人”，黄智生说。根据统计，10条左右的高自杀风险微博发到群内，“能救的最多只有三个”。

线上技术的另一层局限在于，难以在线下得到当事人的配合或亲属的理解。“主动权并不掌握在我们手中”，黄智生说。

2018年4月18日，“树洞机器人”还没上线，团队成员在抓取数据时发现了一条约死信息：一名武汉女孩将在五一假期自杀。黄智生发动群友搜集信息，辗转联系上其父母，他们坚称女儿没病。5月1日，女孩没有按计划实施自杀，但47天后，家人告诉志愿者，她还是选择结束了生命，在微博留下了最后一句“拜拜”。

还有一次，志愿者联系到一位计划自杀者的家长，反而招致威胁。“我女儿好不好，我还不知道？”这名家长说要到法院起诉志愿者，“要是真跳楼，肯定是你们害的！”无奈之下，志愿者只能天天给女孩发私信，关注她的动态。

隐私保护和救援，孰轻孰重？

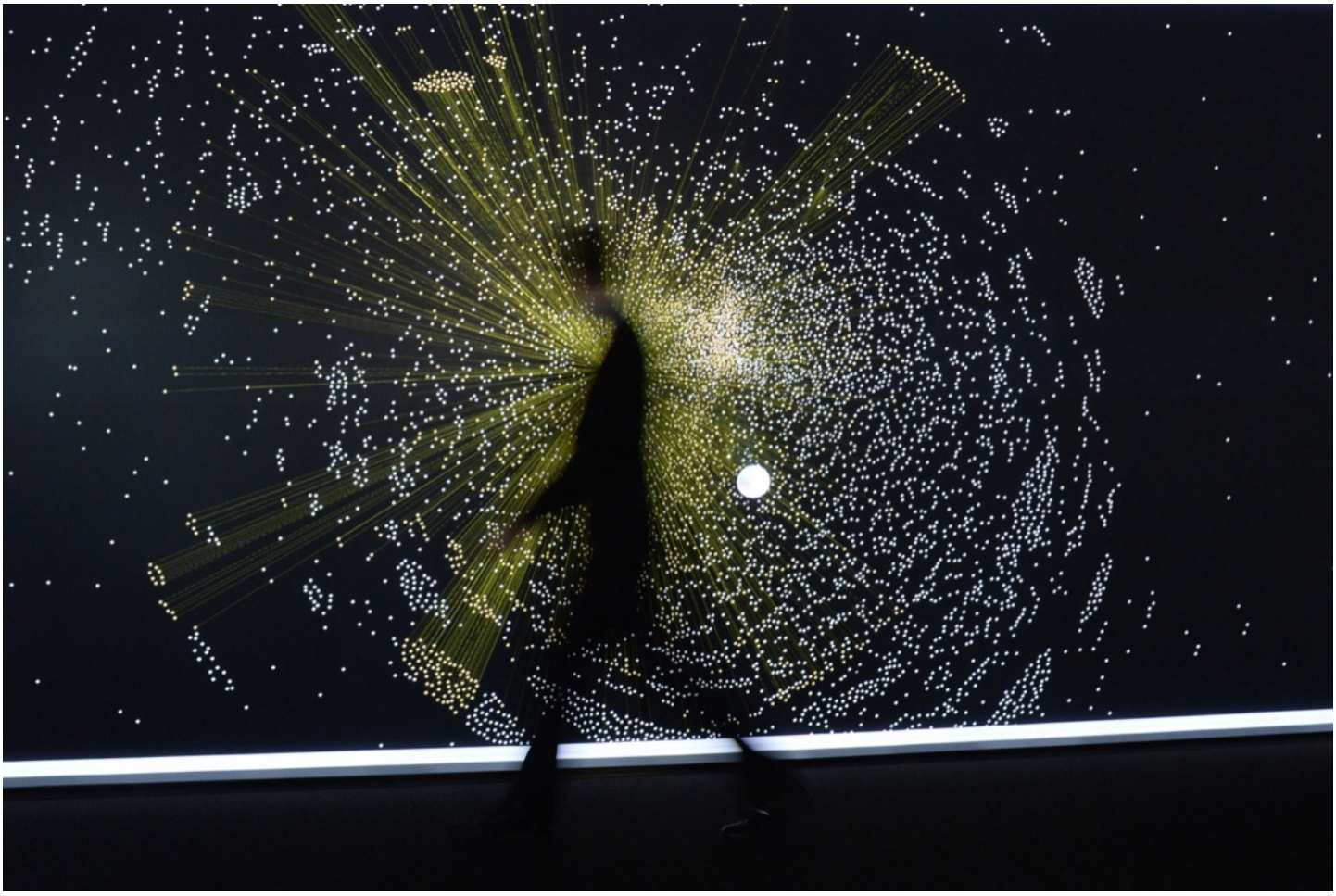
情绪稳定后，曾受助于“树洞行动”的小杨加入了志愿者团队，兼职做文字记录。大家在群里讨论如何联系有自杀倾向的网友时，小杨赫然发现了朋友的名字，随后，她的照片、微博和联系方式逐一被发到群里。“既没有效果，又侵犯个人隐私！”小杨生气地“删除并退出”了该群。

这不是“树洞行动”第一次面临质疑。从微博数据挖掘，与被识别的用户私信，到“人肉搜索”来获取受助者个人信息，隐私问题潜藏在项目的每一个环节。

2017年，Facebook的“自杀检测”功能甫一上线即引发争议。其“秘密地”对平台中的图片和文字进行监测，让用户有“被监视”的感觉。今年2月11日，杂志《内科学年鉴》曾刊文称，Facebook在筛选用户帖子，识别那些有自杀倾向的人以及提醒急救时，缺乏透明度和道德规范。

对于小杨的质疑，“树洞行动”更新的一版“网络自杀救援指导性建议”中指出，志愿者不得向其他无关人员提供受助者的个人信息，群组讨论也不可向第三方提供。但在数据来源方面，黄智生认为，纯公益项目“树洞行动”通过API接口抓取的都是微博公开数据，“不存在隐私问题”。更多的顾虑存在于救援环节。黄智生总结道，如果没有生命危险，隐私保护优先于救援；反之，救援高于隐私保护。

“心理地图PsyMap”项目在隐私方面则更为保守。朱廷劭在“SELF格致论道讲坛”的演讲中指出，不管出于怎样的动机，首先要保证别人的隐私。“侵犯别人隐私为代价做的任何事情都没有意义，甚至是对别人的不尊重”。因此，发出私信后，志愿者需要等待微博用户主动回复或填写问卷，才会进一步私聊。



“心理地图PsyMap”项目在隐私方面则更为保守。发起人朱廷劭指出，不管出于怎样的动机，首先要保证别人的隐私。摄：Christof Stache/AFP/Getty Images

为设计私信内容，朱廷劭团队联合香港大学香港赛马会防止自杀研究中心、北京回龙观医院北京心理危机研究与干预中心人员组织访谈、设计问卷，了解有自杀意念的人希望看到什么内容。团队向4222名有过自杀意念表述的群体发送过问卷邀请，725个回复中78%的人表示不反对收到私信。在实际操作中，正面回复超过总发送量的20%，朱廷劭已很满意。偶尔也有对私信内容产生质疑的情况。有网友问，“你怎么知道我要自杀？”也有人情绪激动，“不要你管！”刘兴云不确定，这些负面回复是否源于隐私被侵犯，“应该不是问题，解释清楚就好。”

2017年9月，人工智能模型识别到一个服药自杀的微博用户，志愿者守在电脑前，用“心理地图PsyMap”的微博帐号跟他从晚上六点聊到凌晨。先开始是试探性地询问，“我现在很担心你的安危，你能不能先把药放在一边”。但对话框那头，对方开始文字直播吞药自杀。

情况变得紧急。志愿者需要联系他的家人或当地警察，但在此之前，他们要征得本人同意。好在对方愿意提供联系方式，也默许警察介入，自杀干预最终成功。

但更多时候，志愿者的意见征询没有得到回应，刘兴云也不知道对话框那头的沉默是否代表悲剧已经发生。“如果他不愿意，那我们也没办法。”刘兴云苦笑一下，从心理咨询和危机干预的专业角度来看，她必须接受伦理问题带来的限制。

边界：“一旦建立情感，就不想再看到他出事了”

“心理地图PsyMap”的14名志愿者一直守着心理咨询的边界，始终使用公共账号联络，不与用户产生其他联系，每天四小时的值班结束后会和对方约定下次时间，除非紧急情况才会延长或另约时间。这些是心理咨询师行业的规范，“不能让咨询者产生依赖”。

刘兴云记得，团队遇见过一个“粘性特别高的小孩”，她注册了两个微博账号，每天咨询时间一到就会出现，常常过了十点还在跟志愿者倾吐自己日常生活中的琐碎情绪。很多人都觉得压力很大。月度的专家指导会上，导师指出，心理咨询是一个长期的过程，我们不能帮她解决长期的情绪问题或根除自杀的想法，只能提供一次心理危机干预。志愿者和那个

女孩反复沟通后，将咨询频率逐渐减少到一周三次、一周一次，建议她向更专业的心理咨询师或医生求助。

相比于“心理地图PsyMap”，“树洞行动”拥有更庞大的队伍，但专业水平参差不齐，缺乏统一安排，全凭志愿者自发行动。

“树洞行动”拥有庞大的队伍，但专业水平参差不齐，缺乏统一安排，全凭志愿者自发行动。图为“树洞行动”发起人黄智生。网上图片

去年八月，在法国某大学计算机系任教的马丽玲和一位存在高自杀风险的网友聊了几次。看着对话框里不时冒出的终极问题，马丽玲很焦虑。她没有任何心理咨询的经验，对方的倾诉让她觉得“完全接不住”。受助者回复的次数越来越少，马丽玲对自己的沟通技巧也没了信心，“不敢再联系其他有自杀意念的人”。

志愿者群里，和马丽玲一样的非专业人士不少，缺乏救助和沟通技巧。有的志愿者为了阻止被情感问题困扰的自杀者，承诺与其成为现实生活中的男女朋友；还有的为了帮助陷入经济困境的网友，自发在群内筹款救助。

针对这些不规范的救助方式，“树洞行动”邀请了专业人士对志愿者指导、培训。此外，项目成员拟定了一份“网络自杀救援指导建议”。32页的PDF详细列出了规范的救援程序以及应该避免的问题，包括什么情况下可以询问个人信息、什么情况下应与警方联系？帮助志愿者在救援的每一个环节找到参考意见。

可边界依然很难守住。在“树洞行动”中，一位拥有多年从业经验的咨询师说，自己正一点一点地突破边界。“一旦建立情感，就不想再看到他出事了。”自杀干预成功后，志愿者会和对方保持联系，关注他们的微博或朋友圈动态，一些人会为对方提供长期的情感陪伴，甚至尝试帮他们解决现实问题。

比如，马丽玲曾尝试为受助者介绍工作，“从根源上帮助他们”。

去年12月，东北某大学计算机系的王老师在“树洞行动救援团”的微信里提供了一些数据标注员的兼职工作。马丽玲很快推荐了几个人，鼓励他们，“咱们有工作了，能赚很多钱！”1月31日，6位处于情绪稳定期的抑郁症患者加入了王老师的人工智能文本处理公司，尝试进行简单的数据标注工作。

跟他们讲解任务要求时，经理小吴有些紧张。“树洞行动”的心理咨询师曾叮嘱她，抑郁症患者很敏感，在沟通方式上需要注意，“要有耐心”。小吴和他们交流基本只发文字，“看一遍没问题再发”。平时习惯性的说法也要调整，“我这么说你能理解吗？”要改成“我有没有表达清楚？”她总是担心自己会在无意间伤害他们。有次和其中一个人私聊，小吴随手发了个

动态表情想表示亲近，对话框里却传来一个撇嘴的表情。小吴有点手足无措，赶紧发文字解释。“之后我再也不敢发表情了，最多加个波浪线。”

到现在，还没有一位受助者通过培训审核。“他们对有些任务的理解有些偏差”，小吴打算，等有较简单的工作再安排给他们。

说起这个结果，马丽玲很懊恼，“还是太不成熟了。”和几位参与试标的受助者沟通后，她意识到这样的安排会给他们带来更大的挫败感，觉得连高科技手段都帮不了自己。后来，马丽玲还在继续帮他们推荐工作，但话术调整成“先试试看，慢慢找解决办法”。她不敢给“树洞行动”贴上“人工智能”、“高科技”的标签，也不敢说聊天或陪伴产生了什么实质性帮助。相比之下，介绍工作可能更为实际。

受助者喜欢在深夜向马丽玲吐露心事，按照法国时间，她正在工作，但仍是每条消息必回。马丽玲不在意心理咨询师口中的“边界”，“我不是专业的心理咨询师，只是一个陪伴他的朋友”，她说，“这两个角色都是重要的。”

随着“树洞行动”的发展，项目对志愿者的专业性要求越来越高，规范化的人员管理和后续的救助都需要更多的资金投入。黄智生认为，把救援与产业结合是条出路。一方面，很多精神疾病的根源在于工作、家庭、学校等方面的问题，彻底治疗需要各方配合，形成一个完整的救助生态链；另一方面，人工智能技术的发展需要一些基础工作的支撑，如数据处理、数据标注等，存在大量的产业需求。黄智生希望建造一个关爱中心，既能提供康复疗养功能，也能为抑郁症患者提供工作机会。“如果我们把救助行动作为一个产业做起来，就不仅仅停留在公益层面，还能带来一定的经济效益。”

目前已有好几家医院和心理咨询机构找到黄智生，希望向人工智能监测到的抑郁症患者提供药物或咨询推介，但他都拒绝了。“盈利必须和公益救助分开”，黄智生说，“一旦带有商业目的，隐私和伦理问题又会出现。”

黄智生曾向中国两个城市的政府寻求支持，希望为获救的自杀者建立关爱中心，为其营造新的生活环境加速病情恢复，但未获得回应。他还曾与华为、百度等科技公司商谈过合作或资金支持，目前暂无实质性进展。黄智生认为，最合适“树洞行动”的方式是注册民营非

企业单位，向基金会申请资金，或者通过其他技术公司盈利为项目提供资金，以此守住公益与商业的界限。

（应采访对象要求，肖雅、马丽玲为化名。）

AI

热门头条

1. 中国「古装剧禁令」风波：为什么一幅微信截图，业界就全都相信了
2. 回应赵皓阳：知识错漏为你补上，品性问题还需你自己努力
3. 连登仔大爆发：“9up”中议政，他们“讲得出做得到”
4. 香港回归22周年，七一升旗礼、大游行、占领立法会全纪录
5. 梁一梦：反《逃犯》修例，港府算漏了的三件事
6. 记者手记：我搭上了罢工当晚的长荣班机
7. 马岳：“反送中”风暴—目中无人，制度失信，残局难挽
8. “突如其来”的新一代：后雨伞大学生如何看社运
9. 专访前大律师公会主席陈景生：香港现在这处境，我最担心十几廿岁的年轻人
10. 读者来函：承认我们的无知，让出一条道路给年轻人吧

编辑推荐

1. 运动中的“救火”牧师：他们挡警察、唱圣诗、支援年轻人
2. 金山上的来客（下）
3. 从争取“劳工董事”到反制“秋后算帐”，长荣罢工之路为何荆棘?
4. 吉汉：暴力抗争先天有道德包袱吗?
5. 金山上的来客（上）
6. 归化球员能“拯救”中国男足吗?
7. 进击的年轻人：七一这天，他们为何冲击立法会?
8. 荣剑：中美不再是中美，中美依然是中美，中美关系下一步

9. 贸易战手记：华府的关税听证会上，我围观了一场中国制造“表彰”会

10. 徐子轩：由盛转衰——G20大阪峰会后，全球政经的新局面

延伸阅读

Google开源人工智能算法将改变世界？

救赎他人和自我，还剩多少可能？

为何随机杀人、自杀潮层出不穷，我们该如何面对这样的社会？

防止人工智能成为人类“终结者”，Google提出“红色按钮”机制

机器学习新应用：Google兄弟公司Jigsaw将帮你摆脱网络语言暴力

KnowYourself：疗愈焦虑时代的中国年轻人，它为何能说中250万粉丝的心事？

“我每天晚上打开KY的推送时，都会觉得它好像在说我最近困惑的事。”

傅景华：人工智能面前，是否人人平等？

平心而论，以收集大型数据配合自动处理系统为手段，把人进行社会分类为目的，再向各类型民众施以不同待遇，这些都并非中国首创。那么，要如何走出所谓是否“妖魔化”的讨论？