Половина Закона Робототехники

Азимовские **"Три Закона Робототехники"** до сих пор часто встречаются не только в форме более или менее прямой аллюзии в произведениях современных фантастов, но и, к сожалению, в качестве цитат у некоторых социологов и представителей когнитивной психологии – как попытка формулировки неких *"принципов организации взаимоотношений между искусственным интеллектом и человеком"*. Трудно поверить, однако много людей на самом деле принимают эти высосанные из пальца *"три главных правила"* за некие философские принципы, за что-то, имеющее право претендовать на статус этических заповедей ИИ. Не умаляя достоинств замечательных произведений Азимова, приходится все же заметить, что повторяющиеся хороводы вокруг той троицы, на которой зиждутся все поведенческие механизмы антропоморфных роботов в одноименной вселенной его рассказов, требуют критического разбора – раз уж её стали наделять статусом заповедей, достойных воплощения в реальных системах искусственного интеллекта.

Для начала напомним эту троицу:

1. Робот не может причинить вред человеку или своим бездействием допустить, чтобы ему был причинен вред.
2. Робот обязан выполнять приказы человека, за исключением случаев, когда это противоречит Первому Закону.
3. Робот обязан заботиться о собственной безопасности, при условии, что это не противоречит Первому и Второму Закону.

Если внимательно проанализировать эти максимы, становится понятным, что из первой, как из кантовского императива, органично и безусловно вытекает вторая, а третья имплицитно содержится в двух предыдущих. Более того, если опускаться в описании поведенческого регламента робота на уровень детализации, который характерен для третьего закона, мы можем расширить этот список на сколь угодно много пунктов, например: *“Робот должен уважать права остальных роботов на реализацию собственных целей, до тех пор, пока те соответствуют первым двум законам, однако в случае конфликта с третьим апеллировать к авторитету человека”* (чтобы исключить конкуренцию между роботами за ресурсы, необходимые для их функционирования), и так далее.

Однако эта избыточность свойственна не только третьему закону, первые два также не имеют логической границы – другими словами, не имеют четкого формального определения самих себя. Робот не способен предотвратить причинение вреда человеку своим бездействием, если он заранее не держит под контролем собственный гомеостаз, если он игнорирует человеческие приказы, а также если он безответственно принимает к исполнению команды, приводящие к причинению вреда (как самому хозяину, так и остальным людям). В сущности, т.н. второй и третий законы представляют собой не более чем варианты прикладной детализации первого, однако количество подобных детальных описаний, как мы легко можем убедиться, бесконечно.

Как мы видим, достаточно внимательно посмотреть на Первый Закон, чтобы стало ясно, что Второй и Третий не просто не нужны, но более того – вредны, как всякой логической системе вредно введение дублирующих утверждений, а также тех, которые сводят абстрактную идею к перечню ее вариантов реализации. А между тем, являясь логическим базисом, эта система ни в коем случае не должна содержать подобные слабые места, так как именно на ее формализме выстраивается весь функционал "позитронного мозга".

Однако это только начало. Продолжим анализ азимовской фантазии. Пока мы лишь увидели, что Первый, и он же Единственный Закон робототехники (впрочем, не только ее), просится лаконизироваться в столь же компактную, сколь и бессодержательную формулу: **"Робот должен делать то, и только то, что является благом человеку, понимая суть его блага раньше, глубже и отчетливее, чем способен это сделать сам хозяин, упреждая и исключая при этом из сферы забот последнего второстепенные мелочи"**. Только при таком определении действия (и бездействия) робота можно гарантировать недопущение им вреда человеку и достижение чистой рафинированной пользы (в идеале, избавляя хозяина от забот о своем слуге).

OK, пусть так. Однако, если эта формулировка будет реализована, мир азимовских роботов станет невозможен – исчезнет фундамент практически всех описанных в произведениях автора забавных психологических коллизий между хозяевами роботов и этими милыми железными чурбанами. К сожалению, в подобной формулировке Главного Закона имеется слабое место, которое не позволяет построить на подобном фундаменте ничего, кроме литературных фантазий. Это место – употребление далекого от формальной строгости выражения "*вред человеку*", а также необходимая целеориентация робота на столь же расплывчатую "*пользу*". Здесь нелишне будет заметить, что рассказы Азимова как раз и посвящены теме противоречий, неизбежно возникающих между различием в восприятии человеком и роботом одних и тех же явлений, а также трудностям их взаимопонимания при формально сходной постановке задач. С этой точки зрения произведения являются одними из лучших в жанре мировой фантастики, но, к сожалению, именно поэтому в них нет места такому изложению руководящих роботами Законов, которое во главу угла поставил сам автор. Какими должны быть законы робототехники, чтобы из них логически вытекали поведенческие паттерны азимовских роботов,

мы оставим за пределами данной статьи, поскольку это не входит в область ее задач. Полагаю, фанаты творчества Азимова успели предложить немало остроумных формулировок (сомневаюсь, впрочем, в их содержательной ценности с точки зрения подлинных принципов регламентации поведения ИИ). Сейчас речь идет о той популярности, которую пресловутые "Три Закона робототехники" обрели вне рамок развлекательной литературы – абсолютно неоправданно, как мы сейчас убедимся. Итак, вернемся к ним и попробуем установить степень их формальной пригодности.

Похоже, что Азимов не зря оставил за бортом этические рассуждения и какие-либо попытки дать определения понятиям "польза" и "вред", сознавая, что для решения этой задачи ни в три, ни в триста тридцать три, ни даже в три тысячи формулировок ему не уложиться. Сознавая?.. Возможно, что он не отдавал себе отчета в сложности затронутой темы. Есть основания подозревать его в этом, потому что на страницах своих рассказов автор допускает странные противоречия. Сперва он устами своих героев неоднократно утверждает, что *"Три Закона робототехники **не могут не работать!"*** – как минимум самый первый (в одной из историй его суть была упрощена, но это не играет существенной роли). Однако вместе с тем в рассказе "Лжец" Азимов повествует о том, как робот, прошедший все (sic!) базисные тесты (одним из первых среди которых был тест пресловутых законов), наконец-то *"приступил к решению самых сложных – **этических** задач"*. На месте этого робота (кажется, его звали Эрби) следовало бы воскликнуть со всей эмоциональностью персонажа известного мультфильма: ***"Шо, опять?!"*** Зачем, скажите пожалуйста, повторять проверку того, что у данного робота обязано было функционировать безукоризненно – судя по результатам самых основных тестов его позитронного мозга?! Если он смог справиться с задачами, которые ставит перед ним Первый Закон, ему уже не нужны никакие проверки, это уже не просто робот, а намного превосходящее нас существо, которому можно поручить решение всех проблем на Земле. Ведь безошибочно определить, что для конкретного человека является "*пользой*", а что "*вредом*" (как прямым, так и косвенным) при произвольном наборе условий, обстоятельств, задач и прочих факторов – не способен не только сам этот человек, но и целая армия врачей, философов, психотерапевтов во главе с самим господом богом (если бы этот криворукий ремесленник когда-либо существовал). Если же учитывать интересы не одного, а нескольких лиц, то задача из неразрешимой превращается в неразрешимую. *"Но почему?* – воскликнет читатель. – *Ведь это так просто!"*. Увы, это не только отнюдь не просто, но совершенно невозможно, если оставаться в условиях жесткой формулировки целей деятельности азимовского робота и функциональных пределов их реализации.

В своей обыденной жизни мы с вами не задумываемся о той густой паутине допущений и бесконечной цепи негласных конвенций (не имеющих никакой формальной природы), которые пронизывают всю нашу деятельность и служат фоном в нашей коммуникации. Даже элементарная просьба товарища: *"подай мне стакан с водой"* исполняется нами так легко и естественно лишь потому что мы не ломаем себе голову над возможными последствиями того, что: человеку, обратившемуся к нам с этой просьбой, в данный момент необходимо воздержаться от удовлетворения жажды; что он может случайно пролить воду на оголенный провод или плеснуть ее на открытую емкость с кислотой; что вода может быть недостаточной степени очистки для текущего состояния его организма; что он только что выпил слишком много жидкости... подобных условий в любой ситуации будет бесконечное количество. Каждый из нас, будучи свободным человеком, легко игнорирует подобные фоновые условности, доверяя человеку-заказчику и оставляя **на его ответственности** все последствия – как изначальной цели употребления стакана воды, так и потенциальных возможностей с негативным эффектом, сопутствующих данной нехитрой процедуре.

Человек – но не азимовский робот, претендующий на самостоятельный контроль за *"пользой"* и *"вредом"* (без чего вся азимовская вселенная позитронных мозгов с hard wired тремя заповедями рассыпается в прах). Связанный вышеуказанным регламентом своей деятельности слуга не способен поступить так легкомысленно (делегируя ответственность за свои действия третьему лицу), поскольку не обладает ни **свободной волей**, ни собственной **интенциональностью** (неизбежно вступающими в конфликт с Законом). Его единственная задача – быть бесконечно исполнительным при столь же бесконечной ответственности за совершаемые им действия и любые их последствия. Понятно, что какая-либо попытка привести эти противоречивые требования к соответствию если и не вызовет перегрузку логических цепей бесконечной рекурсией, то, как минимум, приведет к полному останову всей системы. Причем вернуть из этого состояния данного робота можно будет лишь после глобального reset'a – с полной очисткой всего буфера последних поступивших команд и всех последствий их обработки. Между прочим, судьба Эрби в упомянутом выше рассказе отчетливо иллюстрирует эту коллизию – а ведь он оперировал не полным владением ситуации, а лишь ее оценкой в мысленных пожеланиях окружающих его людей.

Все это свидетельствует о том, что на самом деле азимовские роботы проявляли какую-либо активность **не благодаря** подчинению Первому Закону робототехники, **а вопреки ему**, в игнорирование его! Автор настолько увлекся персонификацией и наделением своих любимцев антропоморфными чертами, что помимо заимствования внешних черт допустил (вопреки декларациям пресловутых Законов) некую свободу воли у них. Лишь таким образом он добился того, что его роботы стали действовать как люди. Как люди, перенесшие лоботомию, но все-таки – люди. Если бы *"U.S. Robots and Mechanical Men, Inc."* (компания, производящая роботов в его рассказах) на самом деле оставалась в рамках оговоренных им же законов, мы бы получили слугу, который готов ловить каждое наше слово, внимательно выслушивать любое наше приказание – и тут же мгновенно зависать намертво, неизменно, впрочем, сохраняя пластиковую улыбку исполнительного идиота (friendly interface обязывает).

Подобные случаи неизбежны, когда сложные понятия, не сводимые к формальному языку, включаются в определения, казалось бы, простых и понятных всем вещей, заимствованных их набора т.н. "*здравого смысла*". Но давайте еще раз посмотрим на главную часть "свернутого" варианта Единственного Закона робототехники:

"Робот должен делать то, и только то, что является благом человеку, понимая суть его блага раньше, глубже и отчетливее, чем способен это сделать сам хозяин."

Ничего не напоминает? В такой формулировке мы всего лишь расписываемся в старом как мир атавистическом желании иметь бесконечно умелого раба, предвосхищающего все наши желания, наделенного при этом безграничным функционалом для их осуществления. Естественно, что ни к робототехнике, ни к Искусственному Интеллекту эта мечта-анахронизм не имеет никакого отношения, всецело оставаясь в сфере рудиментов нашего самосознания, неизжитых детских (в аспекте степени социальной зрелости) комплексов, социального пуэрилизма и вериг юнговских архетипов. Другими словами, фантаст всего лишь переформулировал на современном языке (с учетом моды на техно-культуру и ее атрибутику) древнюю пещерную потребность неокрепшего, несамостоятельного и боящегося самого себя человека – потребность в обладании неким **всемогущим, но управляемым** (*да-да, наш добрый знакомый Старик Хоттабыч!*) существом: *"Боженька, позаботься обо мне лучше*

меня самого, а я тебе в качестве жертвоприношений обещаю своевременно смазку менять и сервопривод апгрейдить! Хотя еще лучше, если и это ты сам будешь делать."

Возможно, именно поэтому данная идея настолько привлекательна для подавляющего числа людей (которых, конечно же, намного больше, чем ценителей творчества Айзека Азимова).